Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Comparative Analysis of Deep Learning Models for Predicting Causative Regulatory Variants

Gaetano Manzo[1], Kathryn Borkowski[1,2] and Ivan Ovcharenko[1,*]

[1]Computational Biology Branch, Division of Intramural Research, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, USA.,

[2]Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genome-wide association studies (GWAS) have identified numerous noncoding variants associated with complex human diseases, disorders, and traits. However, resolving the uncertainty between GWAS association and causality remains a significant challenge. The small subset of noncoding GWAS variants with causative effects on gene regulatory elements can only be detected through accurate methods that assess the impact of DNA sequence variation on gene regulatory activity. Deep learning models, such as those based on Convolutional Neural Networks (CNNs) and transformers, have gained prominence in predicting the regulatory effects of genetic variants, particularly in enhancers, by learning patterns from genomic and epigenomic data. Despite their potential, selecting the most suitable model is hindered by the lack of standardized benchmarks, consistent training conditions, and performance evaluation criteria in existing reviews.

Results: This study evaluates state-of-the-art deep learning models for predicting the effects of genetic variants on enhancer activity using nine datasets stemming from MPRA, raQTL, and eQTL experiments, profiling the regulatory impact of 54,859 SNPs across four human cell lines. The results reveal that CNN models, such as TREDNet and SEI, consistently outperform other architectures in predicting the regulatory impact of SNP. However, hybrid CNN-transformer models, such as Borzoi, display superior performance in identifying causal SNPs within a linkage disequilibrium block. While fine-tuning enhances the performance of transformer-based models, it remains insufficient to surpass CNN and hybrid models when evaluated under optimized conditions.

Availability: Fine-tuned models https://huggingface.co/tanoManzo, data and analysis https://github.com/tanoManzo/AI4Genomic

Contact: ovcharen@nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have revealed that around 95% of disease-associated genetic variants occur in non-coding regions of the human genome, with causative variants commonly affecting regulatory elements that modulate gene expression (Uffelmann *et al.*, 2021), (Visscher *et al.*, 2017), (Knight, 2014). These regulatory variants can profoundly affect phenotypes and alter disease susceptibility via dysregulation of their target genes (Albert and Kruglyak, 2015).

Deep learning models have emerged as powerful tools for predicting the regulatory effects of genetic variants, particularly in enhancers. These methods leverage high-throughput genomic and epigenomic data to learn DNA sequence patterns associating sequence features with regulatory activity. For instance, convolutional neural networks (CNNs) have been used to identify variants that disrupt transcription factor binding or chromatin accessibility, providing insights into their potential phenotypic impact (e.g., DeepSEA (Zhou and Troyanskaya, 2015), SEI (Chen *et al.*, 2022), and TREDNet (Hudaiberdiev *et al.*, 2023)). Similarly, modern, large transformer-based models can very accurately predict cell-type-specific regulatory effects, such as DNA methylation changes and fine-

mapping of disease-associated loci; these primarily include DNABERT series (Ji *et al.*, 2021), (Zhou *et al.*, 2023), Nucleotide Transformer series (Dalla-Torre *et al.*, 2023), (de Almeida *et al.*, 2024), and Enformer (Avsec *et al.*, 2021).

Selecting the most suitable model for detecting the regulatory effects of genetic variants remains a significant challenge despite several surveys offering detailed overviews of the deep learning ecosystem in this domain (Theodoris *et al.*, 2023), (Consens *et al.*, 2023), (Alharbi and Rashid, 2022). These surveys have highlighted the unique strengths and limitations of various models. However, they lack a unified framework for assessment. Specifically, existing reviews often fail to benchmark models on a standardized dataset, train or fine-tune them under consistent conditions, and evaluate their performance using uniform criteria. Furthermore, there is a fundamental difference between benchmarking models on regulatory regions versus regulatory variants, which is rarely addressed. While regulatory region analyses focus on identifying broader functional elements, regulatory variant assessments require evaluating the impact of specific sequence alterations within these regions, presenting distinct challenges and opportunities for model evaluation.

In this study, we evaluate state-of-the-art deep learning models to predict the effects of genetic variants on enhancer activity in the human genome. Our approach involved curating and integrating nine datasets derived from diverse experimental methodologies, including massively parallel reporter assay (MPRA), reporter assay quantitative trait loci (raQTL), and expression quantitative trait loci (eQTL) studies. These datasets encompass 54,859 single-nucleotide polymorphisms (SNPs) in enhancer regions across four human cell lines. To ensure a robust and thorough assessment, we fine-tuned 22 deep learning models for each cell line, systematically exploring a broad spectrum of architectural designs and parameter configurations. Our results indicate that CNN models outperform more "advanced" architectures, such as transformers, on causative regulatory variant detection. However, fine-tuning significantly boosts the performance of transformer-based architectures, revealing their potential to surpass CNNs under optimized conditions.

2 Results

Comparison of Enhancer Variant Prediction Models

We assessed several state-of-the-art deep learning models for their ability to predict the effects of genetic variants on enhancer activity in humans. Our evaluation encompasses a range of architectural and parameter configurations across model families, focusing on 54,859 single-nucleotide polymorphisms located in enhancer regions from nine datasets originating from four cell lines: K562, HepG2, NPC, and HeLa. We fine-tuned the selected models whenever their architecture allows it (e.g., transformer-based models) and validated them on cell-line-specific enhancer detection tasks to ensure optimal performance before proceeding to variant effect prediction. Each model received a standardized DNA sequence input length of 1 kilobase pair (kbp), and outputs were normalized to represent the log2-fold change of the alternative sequence relative to the reference sequence (as detailed in the Material and Methods section).

We performed regression analyses between model predictions and experimental log2-fold changes for variant effects in enhancers across the human genome sequence (Fig. 1, Sup. Fig. 1, Materials and Methods). CNN-based models, such as TREDNet and SEI, demonstrate the best alignment with experimentally recorded variant effects, achieving Pearson correlation coefficients of 0.297 and 0.276, respectively, while the most accurate transformer-based model, Nucleotide Transformer v2, scores only at 0.105, well below the level of statistical uncertainty. These findings highlight the effectiveness of CNNs in recognizing spatial and structural genomic patterns essential for predicting enhancer variant effects. Convolutional layers identify genomic features, such as local motifs and epigenetic markers, that influence genetic variant effects (Yue et al., 2023). This capability is crucial for capturing the complex genomic relationships underlying gene regulation and enhancer activity.

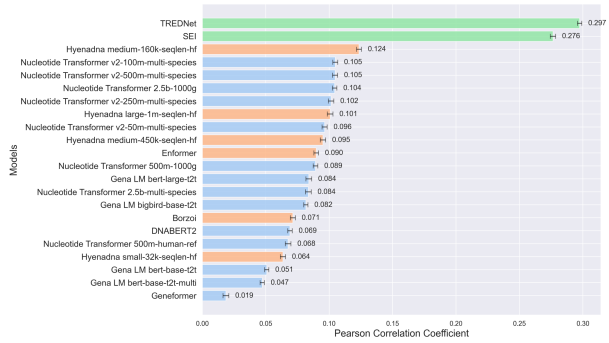

***Fig 1**: Pearson correlation coefficients between model predictions and experimental log2-fold changes for enhancer variant effects across the human genome. Bar colors denote model architectures (CNN: green, transformer: blue, and hybrid: orange). All correlations have p-values < 0.05, and error bars show variance.*

In contrast, transformer-based models fine-tuned for specific cell lines show lower correlation values (0.019–0.105). Despite their ability to model long-range relationships and complex sequences, they struggle to detect the subtle genomic patterns linked to regulatory variant effects. This challenge likely stems from the complexity of genomic data, lower granularity in feature detection compared to CNNs, and the extensive data required to fully exploit transformer capabilities.

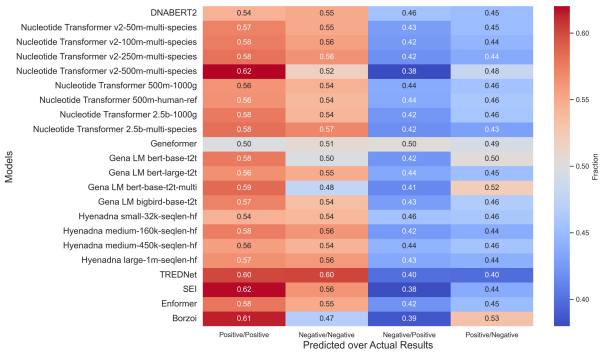

***Fig 2**: Heatmap of model variant predictions (Predicted) versus experimental values (Actual Results). The color intensity represents the fraction of values predicted as positive/negative relative to the experimental positive/negative values. Red indicates higher fractions (desired in the first two columns from the left), while blue indicates lower fractions (desired in the last two columns).*

To bridge the gap between local pattern recognition and long-range dependency modeling, hybrid models combine components such as CNNs, transformers, and LSTMs. Models like the HyenaDNA series and Borzoi show moderate correlations, with HyenaDNA medium-160k-seqlen-hf achieving the highest among hybrids at 0.124. The integration of diverse components enhances performance compared to transformer-based models while optimizing computational costs relative to CNN-based models. However, despite their architectural versatility, these hybrid models do not surpass the performance of CNN-based models, highlighting a trade-off between flexibility and focused efficiency (Fig. 1, Sup. Fig. 1).

The classification task provides a deeper understanding of model performance by evaluating true positive and true negative rates in predicting variants that upregulate or downregulate gene expression (Fig. 2, Sup. Fig. 2). Among 54,859 experimental SNPs, approximately half are linked to upregulation and half to downregulation.

CNN-based models, like TREDNet and SEI, excel in predicting variant effects. TREDNet achieves balanced performance with correctly predicted upregulation (0.60) and downregulation (0.60) rates, while SEI leads in upregulation detection (0.62) but struggles with downregulation (0.56).

Transformer-based models show mixed results. Nucleotide Transformer v2-500m-multi-species matches SEI's upregulation prediction rate (0.62) but underperforms in downregulation (0.52). DNABERT2 performs steadily (0.54, 0.55), while Geneformer shows the weakest outcomes. Borzoi achieves a strong upregulation prediction rate (0.61) but falters in downregulation (0.47). Enformer shows moderate, balanced results (0.58, 0.55), while Hyenadna remains consistent but unremarkable.

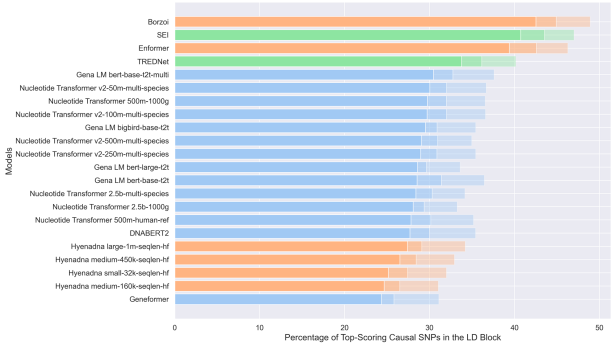

Fig 3*: Percentage of top-scoring causal SNPs predicted within the Linkage Disequilibrium (LD) block by various models. Models are grouped by architecture: CNN (green), transformer (blue), and hybrid (orange). Bars are sorted by top-1 performance, with decreasing transparency indicating lower ranks (top-2 and top-3 results are progressively faded compared to top-1).*

A crucial aspect of each model is its ability to identify causal SNPs within their respective Linkage Disequilibrium (LD) blocks. Essentially, this is the most important function of variant classification models, which can be directly applied to the resolution of causal variants within large blocks of associated variants identified in GWAS studies. To investigate this, we curated a dataset of 10,098 causal SNPs specific to the HepG2 cell line. For each causal SNP, we identified associated LD-block SNPs with an r^2-value above 0.8 within a 500 kbp window, resulting in 263,286 SNPs. With this set, we can directly assess each method's ability to retrieve the causal SNP within this ~26:1 pool of associated SNPs. For each identified SNP, we generated two sequences of 1 kbp: one containing the SNP and the other containing the reference nucleotide. In both cases, the SNP or reference nucleotide were positioned at the center of the sequence. These sequences were then input into several models to assess their ability to predict causal SNPs by calculating the log2 foldchange between the scores of the alternative sequence (containing the SNP) and the reference sequence (Material and Methods section). This procedure was repeated for all LD-block asssociated SNPs and the log2 fold-change score of the known causal SNP was compared to its associated variants. Finally, we computed the percentage of causal SNPs within an LD-block correctly predicted as causal based on their score being either the highest ("top-1" test) or within the top two and three highest scores ("top-2" and "top-3" tests).

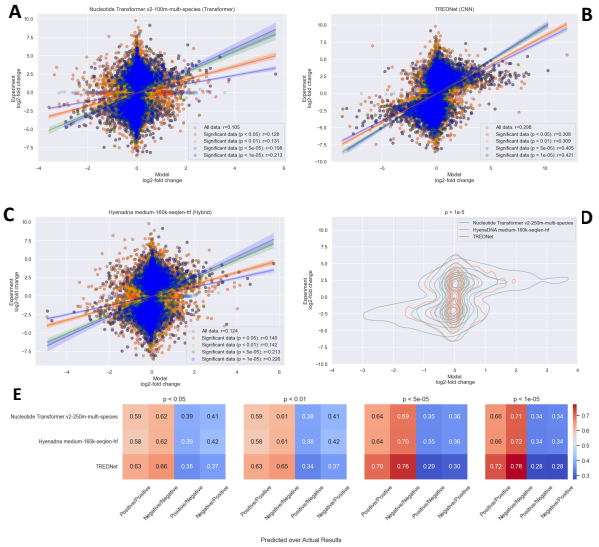

Fig 4*: Model performance relative to experimental data significance, using top models from each architecture category. Panels A-C show scatter plots for TREDNet, Hyenadna, and Nucleotide Transformer, with the correlation between model predictions and experimental results as significance improves. Panel D compares model predictions for highly significant variants. Panel E presents heatmaps of classification metrics at different p-value thresholds, highlighting performance variations in identifying positive/negative outcomes across architectures.*

The Borzoi model achieved the highest accuracy, correctly identifying 42.5% of causal SNPs, followed by the SEI, Enformer, and TREDNet models, with 40.7%, 39.4%, and 33.8% in the top1 test, respectively, (Fig. 3). The causal variant detection rate increased substantially in the top-2 and top-3 test with Borzoi (44.9%, 48.9), SEI (43.6%, 47.0%), Enformer (42.4%, 46.2%), and TREDNet (36.1%, 40.1%). Among the transformer-based models, the Gena LM bert-based-t2t-multi demonstrated the best performance (top-1 30.4%, top-2 32.7%, and top-3 37.5%), marginally surpassing other Nucleotide Transformer and Gena LM variants, which achieved accuracies between 28.0% and 30.0% in top-1 test. The HyenaDNA models, categorized as hybrid architectures, displayed moderate performance, with accuracies ranging from 24.3% to 26.5% in top-1 test, and larger sequence lengths, such as those used in the HyenaDNA large model, led to improved results. The Geneformer model, however, had the lowest accuracy, identifying 24.3% of causal SNPs in top-1 test. Overall, hybrid and CNN architectures outperformed transformer-based models in this application, underscoring their suitability for tasks involving causal SNP detection.

Overall, models like TREDNet and SEI consistently outperform others in both regression analysis and classification tasks, showcasing their strength in capturing local genomic patterns critical for regulatory variant effect prediction. Hybrid models, such as Borzoi and Enformer, offer balanced performance, bridging the gap between local pattern recognition and long-range dependency modeling but fail to surpass CNNs in overall accuracy. Yet, these models excel at accurate detection of causal variants within LD-blocks of associated SNPs. Transformer-based models, despite their potential to model complex sequences and long-range interactions, struggle to match the granularity of CNNs and hybrids.

Impact of Certainty in Experimental Results

The results of experimental assays of regulatory variants depend on the selected degree of certainty in separating significant and insignificant variant effects. To investigate the impact of experimental data significance on correlation with modeling results, we analyzed the top models for each architecture type: TREDNet (CNN), HyenaDNA (hybrid), and Nucleotide Transformer (transformer). The experimental data was binned using the following significance thresholds, $p < 0.5$, $p < 0.1$, $p < 5 \cdot 10^{-5}$, $p < 10^{-5}$.

The results demonstrate that models align better with experimental findings when applied to variants with higher statistical significance. For instance, data points with greater experimental confidence show increased clustering in the first and third quadrants, representing cases where both experimental and predicted results are either positive (upregulation) or negative (downregulation) (Fig. 4, panels A-C, and Sup. Fig. 3.1-3.4).

Among the evaluated architectures, Hyenadna exhibited the greatest improvement in predictive accuracy with the increase of the statistical significance of experimental results (from all data 0.124 to 0.226 with $p < 10^{-5}$, Fig. 4, panel C). However, TREDNet consistently displayed superior overall performance, with a higher density of points aligning with experimental outcomes (from all data 0.298 to 0.421 with $p < 10^{-5}$, Fig. 4, panel B). This trend is further illustrated in the density plots, where TREDNet exhibits increased clustering in the first and third quadrants, demonstrating stronger correlations between model predictions and experimental values compared to both hybrid and transformer models (Fig. 4, panel D). Moreover, the correlation coefficients across varying significance thresholds highlight TREDNet's robustness, achieving the highest values under all conditions. True positives and true negatives increase with the significance of the experimental results (positive/positive 0.63, negative/negative 0.66, with $p < 0.05$; positive/positive 0.72, negative/negative 0.78, with $p < 10^{-5}$), with the hybrid model and the transformer model following in performance (Fig. 4, panel E, and Sup. Fig 2).

Given the strong cell-line specificity of enhancers (Wu and Huang, 2024), determining whether model performance depends on the type of cell line used is essential for understanding their predictive capabilities and limitations. For this, we evaluated whether a model fine-tuned for a specific cell line demonstrates superior performance compared to models from other architectures, such as CNNs, which often rely on generalization rather than cell-line-specific adaptation. Indeed, transformer-based models, pre-trained on extensive datasets and fine-tuned for targeted tasks, can offer both flexibility and accuracy, making them well-suited for capturing the unique regulatory landscape of cell lines. In contrast, CNN-based models, such as TREDNet and SEI, though highly effective for specific tasks, could require additional steps, such as retraining or the addition of specialized layers, to adapt their outputs for cell-line-specific predictions. However, the results indicate that TREDNet and SEI remain the most effective models for predicting variant effects across different cell lines, even after model fine-tuning (Table 1, Sup. Table 1).

TREDNet consistently achieves the highest overall performance, particularly excelling in K562 and HepG2 cell lines (0.315, 0.339), which contain the largest number of experimentally assayed variants (19321 and 16255 respectively). SEI follows as the second-best model overall, also demonstrating strong predictive capabilities (0.297, 0.297, 0.075, and 0.279). In contrast, Nucleotide Transformer models exhibit greater variability in performance across cell lines. For example, the Nucleotide Transformer v2-250m-multi-species model aligns with experimental results in K562 and HepG2 (0.1998 and 0.11) but performs less effectively in NPC and HeLa cells. This variability could stem from differences in pre-training datasets—such as multi-species versus human-specific data—and the fine-tuning approach. Notably, the hybrid model Enformer, which integrates features from transformers and CNNs, performs particularly well in HeLa cells (0.245), ranking second only to SEI (0.279).

Table 1. *Pearson correlation coefficient for various deep learning models across four cell lines: K562 (19321 SNPs), HepG2 (16255 SNPs), NPC (14042 SNPs), and HeLa (5241 SNPs). Bold and underline styles denote the top and second-highest correlations per cell line, respectively.*

	Cell Lines			
Models	**K562 (19321 SNPs)**	**HepG2 (16255 SNPs)**	**NPC (14042 SNPs)**	**Hela (5241 SNPs)**
DNABERT2	0.08592	0.09644	-0.00387	0.11972
Nucleotide Transformer v2-50m-multi-species	0.14681	0.10278	0.01974	0.0743
Nucleotide Transformer v2-100m-multi-species	0.15166	0.12655	0.01638	0.04183
Nucleotide Transformer v2-250m-multi-species	0.16575	0.11014	0.02279	0.02956
Nucleotide Transformer v2-500m-multi-species	0.19861	0.11433	0.02789	0.03601
Nucleotide Transformer 500m-1000g	0.12297	0.11631	0.00141	0.02245
Nucleotide Transformer 500m-human-ref	0.14938	0.11868	0.01302	0.06652
Nucleotide Transformer 2.5b-1000g	0.14717	0.11156	0.005	0.08712
Nucleotide Transformer 2.5b-multi-species	0.15293	0.1388	0.02652	0.06617
Geneformer	0.00522	0.02728	-0.00163	-0.00736
Gena LM bert-base-t2t	0.07685	0.04953	0.02416	0.02883
Gena LM bert-large-t2t	0.11748	0.08822	0.02954	0.06935
Gena LM bert-base-t2t-multi	0.07662	0.06238	0.01037	0.02605
Gena LM bigbird-base-t2t	0.13636	0.09943	0.02051	0.04477
Hyenadna small-32k-seqlen-hf	0.08441	0.09808	0.0016	0.03099
Hyenadna medium-160k-seqlen-hf	0.14837	0.14239	-0.00701	0.04627
Hyenadna medium-450k-seqlen-hf	0.07719	0.13432	-0.00749	0.05511
Hyenadna large-1m-seqlen-hf	0.11761	0.13619	0.00062	0.04606
TREDNet	**0.31544**	**0.33958**	**0.07501**	0.07595
SEI	0.29739	0.29763	0.07307	**0.27951**
Enformer	0.05851	0.1412	0.00843	0.24503
Borzoi	0.03322	0.12503	0.0235	-0.04136

As observed in the previous test, CNN-based models demonstrate superior performance, showcasing robustness across cell-line variants. We also assess the impact of the certainty of experimental results, SEI, which does not require training or fine-tuning for specific cell lines, outperforms fine-tuned models. TREDNet, likely benefiting from its second-phase training tailored to the specific cell line, achieves the best overall performance, further highlighting its ability to leverage cell-line-specific information effectively (Sup. Table 2, Sup. Table 3).

Enhancer Detection Models for Variant Effect Assessment

Models designed to predict variant effects rely primarily on the change in confidence when detecting the enhancer region after introducing the variant. Next, we investigated whether improving the accuracy of enhancer sequence detection enhances predictions of variant effects. While SEI, Borzoi, and Enformer do not require fine-tuning, TREDNet, despite being a CNN-based model, incorporates a two-phase training process. The second phase can be considered fine-tuning (Hudaiberdiev *et al.*, 2023).

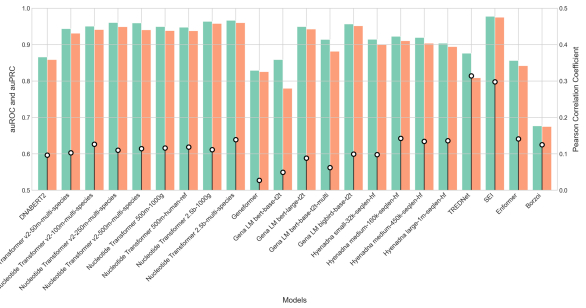

Fig 5*: Performance comparison of different models for enhancer detection in HepG2 cell across multiple metrics. Left axis: Area Under the Receiver Operating Characteristic Curve (auROC, green bars) and Area Under the Precision-Recall Curve (auPRC, orange bars). Right axis: Pearson Correlation Coefficient (black pins).*

Transformer-based architectures, particularly Nucleotide Transformer models pre-trained on multi-species datasets, emerge as strong performers in enhancer classification (Fig. 5). Models such as Nucleotide Transformer v2-250m-multi-species and Nucleotide Transformer 2.5b-multi-species consistently demonstrate high accuracy, as indicated by their auROC and auPRC metrics (auROC: 0.967, 0.963; auPRC: 0.958, 0.951). However, the correlation between the models' outputs and the MPRA experimental results, which measure the fold-change effect between reference and alternative sequences, remains relatively low (ranging from 0.112 to 0.15). SEI stands out as the top performer in detecting enhancer activity, achieving the highest auROC (0.974) and auPRC (0.971) scores among all models. It is also the second best for the correlation between model outputs and experimental results (0.295). TREDNet demonstrates strong performance with a higher correlation (0.318), while balancing high auROC (0.872) and auPRC (0.814) scores. The Hyenadna models show competitive classification metrics, with the Hyenadna medium-160k-seqlen-hf and Hyenadna large-1m-seqlen-hf versions attaining notable auROC (0.933, 0.904) and auPRC (0.921, 0.893) scores. However, their Pearson correlation coefficients are moderate (0.152, 0.151), indicating that while these models classify enhancer activity effectively, further optimization is needed to improve their predictive accuracy for variant-specific effects. Borzoi performs the weakest in detecting enhancer regions (auROC: 0.672, auPRC: 0.65), while Geneformer fares poorly in detecting variants (auROC: 0.027). Yet, this weak enhancer detection performance doesn't translate in inaccurate detection of causal variants, as Borzoi outperforms all other methods in this critical test.

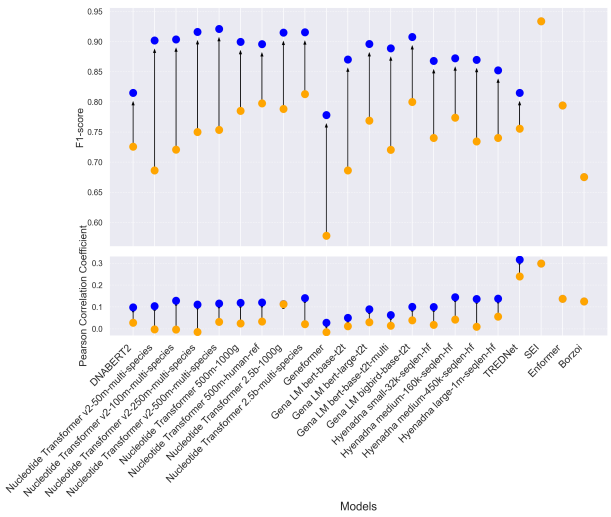

***Fig 6**: Comparison of F1 score (top) and Pearson Correlation Coefficient (bottom) between fine-tuned (blue dots) and one-shot (orange dots) models across various architectures, evaluated on Dataset 3, HepG2 cell line, (14182 SNPs). Arrows indicate the changes in performance between fine-tuned and one-shot implementations for each model.*

Fine-tuned models consistently outperform their one-shot counterparts, achieving F1 scores between 0.85 and 0.92, compared to 0.70 to 0.80 for one-shot models (Fig. 6, top). This improvement in performance is reflected in the correlation between model predictions and experimental results, particularly in detecting enhancer variants (Fig. 6, bottom). The performance enhancement is consistent across most architectures, with CNN-based models showing remarkable results even without fine-tuning. However, SEI, which is available only in its 'one-shot' implementation, achieves a high F1 score and correlation, surpassing many fine-tuned transformer models. Additionally, some transformer-based models, such as Nucleotide Transformer 2.5b-1000g, show no improvement in detecting enhancer variants.

These results suggest that while fine-tuning generally enhances model performance, optimizing a model for better detection of enhancer activity does not necessarily improve its ability to detect variant regulatory effects. The underlying architecture plays a crucial role in determining baseline effectiveness. Consistent with previous analyses, CNN-based and hybrid architectures seem inherently better suited for detecting the variant effects of regulatory elements.

3 Discussion

The comparative evaluation of deep learning models for predicting the effects of genetic variants on enhancer activity reveals insights into the strengths and limitations of different model architectures. Our results suggest that CNN-based models, such as TREDNet and SEI, outperform transformer-based models and hybrid architectures across a range of tasks, including regression and classification. The hybrid model Borzoi achieves excellent results in causal SNP prediction. These models excel due to their ability to capture local genomic features that are crucial for regulatory variant effect prediction.

Performance of CNN-Based Models

CNN-based models, TREDNet and SEI, consistently achieve the best performance in both regression (Fig. 1) and classification tasks (Fig. 2). TREDNet shows robust predictive capability across both upregulated and downregulated variants, with balanced true positive and true negative rates. The performance of SEI, which excels at detecting upregulated variants, further underscores the utility of CNNs in recognizing local genomic patterns, such as motifs and epigenetic markers, which play a pivotal role in enhancer activity (Yue et al., 2023). This capacity to detect local patterns is crucial for understanding the effect of genetic variants on enhancer activity and regulatory elements. CNNs ability to model spatial relationships in genomic sequences, coupled with their relatively simpler architecture, makes them well-suited for this task, as evidenced by the high Pearson correlation coefficients observed in our regression analysis (0.297 for TREDNet and 0.276 for SEI). The trade-off in CNN-based models lies in their inability to model long-range dependencies as effectively as transformer and hybrid models, which could explain their relatively lower performance in tasks requiring such global sequence context.

Transformer-Based Models and Their Limitations

Transformer-based models, such as the Nucleotide Transformer and DNABERT 2, demonstrate some of the potential advantages of deep learning models, such as their ability to capture long-range dependencies and complex sequence patterns. However, these models perform poorly in comparison to CNN-based models, particularly when predicting variant effects in enhancers. Our findings indicate that transformer-based models struggle with the granularity required for regulatory variant effect prediction, particularly in the case of subtle genomic patterns that influence enhancer activity (Fig. 1). While transformer models achieve moderate F1 scores in detecting enhancer sequences (e.g., Nucleotide Transformer v2-500m-multi-species), they fail to match the performance of CNN- and hybrid-based models in causal SNP prediction (Fig. 3), with accuracies between 28.0% and 30.0%. This limitation may arise from the difficulty in detecting fine-grained regulatory elements within a complex genomic

context, a task that CNN/hybrid models handle more effectively due to their local focus.

Furthermore, transformer-based models require substantial amounts of data to fully exploit their capabilities. The limited dataset available in this study, combined with the model's need for high data diversity to capture long-range interactions, might account for their lower performance in enhancer variant effect prediction. This highlights a potential area for improvement in transformer model training, where expanding the dataset or refining the model to focus on more granular regulatory signals could improve performance.

Hybrid Models: Local and Long-Range Dependency Modeling

Hybrid models, such as HyenaDNA and Borzoi, offer a balanced approach by integrating the strengths of CNNs, transformers, and LSTMs. These models aim to bridge the gap between local pattern recognition and long-range dependency modeling. HyenaDNA, in particular, shows moderate correlations between the model's output and experimental results (Fig. 1), suggesting that hybrid architectures may offer a compromise between computational efficiency and predictive accuracy. Borzoi stands out as a hybrid model that significantly outperforms other models in detecting causal SNPs (Fig. 3), demonstrating its potential for tasks requiring the integration of both local and global sequence contexts. This exceptional performance in causal SNP detection highlights Borzoi's strength in identifying causal variants, which is a key task in detection of disease-causal variants in analysis of GWAS data. However, despite this success in causal SNP prediction, Borzoi, like other hybrid models, does not yet surpass CNN-based models in overall predictive accuracy. This further suggests that hybrid models should not be compared to CNN models based purely on the predictive power of regulatory variants, but evaluated fully in the context of the specific biological tasks.

Impact of Certainty in Experimental Results

The certainty in experimental results plays a significant role in improving the alignment between model predictions and experimental outcomes. Our analysis indicates that models, particularly TREDNet, align better with experimental data when higher statistical confidence is applied to variant effects (Fig. 4). This trend highlights the importance of refining experimental datasets and improving their statistical rigor to enhance model predictions. For instance, higher p-value thresholds (e.g., $p < 10^{-5}$) result in better clustering of experimental and predicted results, suggesting that models are more accurate when validated against robust experimental data. TREDNet demonstrates superior performance across varying significance thresholds, maintaining strong predictive accuracy even when experimental results are less certain. This finding emphasizes the importance of robustness in CNN-based models and their ability to generalize well across a variety of experimental conditions.

Cell-Line-Specific Adaptation and Generalization

The cell-line-specific adaptation of deep learning models is a crucial aspect of enhancer variant effect prediction. Our results reveal that transformer-based models, when fine-tuned for specific cell lines, can capture the unique regulatory landscape of these cell lines, but their performance still lags behind CNN-based models (Table 1). For instance, models like the Nucleotide Transformer v2-250m-multi-species achieve reasonable results in cell lines like K562 and HepG2 but perform poorly in others, such as NPC and HeLa. This variability may be attributed to the pre-training datasets used for these models, which may not fully capture the regulatory environment of specific cell lines.

CNN-based models demonstrate strong performance across cell lines without the need for fine-tuning, suggesting that these models are more robust to cell-line-specific variations. The results underscore the value of CNNs in regulatory variant prediction tasks, where cell-line-specific adaptations may not always be feasible due to limited data or computational constraints.

Enhancer Detection Models for Variant Effect Assessment

Models designed for enhancer detection also play a critical role in variant effect prediction. While models such as Nucleotide Transformer v2-250m-multi-species achieve high classification accuracy for enhancer regions (auROC: 0.967, auPRC: 0.958), their correlation with experimental results remains low, underscoring the need for further optimization in predictive accuracy for regulatory variants (Fig. 5). In contrast, CNN-based models like SEI show strong performance in both enhancer detection and variant effect prediction, further emphasizing the importance of local genomic pattern recognition.

Fine-tuning models consistently results in improved performance, particularly for transformer-based architectures. However, SEI, despite being available only in a 'one-shot' implementation, outperforms many fine-tuned Transformer models.

In conclusion, our findings suggest that SEI and TREDNet are the most effective models for predicting the effects of genetic variants on enhancer activity. Hybrid models like Borzoi demonstrate strong potential in extracting causal SNPs from GWAS data, showcasing their value in applications that require the integration of both local and global sequence contexts. Transformer-based models, on the other hand, are particularly effective at identifying enhancer regions due to their ability to capture long-range dependencies, although they still face challenges with fine-grained genomic features and often require substantial data for optimal performance. These results highlight the strengths of different model architectures for various aspects of regulatory variant prediction and underscore the need for further optimization to enhance their performance. Future work should aim to improve feature detection in transformer models and refine the flexibility of hybrid architectures to better capture both local and long-range genomic dependencies.

4 Material and Methods

3.1 Deep Learning Models

This study utilizes several deep learning models designed for genomic sequence analysis, with a particular focus on applications to the Homo sapiens genome. The models represent the forefront of innovation in deep learning for genomics, leveraging diverse architectures and pre-training paradigms to handle the complexity of human genomic data. To maintain focus on models explicitly designed for the human genome, this study excluded models like EVO and GPN, which are primarily aimed at other species or biological domains.

DNABERT 2

DNABERT 2 is a highly efficient foundation model specifically designed for multi-species genome analysis, addressing the increasing demand for cross-species genomic studies. It employs Byte Pair Encoding (BPE) for tokenization, which ensures compact and efficient sequence representation, reducing the computational overhead commonly associated with processing long genomic sequences. The model incorporates advanced features such as Attention with Linear Biases (ALiBi) and Flash Attention mechanisms, which not only improve its computational efficiency but also enhance its scalability, enabling it to handle extensive genomic datasets with reduced latency.

Despite its relatively smaller size compared to larger foundation models, DNABERT 2 consistently delivers performance on par with its more

resource-intensive counterparts. This makes it particularly well-suited for applications where computational resources are limited, such as real-time analysis, edge computing, or research in under-resourced settings. Furthermore, its architecture is optimized for tasks like sequence classification, motif discovery, and variant effect prediction across diverse species, highlighting its versatility and robustness (Zhou *et al.*, 2023).

Nucleotide Transformer Series

The Nucleotide Transformer series encompasses models of varying parameter sizes—50M, 100M, 250M, 500M, 2.5B and v2—designed to cater to a wide range of genomic research applications. Pre-trained on multi-species genomes, these models are optimized for broad applicability, enabling researchers to address diverse challenges in comparative genomics, functional annotation, and variant effect prediction across species.

The models utilize a 6-mer tokenization strategy, which effectively captures patterns in genomic sequences, allowing for a nuanced representation of sequence context. Key architectural enhancements, such as rotary embeddings and Swish activation functions, further bolster the models' ability to model intricate genomic features. These features enhance the precision and scalability of the models, making them well-suited for tasks requiring detailed analysis of local and long-range genomic dependencies.

The Nucleotide Transformer series demonstrates a strong balance between scalability and performance, with models of different sizes tailored to meet the demands of various computational environments, from resource-constrained setups to high-performance clusters. Their multi-species training regime ensures robust generalization across diverse genomic datasets, solidifying their utility for cross-species studies and evolutionary biology. Collectively, these advancements position the Nucleotide Transformer models as cutting-edge tools in the genomic research toolkit, providing both versatility and high accuracy for modern bioinformatics workflows (Dalla-Torre *et al.*, 2023), (de Almeida *et al.*, 2024).

Geneformer

Geneformer is a model tailored for analyzing gene network dynamics and classifying cell states, offering a specialized tool for single-cell transcriptomics. Pre-trained on extensive single-cell transcriptome datasets, Geneformer delivers robust and accurate predictions, even in the presence of technical noise or variability inherent to single-cell data.

The model's rank-based analytical approach enhances its stability and reliability, ensuring consistent performance across diverse datasets and experimental conditions. This makes it particularly well-suited for single-cell studies, where precision and resilience to noise are critical for identifying subtle patterns in gene expression (Theodoris *et al.*, 2023).

GENA-LM Series

The GENA-LM series, encompassing BERT-base, BERT-large, and BigBird variants, is meticulously designed to optimize genomic sequence analysis. By employing diverse architectural frameworks, the series achieves a strategic balance between computational efficiency and predictive performance, catering to a wide range of genomic tasks.

BERT-base and BERT-large models excel in capturing intricate sequence patterns through their transformer-based architecture, while BigBird introduces a sparse attention mechanism, enhancing scalability for processing extended genomic sequences. This versatility allows the GENA-LM series to adapt effectively to tasks ranging from local feature detection to modeling long-range genomic dependencies (Fishman *et al.*, 2023).

Enformer

Enformer is a groundbreaking deep learning model developed for predicting gene expression directly from DNA sequences. Featuring 11 Transformer layers and approximately 600 million parameters, it is engineered to process sequences of up to 200,000 base pairs. This extensive sequence length enables the model to predict over 5,000 epigenetic and transcriptional features, effectively capturing long-range regulatory interactions spanning distances of up to 100 kilobases.

Enformer's capacity to model these interactions provides unparalleled insights into the regulatory landscape of the genome, making it a vital tool for the study of regulatory elements and their effects on gene expression. Its robust architecture positions it as a key asset in advancing our understanding of genomic regulation and its implications in health and disease (Avsec *et al.*, 2021).

HyenaDNA Series

The HyenaDNA series comprises advanced models specifically designed for processing exceptionally long genomic sequences, ranging from 32,000 to 1,000,000 base pairs. These models excel in runtime scalability, making them particularly effective for analyzing long-range genomic interactions. This capability is crucial for understanding complex regulatory networks and their role in gene regulation and expression. By facilitating the exploration of extended genomic regions, the HyenaDNA models provide valuable insights into the intricate interplay of regulatory elements (Nguyen *et al.*, 2023).

Borzoi

Borzoi is an advanced model specifically designed to predict cell- and tissue-specific RNA-seq coverage directly from DNA sequences. By integrating multiple layers of regulatory predictions, it excels in capturing the intricate mechanisms of gene regulation. Additionally, Borzoi's ability to accurately score variant effects makes it a vital tool for regulatory genomics, particularly in uncovering cis-regulatory patterns. Its robust predictive capabilities position it as a key resource for advancing our understanding of the regulatory genome (Linder *et al.*, 2023).

SEI

SEI is a deep learning model designed to predict the effects of genetic variants on cis-regulatory activity across 21,907 chromatin profile targets. It processes 4,096bp input sequences, which are one-hot encoded, through a multi-layer architecture that includes residual dual linear and nonlinear paths for enhanced learning efficiency and expressiveness, residual dilated convolution layers to capture multi-scale sequence patterns, and an efficient global integration layer to incorporate long-range dependencies. The model concludes with spatial basis function and output layers, enabling precise prediction of chromatin profiles. This architecture ensures a balance between computational efficiency and expressiveness, making SEI a robust tool for variant effect prediction and regulatory genomics (Chen *et al.*, 2022).

TREDNet

TREDNet is specifically designed for enhancer prediction and variant prioritization. Its two-phase architecture integrates a CNN for epigenomic signal prediction with a secondary CNN for enhancer prediction. The first phase predicts key epigenomic features, such as histone modifications, DNase I hypersensitive sites, and transcription factor binding, across multiple cell types. The second phase refines these predictions to identify enhancers with high precision. Trained on comprehensive datasets from ENCODE and NIH Roadmap Epigenomics, TREDNet is capable of performing in silico saturated mutagenesis and prioritizing variants. This makes it a versatile and robust tool for enhancer analysis and regulatory variant interpretation (Hudaiberdiev *et al.*, 2023).

3.2 Datasets

Training Data

We constructed a dataset of positive and control enhancer sequences from four human cell lines: HepG2, Hela, K562, and Neural Progenitor Cells (NPC). These cell lines were selected for their relevance in genomic research and the availability of comprehensive epigenomic data, such as DNase-Seq and histone modification profiles, sourced from the ENCODE project using genome annotations for hg19 and hg38.

Positive enhancer sequences were defined through a two-step process. First, open chromatin regions were identified using DNase-Seq data. These regions were then intersected with H3K27ac histone modification signals, a hallmark of active enhancers. To refine the dataset, we excluded non-enhancer regions such as exonic regions (coding sequences), promoter regions (near transcription start sites), and low-confidence regions listed in the ENCODE Blacklist (e.g., repetitive sequences). Specific blacklisted peaks from UCSC browser tables (wgEncodeDacMapabilityConsensusExcludable) were also removed.

Control sequences were selected using a similar filtering process but specifically excluded regions overlapping both DNase-Seq peaks and H3K27ac signals to ensure they represented non-active genomic regions.

Both positive and control sequences underwent identical exclusion criteria to eliminate confounding genomic features such as exons and promoters. All sequences were standardized to 1 kb in length to balance sufficient coverage of regulatory regions with computational efficiency. This size is widely used in enhancer studies as it captures relevant regulatory elements while avoiding extraneous genomic features.

The dataset includes an equal number of positive and control sequences for each cell line: HepG2 (14,062 each), NPC (12,466 each), K562 (19,968 each), and Hela (31,247 each). Metadata from the ENCODE project was used to retrieve biosample IDs and associated DNase-Seq and histone modification data. Reference files provided coordinates for exons, blacklist regions, and promoters, ensuring only valid enhancer or control regions were included in the final dataset.

Validation Data

Datasets 1 and 2 were sourced from the study by Arensbergen et al. (van Arensbergen *et al.*, 2019), which utilized the Survey of Regulatory Elements (SuRE) reporter assay to identify reporter assay quantitative trait loci (raQTLs)—SNPs influencing regulatory element activity. For K562 cells, 19,237 raQTLs were identified with an average allelic fold change of 4.0-fold, while HepG2 cells yielded 14,183 raQTLs with a higher average fold change of 7.8-fold. Most raQTLs were cell-type-specific, showing limited overlap between the two cell lines.

Dataset 3 was derived from Kircher et al. (Kircher *et al.*, 2019) and focuses on a ~600 bp enhancer region at the SORT1 locus (1p13). This region includes SNP rs12740374, which creates a C/EBP binding site that influences SORT1 expression. The enhancer was tested in HepG2 cells using a luciferase reporter assay. Data for flipped sequences were excluded due to the orientation-independent nature of enhancers, and deletion datasets were omitted due to computational model constraints, which focus on single-nucleotide variants.

Dataset 4, from Ulirsch et al. (Ulirsch *et al.*, 2016) used MPRA to investigate 2,756 variants linked to red blood cell traits in K562 cells. The study identified 84 highly significant SNPs with functional effects, enriched in open chromatin regions shared between K562 cells and primary human erythroid progenitors (HEPs). This dataset provides insights into erythroid-specific regulatory activity.

Dataset 5, generated by Vockley et al. (Vockley *et al.*, 2015), includes MPRA results for 284 SNPs tested in HepG2 cells. These SNPs, located within regions of high linkage disequilibrium identified through eQTL analyses, offer a detailed map of allele-specific regulatory effects relevant to liver-specific gene regulation and genotype-phenotype associations.

Dataset 6, from Weiss et al. (Weiss *et al.*, 2021), includes lentiMPRA results for 14,042 single-nucleotide variants fixed or nearly fixed in modern humans but absent in archaic humans (e.g., Neanderthals). Tested across three cell types—embryonic stem cells, neural progenitor cells, and fetal osteoblasts—the study identified 1,791 sequences with regulatory activity, including 407 with differential expression between modern and archaic humans. This dataset focuses on NPC data.

Datasets 7, 8, and 9, from Birnbaum et al. (Birnbaum *et al.*, 2014), include MPRA results for three coding exons—SORL1 exon 17, TRAF3IP2 exon 2, and PPARG exon 6—tested in HeLa cells. Respectively, these datasets cover 1,962, 1,614, and 1,665 SNPs. The study explored enhancer activity within these coding exons (eExons), finding that ~12% of mutations altered enhancer activity by at least 1.2-fold, with ~1% causing changes of less than 2-fold. This version eliminates repetition while maintaining all critical details about the datasets and their relevance.

3.3 Methods

Data Pre-Processing

The datasets used in this study were preprocessed through a unified pipeline to standardize and extract relevant features, ensuring consistency across data from HepG2, K562, NPC, and HeLa cell lines. Chromosome and position information were extracted for all datasets, along with reference and alternative alleles. Log2 activity metrics, such as fold-change values or effect sizes, were calculated to quantify regulatory activity. Signed p-values were derived from raw or adjusted p-values to capture both the statistical significance and directionality of regulatory effects. This preprocessing ensured that variations in dataset structure did not affect downstream analyses. To further refine the data, hg19 genomic coordinates were used for variant mapping, and enhancer activity metrics were consistently applied across datasets. Differences in expression or activity levels between experimental conditions (e.g., mutant vs. control or modern vs. archaic sequences) were computed where applicable. All datasets were ultimately converted into a standardized format containing chromosome information, positions, reference/alternative alleles, log2 activity ratios, signed p-values, raw p-values, and other relevant regulatory features.

Fine-tuning

To adapt pre-trained DNA language models for specific regulatory prediction tasks, we fine-tuned them using sequences derived from the previously described datasets. A sequence length of 1 kilobase pairs was chosen to balance genomic context with model compatibility. The fine-tuning process was implemented using Hugging Face's transformers library, which provided a flexible framework for seamless model training and evaluation. Most models, including LongSafari/hyenadna-small-32k-seqlen-hf, Geneformer, and other nucleotide transformers, were sourced from the Hugging Face Model Hub, leveraging the Trainer API for efficient fine-tuning. Models not integrated into the Hugging Face ecosystem, such as TREDNet, Enformer, SEI, and Borzoi, were obtained from GitHub repositories and required custom pipelines for data preprocessing and inference.

Fine-tuning was conducted on four biosample-specific datasets: HeLa (*BioS2*), neural progenitor cells (NPC, *BioS45*), HepG2 (*BioS73*), and K562 (*BioS74*). For each dataset, sequences were cleaned to remove ambiguous bases (Ns) and split into training, validation, and test sets. Tokenization was performed using model-specific tokenizers with a maximum sequence length of 512 tokens. For non-fine-tuned models like TREDNet and Enformer, the same one kbp sequence length was used, with padding applied as needed to meet input requirements. While this approach ensured consistency in sequence representation across all models, it limited the potential of architectures optimized for different input lengths or formats.

Key hyperparameters for fine-tuning included a batch size of 8, a learning rate of 10-5, and a maximum of 200 epochs. Early stopping with patience of 5 epochs was employed to prevent overfitting, and training progress was logged every 500 steps for monitoring purposes. Fine-tuned models were saved in dedicated output directories named according to model checkpoints and biosample IDs and subsequently pushed to the Hugging Face Model Hub for reproducibility. GPU memory was cleared after each training session to optimize resource usage.

This workflow enabled adapting general-purpose DNA language models to cell-type-specific regulatory prediction tasks while maintaining compatibility with non-fine-tuned models through consistent sequence processing. Training and fine-tuning were conducted on NIH BioWulf node clusters equipped with A100 GPUs, which provided sufficient computational power for large-scale datasets. For instance, training on the largest dataset (~130,000 sequences for positive and control enhancer regions) utilized 8 A100 GPUs and was completed within hours.

Validation

The prediction process involved generating regulatory effect predictions for each dataset by calculating log2 ratios of alternative to reference allele probabilities for each variant. This metric provided a standardized way to evaluate the relative regulatory effects of alternative versus reference alleles. For each dataset, predictions for reference and alternative alleles were processed by first calculating margin logits, representing the difference in predicted class probabilities. These logits were transformed into probabilities using the sigmoid function. The log2 ratio of these probabilities was then computed to quantify the regulatory differences between the two alleles.

Additionally, precomputed predictions stored in pickle files were loaded for specific models and datasets. Predictions for reference and alternative alleles were used to compute a ratio, representing the ratio of alternative

to reference predictions. This ratio was then log2-transformed to ensure consistency in the output format. Additional preprocessing steps were applied for some datasets based on specific requirements, such as loading specialized prediction files.

The prediction pipeline was applied across all datasets by iterating through experiments defined in a metadata table. For each experiment, data were preprocessed, and predictions were computed or loaded accordingly. The results were compiled into a unified dictionary containing log2 prediction ratios for all variants across experiments. As with the fine-tuning process, the validation of variant/reference sequences was performed using sequences of 1 kilobase pairs (1kb), ensuring consistency in genomic context and input length across models.

Several metrics were used to evaluate the models' performance, including Area Under the Curve (AUC), Precision-Recall Curve (PRC), F1-score, and correlation coefficients, such as Pearson and Spearman correlation. These metrics allowed for a comprehensive assessment of model performance, covering both classification accuracy and the relationship between predicted and observed values.

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Funding

This research was supported by the Division of Intramural Research of the National Library of Medicine, National Institutes of Health; 1-ZIA-LM200881-12 (to I.O.).

References

Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. Nat. Rev. Genet., 16, 197–212.

Alharbi,W.S. and Rashid,M. (2022) A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics, 16, 26.

de Almeida,B.P. et al. (2024) SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models. BioRxiv.

van Arensbergen,J. et al. (2019) High-throughput identification of human SNPs affecting regulatory element activity. Nat. Genet., 51, 1160–1169.

Avsec,Ž. et al. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods, 18, 1196–1203.

Birnbaum,R.Y. et al. (2014) Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. PLoS Genet., 10, e1004592.

Chen,K.M. et al. (2022) A sequence-based global map of regulatory activity for deciphering human genetics. Nat. Genet., 54, 940–949.

Consens,M.E. et al. (2023) To Transformers and Beyond: Large Language Models for the Genome. arXiv.

Dalla-Torre,H. et al. (2023) The nucleotide transformer: building and evaluating robust foundation models for human genomics. BioRxiv.

Fishman,V. et al. (2023) GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences. BioRxiv.

Hudaiberdiev,S. et al. (2023) Modeling islet enhancers using deep learning identifies candidate causal variants at loci associated with T2D and glycemic traits. Proc Natl Acad Sci USA, 120, e2206612120.

Ji,Y. et al. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37, 2112–2120.

Kircher,M. et al. (2019) Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun., 10, 3583.

Knight,J.C. (2014) Approaches for establishing the function of regulatory genetic variants involved in disease. Genome Med., 6, 92.

Linder,J. et al. (2023) Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. BioRxiv.

Nguyen,E. et al. (2023) HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. arXiv.

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

Uffelmann,E. et al. (2021) Genome-wide association studies. Nat. Rev. Methods Primers, 1, 59.

Ulirsch,J.C. et al. (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell, 165, 1530–1545.

Visscher,P.M. et al. (2017) 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet., 101, 5–22.

Vockley,C.M. et al. (2015) Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res., 25, 1206–1214.

Weiss,C.V. et al. (2021) The cis-regulatory effects of modern human-specific variants. eLife, 10.

Wu,C. and Huang,J. (2024) Enhancer selectivity across cell types delineates three functionally distinct enhancer-promoter regulation patterns. BMC Genomics, 25, 483.

Yue,T. et al. (2023) Deep learning for genomics: from early neural nets to modern large language models. Int. J. Mol. Sci., 24.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods, 12, 931–934.

Zhou,Z. et al. (2023) DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv.