Building foundation models for regulatory genomics requires rethinking large language models

Ziqi Tang [1] **Peter K Koo** [1]

Abstract

As large language models (LLMs) continue to gain attention as foundation models in numerous domains, it is intriguing to consider their potential impact on functional genomics. Specifically, we aim to determine whether existing genomic LLMs have reached the esteemed "foundation" status and can effectively encode meaningful representations that address a diversity of cell-type specific prediction tasks in regulatory genomics, such as enhancer activity, non-coding variant effect predictions, and RNA stability. Our study reveals that current LLMs fall short of the performance achieved by more specialized supervised models. These findings underscore the lack of versatility of existing genomic LLMs and raise potential avenues moving forward, including how to train them effectively, understanding what information they encode, and how this knowledge can be leveraged for functional genomics.

1. Introduction

Large language models (LLMs) have showcased exceptional achievements in natural language processing (Devlin et al., 2018; OpenAI, 2023) and have also made significant strides in the field of protein sequence analysis (Rives et al., 2019; Madani et al., 2020; Elnaggar et al., 2021). These models, often referred to as "foundation models" acquire highly effective embeddings or representations of input data through self-supervised learning, which proves immensely valuable across various downstream prediction tasks. Notably, LLMs trained on protein sequences have been leveraged to predict protein structures from a single sequence (Lin et al., 2023; Chowdhury et al., 2022; Wang et al., 2022; Wu et al., 2022) and facilitate protein design (Madani et al., 2023; Ferruz & Höcker, 2022), among numerous other applications.

[*]Equal contribution [1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, NY, USA. Correspondence to: Ziqi Tang <ztang@cshl.edu>, Peter Koo <koo@cshl.edu>.

The 2023 ICML Workshop on Computational Biology. Honolulu, Hawaii, USA, 2023.

The representations learned by protein LLMs encompass biochemical features (Vig et al., 2020) and incorporate evolutionary information (Rao et al., 2020; Bhattacharya et al., 2021; Hie et al., 2022; Lupo et al., 2022), such as conserved motifs, covariation associated with protein contacts, and phylogenetic relationships.

The objective of masked-language modeling (MLM) facilitates the acquisition of contextual embeddings by the LLM. In the case of sequence data, each element is assigned a vector representation that captures not only its own characteristics but also the information pertaining to other elements in the sequence and their interdependencies. In contrast, the conventional one-hot encoding approach treats each element independently, employing identical representations for identical characters (such as amino acids for proteins or nucleotides for DNA) regardless of their respective positions. Consequently, the responsibility of learning contextual information falls solely on the machine learning model.

In the field of genomics, there has been a recent emergence of LLMs, including BERT-style models like DNABERT (Ji et al., 2021) and Nucleotide Transformers (NT) (Dalla-Torre et al., 2023), among others (Zaheer et al., 2020; Chen et al., 2022; 2023), as well as convolution-based models like Genomic Pre-trained Network (GPN) (Benegas et al., 2022). These models have exhibited the ability to capture gene features in their representations and can be utilized to predict single-nucleotide variant effects. Since most LLMs are trained on reference genomes, it remains unclear to what extent they are able to learn cell-type specific information, which is crucial for comprehending cell-type specific regulatory genomics. Consequently, to achieve improved performance on regulatory genomics prediction tasks, it is often necessary to fine-tune the LLM weights on the desired downstream task (Mo et al., 2021; Yamada & Hamada, 2022; Yang et al., 2022). This reliance on fine-tuning poses challenges, as foundation models are typically large and fine-tuning on individual tasks demands substantial GPU resources, which may not be readily available to many academic labs. Therefore, the extent to which existing LLMs can genuinely serve as foundation models for functional genomics, without necessitating additional fine-tuning, remains an open question.

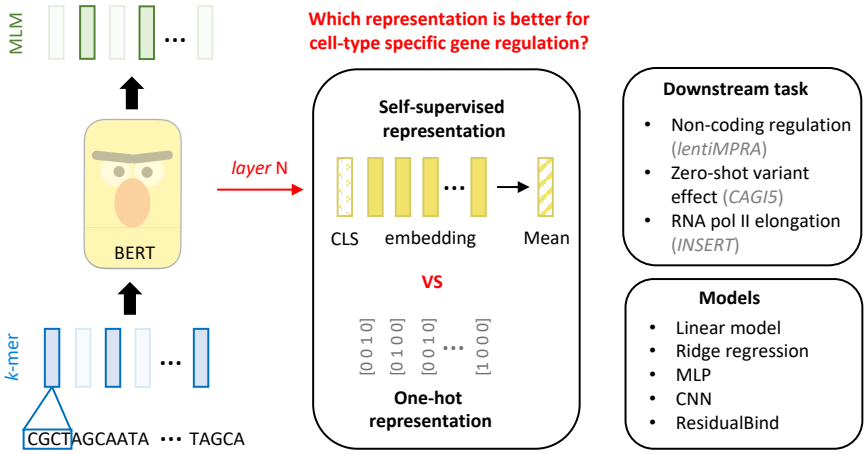

Figure 1. Experiment overview. Comparison of LLM embeddings versus one-hot representations for functional genomics prediction tasks.

Here we assess the potential of genomic LLMs as foundation models, with a particular emphasis on comparing the informativeness of their learned sequence embeddings with traditional one-hot representations. We performed three systematic experiments to gauge the predictive abilities of LLM embeddings: 1) functional characterization of regulatory elements, 2) zero-shot non-coding variant effect generalization, and 3) RNA pol II elongation potential. Our results indicate that present LLMs exhibit significantly lower performance than supervised models in human functional genomics prediction tasks and zero-shot variant effect prediction, thus demonstrating their lack of versatility.

2. Experiment 1: functional characterization of cis-regulatory elements

Cis-regulatory elements (CREs) are vital regions in the genome that are hot-spots for transcription factor binding and this, in turn, plays a crucial role in transcriptional regulation. Given their ubiquitous presence throughout the genome, one would anticipate that LLMs trained on genome-wide data should learn essential characteristics of CREs. Consequently, LLMs are expected to encode representations that are informative for predicting CRE activity, but the extent to which they do remains unclear.

To assess the effectiveness of LLM embeddings in predicting cell-type specific CRE activity, we utilized experimental lentiMPRA data for K562 and HepG2 cell lines (Agarwal et al., 2023). Each lentiMPRA dataset consists of over 100k sequences each 230 nt long paired with a target scalar value that represents the transcriptional activity of the sequence (see Appendix A). In this initial experiment, we evaluated the performance of two models: the Nucleotide Transformer and a custom Genomic Pre-trained Network (GPN). The Nucleotide Transformer is a 2.5 billion parameter BERT model that was trained on 3,202 diverse human genomes from the 1000 Genomes Project and exhibited the best performance in variant effect performance in humans (Dalla-Torre et al., 2023). The GPN model is a convolution-based model that was originally trained on the Arabidopsis genome (Benegas et al., 2022). We trained a custom GPN model on the human reference genome following a similar procedure as the original study (see Appendix A for details).

As there is no principled strategy to harness the predictive power of LLM embeddings, we trained a linear regression, a ridge regression, and a multi-layer perceptron (MLP) using either the classification(CLS) token or the mean embedding generated by each layer of Nucleotide Transformer. This involved acquiring the representations for each training sequence to be used as input data in lieu of the standard one-hot sequence. Additionally, we trained a basic convolutional neural network (CNN) that incorporated the complete embedding as input (see Appendix A). As the GPN model follows a pure convolution-based structure, we only considered the penultimate embeddings as input to basic CNN. To benchmark the performance against models trained on one-hot data, we trained three additional models: 1) the basic CNN trained directly on one-hot data, 2) a re-impelmented MPRAnn originally designed for the lentiMPRAs dataset (Agarwal et al., 2023), and 3) a Residualbind model (Koo et al., 2021) to represent performance achievable by a more sophisticated model.

Our results demonstrate that models trained on LLM embeddings generally under-perform compared to one-hot based models (Fig.2a). For Nucleotide Transformer, we observed variation in performance across layers, suggesting that the penultimate layer is not necessarily optimal for feature extraction. Moreover, analyzing the full embeddings led to

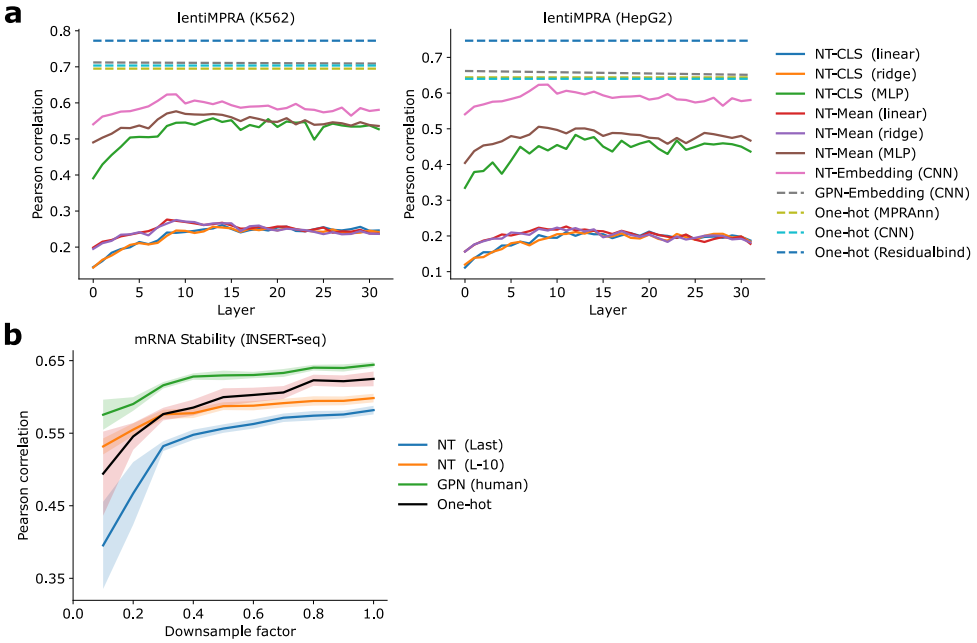

Figure 2. Test set performance on (a) LentiMPRA and (b) INSERT-seq for downstream models trained using different input features generated by nucleotide Transformer (NT) and GPN. Downstream model used is shown in parenthesis.

better overall performance, albeit with GPN achieving comparable performance as one-hot models. The performance gap indicates that LLM embeddings are not more informative than one-hot sequences for cell-type specific gene regulation, in fact, BERT-based LLMs seem to lose valuable information.

3. Experiment 2: zero-shot single-nucleotide variant effect generalization

A major use case of highly accurate sequence-function models is in helping to understand how genomic variation leads to functional changes. These models are not directly trained on the variants, making it a case of zero-shot variant effect generalization. Previous studies have indicated that LLMs have a certain degree of predictive capability for variant effects (Benegas et al., 2022; Dalla-Torre et al., 2023). However, it is not intuitive how LLMs could yield good zero-shot predictions of cell type specific variant effects since they are trained without any cell-type information.

To assess zero-shot variant effect generalization, we focused on single-nucleotide variants of regulatory elements characterized through a saturation mutagenesis MPRA as part of the Critical Assessment of Genome Interpretation 5 (CAGI5) (Shigaki et al., 2019), considering only cell lines that match the lentiMPRA dataset: 1 MPRA experiment for K562 and 3 experiments for HepG2.

To evaluate the single-nucleotide variant effect predictive capabilities of each LLM, we followed the procedures outlined in their respective studies (see Appendix A). Our study extends to all Nucleotide Transformer models, which vary in size and training data scale. This includes a 2.5 billion parameter BERT model trained on 3,202 diverse human genomes as well as 850 genomes from various species. We also examined a custom GPN model solely trained on the human reference genome. Additionally, we assessed the variant effect predictions of the best-performing embedding-based models for GPN and Nucleotide Transformer on the lentiMPRA dataset. For comparison, we also included one-hot models trained on the lentiMPRA dataset as a baseline.

Notably, our findings revealed that all LLMs exhibited poor performance in zero-shot variant effect generalization (Table 1). This suggests that, unlike the original studies, LLMs may not be as effective in capturing variant effects across all genomic regions, particularly in non-coding regions where they struggle to capture cell-type specific variant effects.

Interestingly, CNNs trained on lentiMPRA data using LLM embeddings demonstrated improved variant effect generalization, with GPN approaching performance levels surpassing certain CNNs trained directly on one-hot data. Nonetheless, the ResidualBind model (Koo et al., 2021), trained on one-hot lentiMPRA data, and the original Enformer model (Avsec et al., 2021), trained on a multitude of epigenomic tracks, yielded substantially better performance. Thus, specialized supervised models might serve as a better foundation model compared to current LLMs.

Table 1. Zero-shot variant effect generalization on CAGI5 dataset. The values represent the average Pearson correlation of predictions with experimental values (1 experiment for K562; 3 for HepG2).

INPUT	MODEL	HEPG2	K562
ZERO-SHOT	NT (2B51000G)	-0.0167	-0.0234
	NT (2B5SPECIES)	-0.0003	-0.0625
	NT (500MHUMAN)	-0.0486	0.0324
	NT (500M1000G)	0.0270	0.0539
	GPN (HUMAN)	-0.0034	0.0390
EMBEDDING	GPN (CNN)	0.3774	0.4567
	NT (CNN)	0.1798	0.3875
ONE-HOT	CNN (LENTIMPRA)	0.3128	0.4257
	RESIDUALBIND (LENTIMPRA)	0.4860	0.5510
	MPRANN (LENTIMPRA)	0.3014	0.3686
	ENFORMER (DNASE)	0.5104	0.6845

4. Experiment 3: RNA elongation potential

While LLMs exhibited subpar performance in non-coding regions, they have shown some aptitude for learning gene definitions and splice sites. Hence we now focus on predicting RNA pol II elongation potential measured via INtegrated Sequences on Expression of RNA and Translation using high-throughput sequencing (INSERT-seq) (Vlaming et al., 2022). This dataset is relatively small, consisting of 10,774 sequences each 173 nt long, and primarily encompasses gene elements such as 3' untranslated regions. Given the modest size of the dataset, training large models in a supervised manner can easily result in overfitting.

Pretrained LLMs have the potential to acquire a diverse range of generalizable features that can be applied to various downstream tasks. In cases where the dataset is large, such as the lentiMPRA dataset, which consists of over 100k sequences, standard CNNs can directly learn predictive features from one-hot representations, rendering pretrained LLMs less advantageous. To investigate the validity of this assumption with smaller datasets, we conducted systematic downsampling of the INSERT-seq dataset in incremental steps of 10%. For each down-sampled dataset, we systematically trained identical CNNs but using different input representations: one-hot encoding, embeddings from GPN, and embeddings from Nucleotide Transformer (trained on 1000 Genomes Project), both from the penultimate layer and layer 10, the top performer in the lentiMPRA study.

Interestingly, GPN displayed superior performance across all downsampling factors, whereas Nucleotide Transformer models exhibited lower performance compared to one-hot based CNN (Fig. 2b). The improved performance by GPN suggests that LLMs can specialize in some genomic regions better than others. In this dataset, capturing 5' splice sites is a critical feature (Vlaming et al., 2022). Thus, understanding what features LLMs learn well can help to identify suitable downstream tasks for which they can thrive.

5. Discussion

Here we assessed the predictive capabilities of LLM representations trained on whole genomes via masked language modeling. Our findings reveal that LLMs generally fail to capture crucial features of cell-type specific cis-regulatory activity in humans and may even result in the loss of valuable information. Furthermore, their zero-shot prediction performance significantly lags behind that of supervised models. These observations indicate that despite the achievements of LLMs in various domains, they still have a long way to go before unlocking their potential in genomics.

The success of LLMs in defining gene features and capturing certain motifs in previous studies raises a crucial question: how can we reconcile these achievements with their limited performance in capturing cell-type specific information? Although the human genome is a blueprint for all cells in the body, each cell possesses a unique regulatory code, which is projected onto a single DNA sequence. This inherent complexity makes it challenging to disentangle cell-type specific information through a masked language modeling objective. Nevertheless, gene definition remains relatively consistent across cell types. Thus, basic tasks requiring RNA features may bode well for LLMs. Also, LLMs have demonstrated considerable success in simpler organisms like yeast, bacteria, and Arabidopsis. However, our findings indicate that the extension of these models, such as GPN, to human genomes does not yield comparable results.

One question that arises is how to reconcile the performance gap of LLMs with previous studies that have shown comparable performance with supervised models (Ji et al., 2021; Dalla-Torre et al., 2023). In those studies, LLMs were fine-tuned for a specific downstream task, meaning that the pretraining strategy was simply an initialization. While this can still be effective, it remains unclear whether it provides a more effective initialization or transfer learning strategy compared to pretrained models that are trained in a multi-task setting with supervised learning. Without fine-tuning, it is not clear what specific prediction tasks LLMs excel at.

Although LLMs have demonstrated promise across diverse domains, their potential in regulatory genomics is still uncertain. A crucial factor that will shape their applicability is model interpretability (Toneyan et al., 2022). Understanding what features are encoded in the learned representations will provide valuable insights into the specific tasks for which they are most suitable. Moreover, rethinking masked language modeling in non-coding regions may catalyze a path forward to thinking about how to deal with the seemingly dense regions of random DNA with low-order statistical properties, i.e. dinucleotide frequencies, and motifs that carry high information content but are sparsely located.

References

Agarwal, V., Inoue, F., Schubach, M., Martin, B., Dash, P., Zhang, Z., Sohota, A., Noble, W., Yardimci, G., Kircher, M., et al. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv*, pp. 2023–03, 2023.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Benegas, G., Batra, S. S., and Song, Y. S. Dna language models are powerful zero-shot predictors of non-coding variant effects. *bioRxiv*, pp. 2022–08, 2022.

Bhattacharya, N., Thomas, N., Rao, R., Dauparas, J., Koo, P. K., Baker, D., Song, Y. S., and Ovchinnikov, S. Interpreting potts and transformer protein models through the lens of simplified attention. In *Pacific Symposium on Biocomputing 2022*, pp. 34–45. World Scientific, 2021.

Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., Shen, T., et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pp. 2022–08, 2022.

Chen, K., Zhou, Y., Ding, M., Wang, Y., Ren, Z., and Yang, Y. Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *bioRxiv*, pp. 2023–01, 2023.

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and machine intelligence*, 44(10):7112–7127, 2021.

Ferruz, N. and Höcker, B. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6): 521–532, 2022.

Hie, B. L., Yang, K. K., and Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285, 2022.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P., and Paul, S. B. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Computational Biology*, 17 (5):e1008925, 2021.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Lupo, U., Sgarbossa, D., and Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nature Communications*, 13 (1):6298, 2022.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Mo, S., Fu, X., Hong, C., Chen, Y., Zheng, Y., Tang, X., Shen, Z., Xing, E. P., and Lan, Y. Multi-modal self-supervised pre-training for regulatory genome across cell types. *arXiv preprint arXiv:2110.05231*, 2021.

OpenAI. Gpt-4 technical report, 2023.

Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. doi: 10.1101/2020.12.15.422761. URL https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*,

2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

Shigaki, D., Adato, O., Adhikari, A. N., Dong, S., Hawkins-Hooker, A., Inoue, F., Juven-Gershon, T., Kenlay, H., Martin, B., Patra, A., et al. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Human Mutation*, 40(9):1280–1291, 2019.

Toneyan, S., Tang, Z., and Koo, P. K. Evaluating deep learning for predicting epigenomic profiles. *Nature Machine Intelligence*, pp. 1–13, 2022.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. Bertology meets biology: interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Vlaming, H., Mimoso, C. A., Field, A. R., Martin, B. J., and Adelman, K. Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of rna polymerase ii elongation potential. *Nature Structural & Molecular Biology*, 29(6):613–620, 2022.

Wang, W., Peng, Z., and Yang, J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nature Computational Science*, 2(12): 804–814, 2022.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

Yamada, K. and Hamada, M. Prediction of rna–protein interactions using a nucleotide language model. *Bioinformatics Advances*, 2(1):vbac023, 2022.

Yang, M., Huang, L., Huang, H., Tang, H., Zhang, N., Yang, H., Wu, J., and Mu, F. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Research*, 50(14):e81–e81, 2022.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

A. Methods

A.1. Experiment 1

lentiMPRA. The lentiMPRA dataset was generated using lentivirus-based massively parallel reporter assays to measure the regulatory activity of candidate CREs in K562 and HepG2 cell lines (Agarwal et al., 2023). The HepG2 library consists of 139,984 sequences, each 230 nucleotides long, and the K562 library contains 226,253 sequences. We split the dataset into train, valid and test sets randomly, the same split was applied to all experiments using this dataset.

GPN. The original GPN model is a convolutional neural network trained on Arabidopsis genome sequences via masked language modeling with an input size of 512 nucleotides (Benegas et al., 2022). It consists of 25 convolutional blocks, where each convolutional block includes a dilated convolutional layer followed by a feed-forward layer, connected by intermediate residual connections and layer normalization. The dilation rate for each convolutional layer was increased exponentially from 1 up to 32 and then cycled. The embedding dimension was kept fixed at 512 throughout the layers.

For our custom GPN (human) model, we created training datasets using the human reference genome (hg38). The genome was split into contigs and filtered for a minimum length of 512 nucleotides, with chromosome 8 held out as test set. We trained the GPN structure on the human genome dataset using the same set of training hyper-parameters and masked language modeling task as the original study (Benegas et al., 2022).

MLP. A multi-layer perceptron model was used to train on CLS token embeddings or the average embedding across sequences for Nucleotide Transformer models. The model is constructed by two fully connected blocks. The first block includes a fully-connected layer with 512 units and ReLU ativation, followed by batch normalization and a dropout rate of 0.5. The second block consists of a fully-connected layer with 256 units and the same activation, batch normalization, and dropout layers. The model was trained on lentiMPRA dataset with Adam optimizer, learning rate of 0.0001, learning rate decay patience of 5 epochs with a decay factor of 0.2, and early stopping patience of 10 epochs.

Basic CNN. We designed a baseline CNN model with the following structure:

1. batch normalization
2. convolution (196 filters, size 1)
3. convolution (196 filters, size 7, batch norm, exponential)
   dropout (0.2)
   max-pooling (size 5)
4. convolution (256 filters, size 7, batch norm, relu)
   dropout (0.2)
   max-pooling (size 4)
5. fully-connected (512 units, batch norm, relu)
   dropout (0.5)
6. fully-connected (256 units, batch norm, relu)
   dropout (0.5)
7. output layer (1 unit, linear)

We trained this CNN model with Adam optimizer, mean squared error loss function, learning rate of 0.0001 with a learning rate decay patience of 5 epochs with a decay rate of 0.2, and early stopping with patience of 10 epochs for both one-hot sequence and language model embedding-based training on the lentiMPRA dataset.

A.2. Experiment 2

CAGI dataset. The CAGI5 challenge dataset was used to evaluate the performance of the models on zero-shot single-nucleotide variant effect generalization (Shigaki et al., 2019). Among all the experiments, we only included the ones executed in HepG2: LDLT, SORT1, F9; and K562: PKLR. We extracted 230 nucleotide sequences from the reference genome centered on each regulatory region of interest. Alternative alleles are then substituted correspondingly to construct the CAGI test sequences. Pearson correlation was calculated between the model prediction scores and experimentally measured effect size per experiment. For HepG2 performances, we report the average Pearson's r across the three experiments.

Zero-shot variant effect prediction methods. For Nucleotide Transformer models, we derived the zero-shot predictions using cosine similarity as suggested in the original study (Dalla-Torre et al., 2023). For each variant, we passed the sequences with the centered reference allele and the alternative allele through the model to extract embeddings. The cosine similarity between the two sequence embeddings was calculated and used as the zero-shot score. A negative correlation is expected between the score and effect size. Since this distance based zero-shot score only reflects the magnitude, not the direction, of function change, we calculated the Pearson correlation using the absolute value of the effect size.

For the GPN models, we input sequences with the center variant loci masked, following similar procedure as the original study (Benegas et al., 2022). From the predicted allele probabilities for the masked loci, we calculate the zero-shot prediction score as the log-likelihood ratio between the alternate and reference alleles. Again, since the likelihood ratio doesn't reflect the direction of function change associated with the variants, we calculated the correlation score using the absolute value of effect size.

Finally, for the embedding-based and one-hot based models, we used the difference in predictions between the alternative and reference allele sequence as the zero-shot prediction score. For Enformer, we use the cell-type agnostic approach of averaging the effect size across all DNase-seq tracks. To reduce predictions to scalars, we summed across the profile predictions.

A.3. Experiment 3

INSERT-seq. The INSERT-seq was executed in mouse embryonic stem cells investigated the impact of transcribed sequences on the RNA polymerase II enlongation potential and expression. We used the 173 nucleotides long insert sequence as the model input to predict the totalRNA output, which measures the relative abundance in RNA relative to genomic DNA.

Model Structure. For the RNA pol II elongation potential dataset, we developed a residual convolutional network structure and used it for all embedding and one-hot-based models. The model was trained using mean square error loss function, Adam optimizer, learning rate of 0.0001, learning rate decay patience of 5 epochs with a decay rate of 0.2, and early stopping patience of 10 epochs.

1. convolution(48 filters, size 1)
2. convolution (96 filters, size 19, batch norm, exponential)
   dropout (0.1)
3. dilated residual block (96 filters, size 3, relu)
   convolution (batch norm)
   dropout (0.1)
   convolution (batch norm, dilation rate 2)
   dropout (0.1)
   convolution (batch norm, dilation rate 4)
   residual connection to block input
   relu activation
   max-pooling (size 10)
   dropout(0.1)
4. convolution (128 filters, size 7, batch norm, relu)
   global average-pooling
   dropout (0.1)
5. fully-connected layer (128 units, relu)
   dropout (0.5)
   6. output layer (1 unit, linear)