Table S1. Summary of existing gLM architecture and training choices

Name	Base Model	Training Data	Pre-training Task	Tokenization	Finetuned for evaluations
GPN [20]	CNN	8 plant genomes	MLM	Base-res	No
GPN-MSA [34]	BERT	Conserved regions in 100 vertebrates MSA (PhastCons)	MLM	Base-res	No
Nucleotide Transformer [23]	BERT	1000 Genomes, Human genome and multi-species genomes	MLM	Non-overlapping k-mer	Yes
DNABERT [24]	BERT	Human genome	MLM	Overlapping k-mer	Yes
DNABERT2 [26]	BERT-Flash attention	Human genome and multi-species genomes	MLM	Byte-pair encoding	Yes
BigBird [32]	Sparse attention	Human genome	MLM+CLM	Byte-pair encoding	Yes
hyenaDNA [21]	Hyena	Human genome	CLM	Base-res	Yes
OmniNA [31]	LLaMA	Multi-species and text annotations	CLM (multimodal)	Byte-pair encoding	Yes
GENA-LM [33]	BERT	T2T, 1000 Genomes, and multi-species genomes	MLM	Byte-pair encoding	Yes
LOGO [38]	ALBERT	Human genome	MLM	Non-overlapping k-mer	Yes
DNA-GPT [25]	GPT	Genome from 9 species	CLM, GC-content, and sequence order	Non-overlapping k-mer	Yes
GROVER [27]	BERT	Human genome	MLM	Byte-pair encoding	Yes
UTR-LM [29]	BERT	5' UTR sequences (Ensembl)	MLM and supervised (secondary structure and minimum free energy)	Base-res	Yes
SpliceBERT [30]	BERT	pre-mRNA from 72 vertebrates	MLM	Base-res	Yes
RegLM [22]	Hyena	Yeast promoter sequences	CLM	Base-res	Yes
RNA-FM [39]	BERT	RNAcentral (promoters)	MLM	Base-res	Yes
Species-aware DNA LM [28]	BERT	UTR of 1500 yeast genomes	MLM	Base-res	No
FloraBERT [41]	BERT	Multi-species plant genome	MLM	Byte-pair encoding	Yes
RandomMask [42]	BERT	Human genome	MLM (expanding size)	Overlapping k-mer	Yes
GenSLMs[40]	BERT	BV-BRC dataset	Contrastive loss	Codon	Yes
Evo [85]	StripedHyena	prokaryotic genomes and plasmids	CLM	Base-res	Both
Mamba [43]	Mamba	human genome	CLM	Base-res	Both
Caduceus [46]	Bi-directional Mamba	prokaryotic genomes and plasmids	MLM	Base-res	Yes
PlantCaduceus [47]	SSM (Bi-directional Mamba)	16 Angiosperm genomes	MLM	Nucleotide	Yes
AgroNT [48]	Transformer (BERT)	48 plant genomes	MLM	Non-overlapping k-mer	Yes
Epigenomic BERT [49]	Transformer (BERT)	Human genome, epigenetic state information (IDEAS)	MLM	Non-overlapping k-mer	Yes
RiNALMo [60]	Tansformer (BERT)	non-coding RNA	MLM	Nucleotide	Both

Supplementary Table S2. Performance comparison on cell-type-specific regulatory activity prediction tasks from lentiMPRA data. Predictive performance using a baseline CNN trained using different gLM embedding inputs, one-hot sequences, or supervised embeddings from Sei. MPRAnn and ResNet represent the performance of more sophisticated models that are trained using one-hot sequences. The performance of NT, Hyena, and DNABERT2 fine-tuned directly on the lentiMPRA data is also shown. Note that NT fine-tuning is different from the self-supervised embedding due to limited resources for fine-tuning. Enformer* represents a LASSO regression – i.e., linear probing based on Enformer's predictions, instead of a downstream CNN trained on the full embeddings like SEI.

Training Task	Model	HepG2	K562
Self-Supervised Embedding	NT (2B51000G)	0.473	0.549
Self-Supervised Embedding	GPN	0.637	0.678
Self-Supervised Embedding	HyenaDNA	0.556	0.650
Self-Supervised Embedding	DNABERT2	0.434	0.568
Self-Supervised Embedding	Random	0.335	0.435
Fine-Tuned	NT (500M1000G)	0.425	0.337
Fine-Tuned	HyenaDNA	0.772	0.777
Fine-Tuned	DNABERT2	0.782	0.781
Supervised Embedding	SEI	0.735	0.751
Supervised Embedding	Enformer*	0.728	0.731
One-hot	One-hot	0.638	0.691
One-hot	ResNet	0.753	0.789
One-hot	MPRAnn	0.713	0.742

Table S3. Expanded zero-shot variant effect generalization on CAGI5 dataset. The values represent the Pearson correlation between the variant effect predictions with experimental saturation mutagenesis values of a given CRE measured via MPRAs. Values are reported for a single CRE experiment for K562 and the average of three CRE experiments for HepG2. Notably, this table includes the performance based on NT, HyenaDNA and DNABERT2 when fine-tuned on lentiMPRA data.

Training Task	Model	Variant Effect Prediction	HepG2	K562
Self-Supervised Pre-Training	NT (2B51000G)	cosine distance	0.125	0.007
Self-Supervised Pre-Training	NT (2B5Species)	cosine distance	0.112	0.135
Self-Supervised Pre-Training	NT (500MHuman)	cosine distance	0.020	0.088
Self-Supervised Pre-Training	NT (500M1000G)	cosine distance	0.041	0.068
Self-Supervised Pre-Training	GPN (Human)	log2-ratio	0.002	0.037
Self-Supervised Pre-Training	HyenaDNA	cosine distance	0.064	0.021
Fine-Tuned	NT (500M1000G)	difference from wild type	0.155	0.104
Fine-Tuned	HyenaDNA	difference from wild type	0.286	0.315
Fine-Tuned	DNABERT2	difference from wild type	0.357	0.430
LentiMPRA-Embedding	CNN-GPN	difference from wild type	0.377	0.457
LentiMPRA-Embedding	CNN-NT (2B51000G)	difference from wild type	0.137	0.240
LentiMPRA-Embedding	CNN-SEI	difference from wild type	0.559	0.719
LentiMPRA-one-hot	CNN	difference from wild type	0.313	0.426
LentiMPRA-one-hot	Residualbind	difference from wild type	0.486	0.551
LentiMPRA-one-hot	MPRAnn	difference from wild type	0.301	0.369
Supervised one-hot	SEI	cosine distance	0.545	0.641
Supervised one-hot	Enformer (DNase)	difference from wild type	0.510	0.685

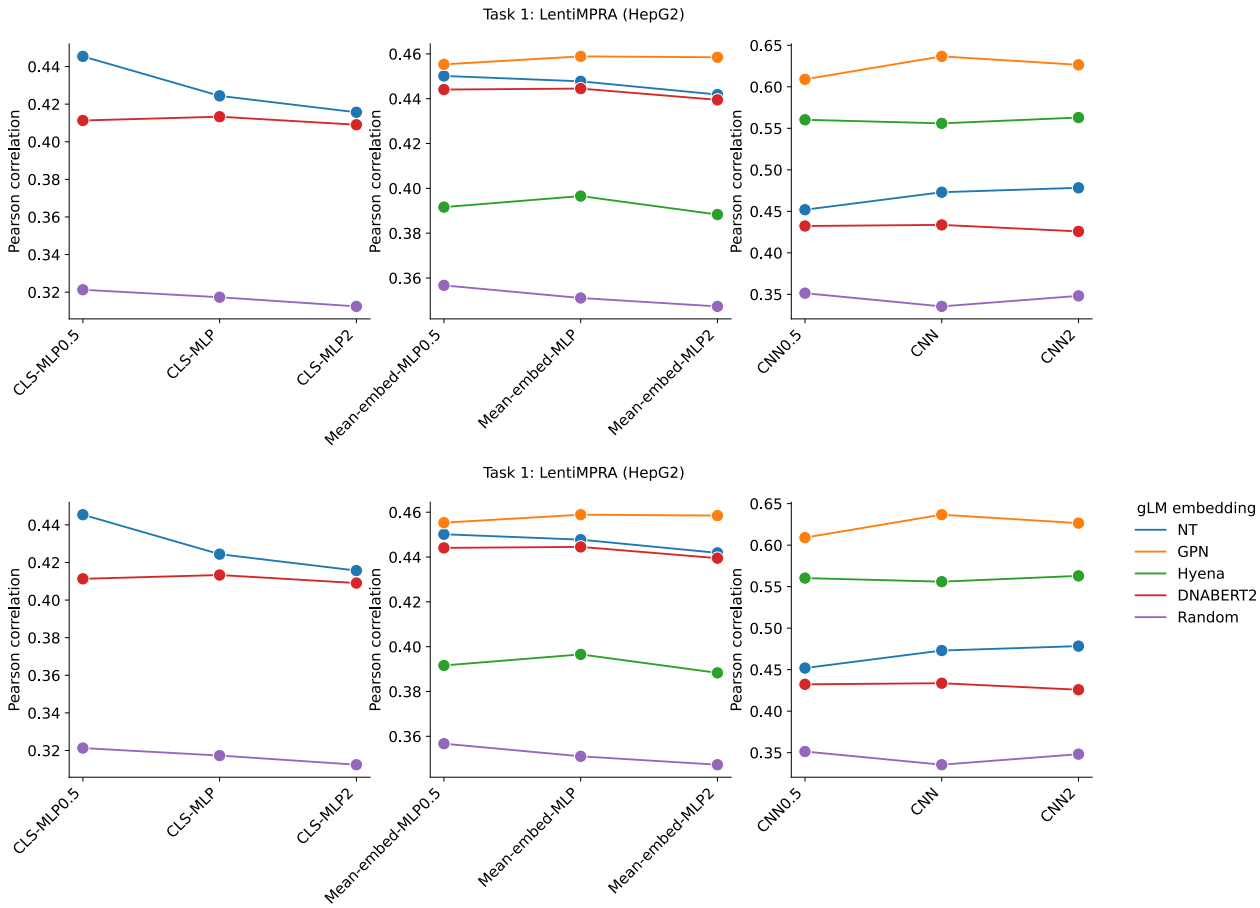

Fig S1. Comparison of performance for various hyperparameter choices for downstream models on lentiMPRA data. Four types of downstream models, including ridge regression (Ridge), multi-layer perceptron (MLP), and convolutional neural network (CNN), were trained for each gLM. The numbers after the model names (i.e., 0.5 or 2) represent the scaling factor for the number of parameters in each hidden layer relative to the original model structure. These results show that, up to a factor of 2, the performance stays roughly similar.

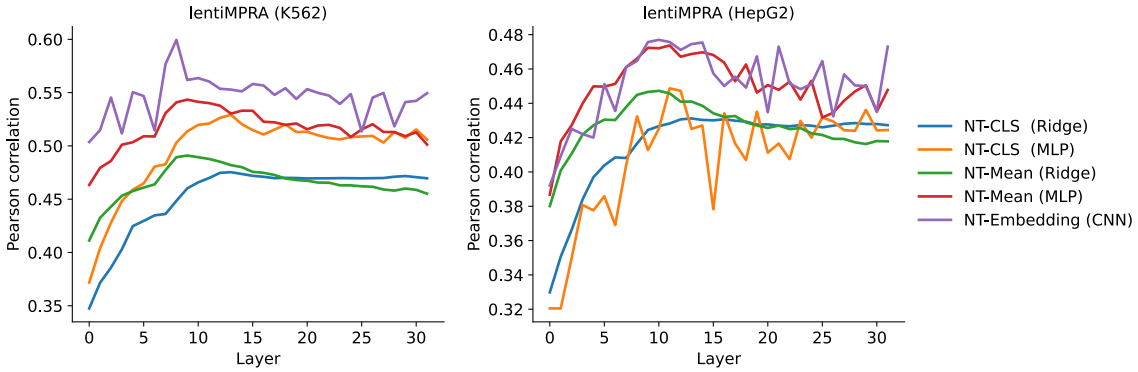

Fig S2. Layer-wise performance of Nucleotide-Transformer on the lentiMPRA dataset. Test performance of various machine learning models trained using embeddings from different layers of Nucleotide-Transformer. Embeddings include the CLS token, mean embedding (Mean), and the full embedding (Embedding). Machine learning models include ridge regression (ridge), multi-layer perceptron (MLP) and a convolutional neural network (CNN).

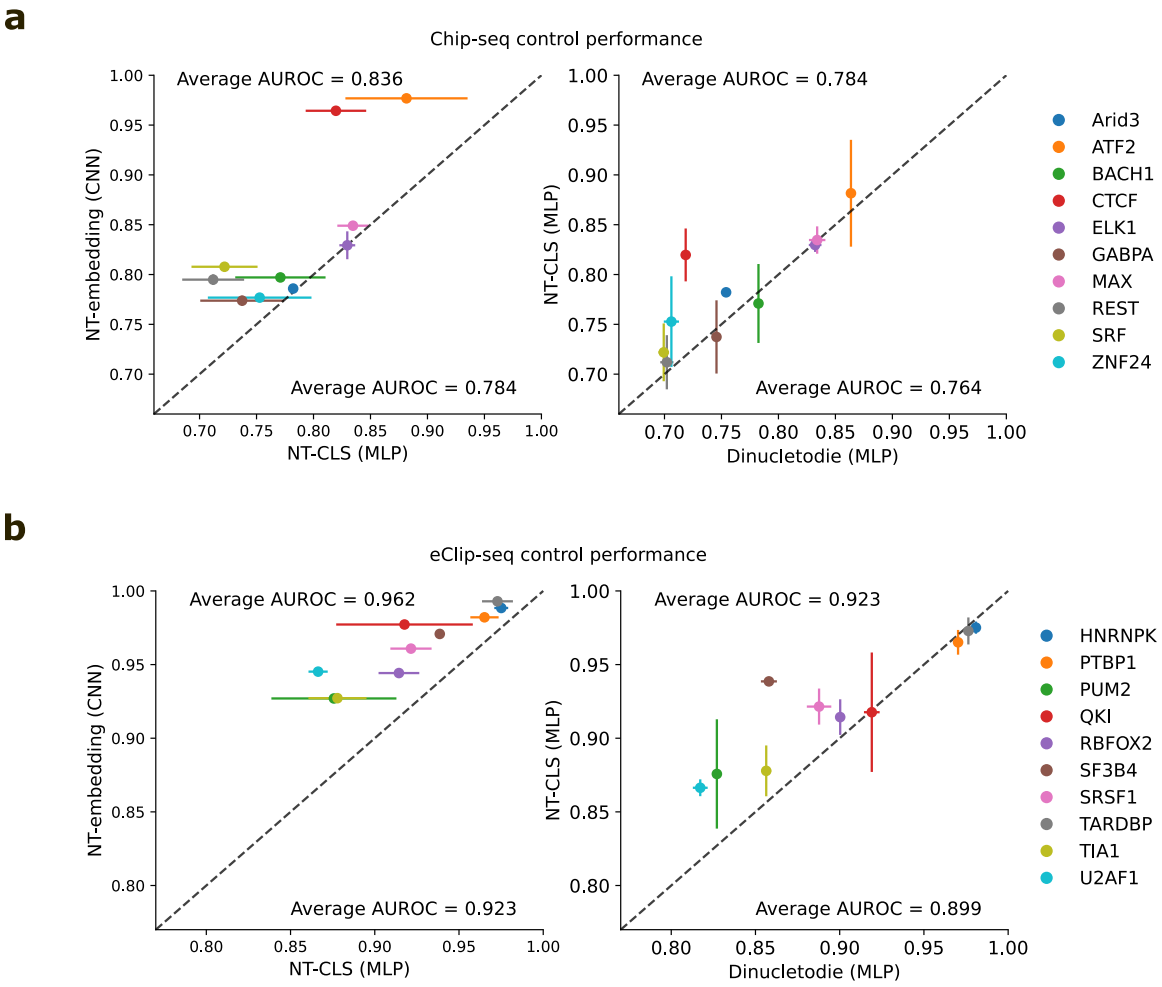

Fig S3. Control experiments with different embeddings. Performance comparison between a CNN trained using full embeddings of the penultimate layer from Nucleotide-Transformer, an MLP trained using Nucleotide-Transformer's CLS token, and an MLP trained using dinucleotide frequencies of the sequence on (**a**) ChIP-seq data and (**b**) eCLIP-seq data. Performance is measured by the average area-under the receiver-operating characteristic curve (AUROC) and error bars represent the standard deviation of the mean across 5 different random initializations. Text values represent the average AUROC across all ChIP-seq or CLIP-seq datasets.

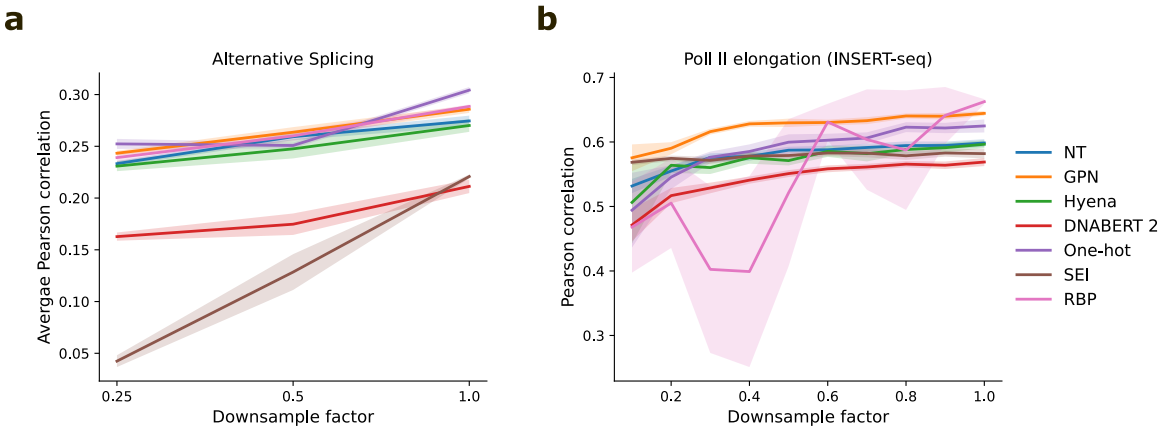

Fig S4. Down-sampling performance on RNA regulation tasks. Average performance of machine learning models on (**a**) alternative splicing, task 4, and (**b**) RNA Pol II elongation potential, task 5, down-sampled by various factors. Shaded region represents standard deviation of the mean across 5 different random initializations.

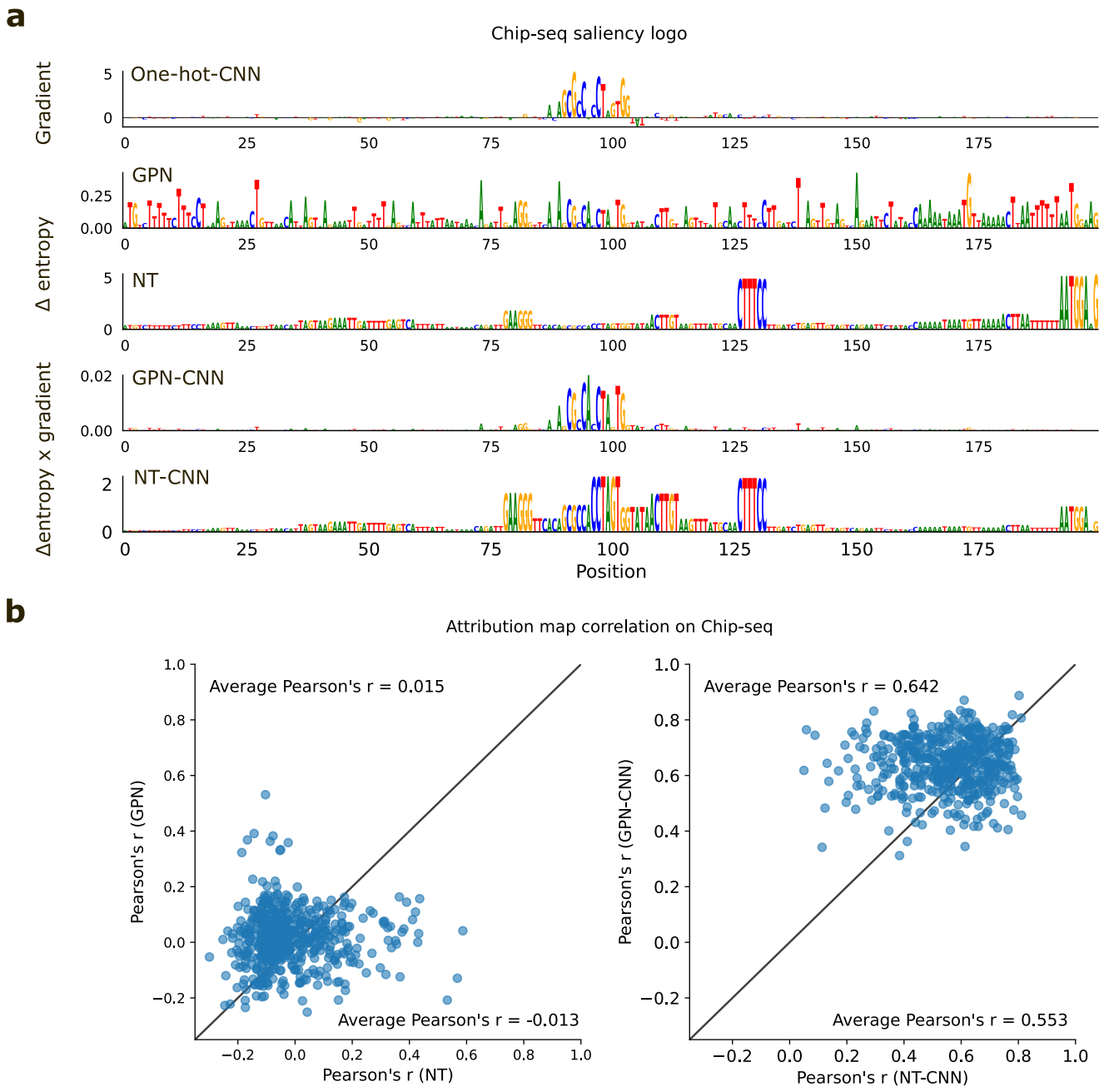

Fig S5. Attribution analysis comparison for sequences from CTCF ChIP-seq data. **a**, Representative example of attribution maps for a CTCF binding sequence. Attribution maps include (top to bottom): the gradient-times-input of a one-hot-trained CNN; the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained GPN; the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained Nucleotide-Transformer; the gradient of a CNN-trained using GPN embeddings multiplied by the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained GPN; and the gradient of a CNN-trained using Nucleotide-Transformer embeddings multiplied by the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained Nucleotide-Transformer. **b**, Scatter plot comparison of the attribution map correlations for different pre-trained gLMs (left) and CNNs trained using gLM embeddings (right). Attribution map correlations reflect the Pearson correlation coefficient between the attribution map generated by the gLM-based attribution method with the Saliency Map generated by a one-hot-trained CNN. Each dot represents a different sequence in the CTCF ChIP-seq dataset (N=500).

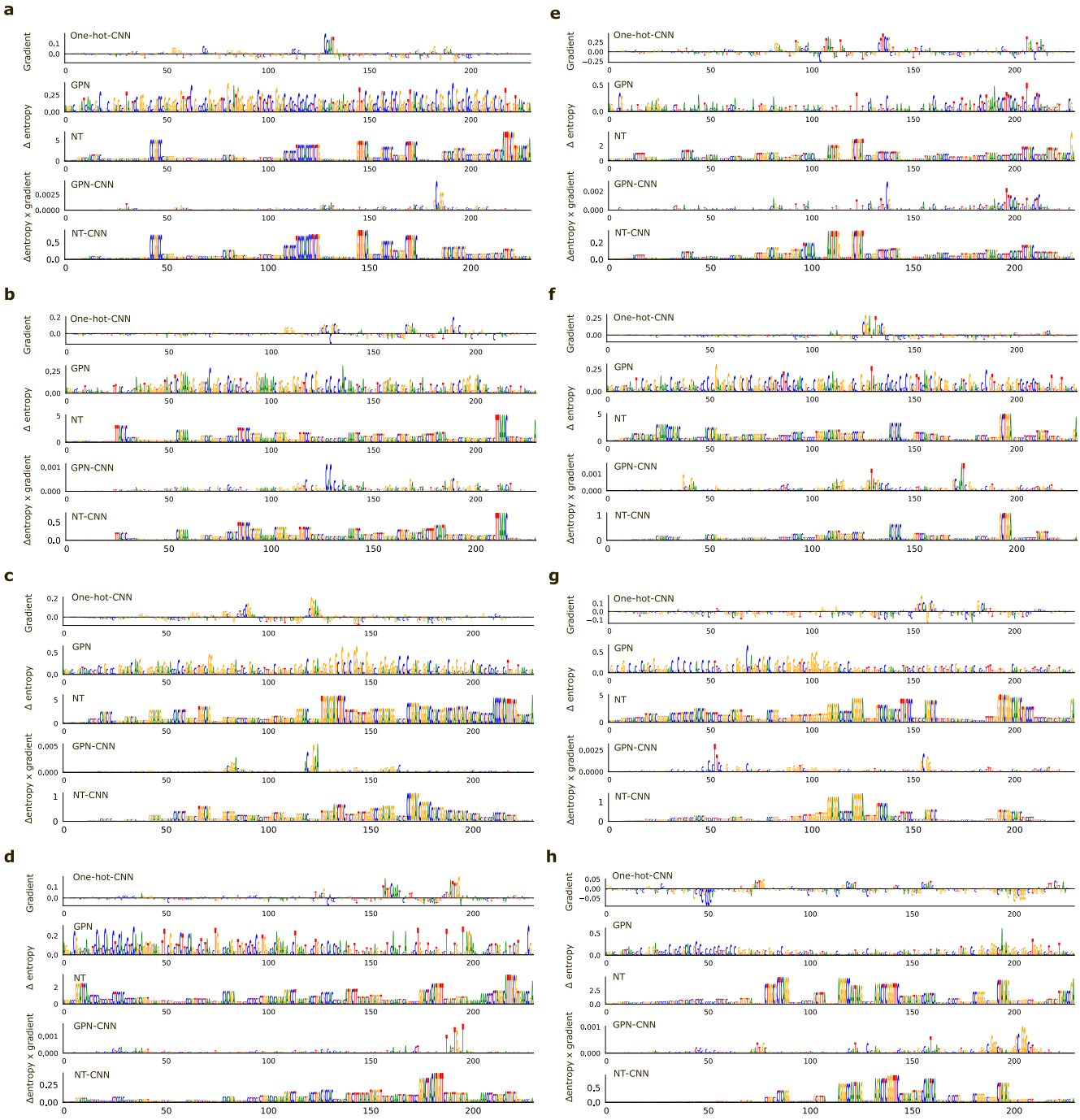

Fig S6. Representative examples of attribution maps for sequences from the lentiMPRA dataset. In each panel, attribution maps are shown for different sequences in order of (top to bottom): the gradient-times-input of a one-hot-trained CNN; the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained GPN; the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained Nucleotide-Transformer; the gradient of a CNN-trained using GPN embeddings multiplied by the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained GPN; and the gradient of a CNN-trained using Nucleotide-Transformer embeddings multiplied by the delta entropy of predicted nucleotides via single-nucleotide masking from a pre-trained Nucleotide-Transformer.