BEND: BENCHMARKING DNA LANGUAGE MODELS ON BIOLOGICALLY MEANINGFUL TASKS

Frederikke I. Marin[1,2]*, Felix Teufel[3,4]*, Marc Horlacher[5], Dennis Madsen[3], Dennis Pultz[1], Ole Winther[4,6], Wouter Boomsma[2]
[1]Bioinformatics & Design, Enzyme Research, Novozymes A/S
[2]Department of Computer Science, University of Copenhagen
[3]Digital Science & Innovation, Novo Nordisk A/S
[4]Department of Biology, University of Copenhagen
[5]Computational Health Center, Helmholtz Center Munich
[6]DTU Compute, Technical University of Denmark

ABSTRACT

The genome sequence contains the blueprint for governing cellular processes. While the availability of genomes has vastly increased over the last decades, experimental annotation of the various functional, non-coding and regulatory elements encoded in the DNA sequence remains both expensive and challenging. This has sparked interest in unsupervised language modeling of genomic DNA, a paradigm that has seen great success for protein sequence data. Although various DNA language models have been proposed, evaluation tasks often differ between individual works, and might not fully recapitulate the fundamental challenges of genome annotation, including the length, scale and sparsity of the data. In this study, we introduce **BEND**, a **Ben**chmark for **D**NA language models, featuring a collection of realistic and biologically meaningful downstream tasks defined on the human genome. We find that embeddings from current DNA LMs can approach performance of expert methods on some tasks, but only capture limited information about long-range features. BEND is available at https://github.com/frederikkemarin/BEND.

1 INTRODUCTION

Within the last two decades, the cost of sequencing whole genomes has significantly decreased, having led to an extraordinary wealth of genomic DNA sequences. This has improved our understanding of genetic variation among human genomes and introduced genomes of hitherto understudied species. However, the generation of experimental data to annotate and understand these genomic sequences has not kept pace.

At the same time, Natural Language Processing (NLP) has demonstrated the power of large-scale models to capture signals in sequences by masking and reconstructing them in a self-supervised manner. The success of masked language modeling (MLM) has extended to the biological domain Rao et al. (2019); Bepler & Berger (2021); Madani et al. (2023); Rives et al. (2019), with protein language models (pLMs) now being widely used for prediction tasks on protein sequences. The availability of unlabeled genomic sequences and, in many organisms, limited labeled data appear to make language modeling a natural fit for DNA. DNA language models (LMs) have indeed started to emerge, but while the paradigms of NLP have been easy to transfer to proteins, the same may not be true for modeling genomes, as they present unique challenges: signals can have an extremely long length range, high-signal regions are sparse, and even in those regions the density of signal is lower compared to proteins.

In this paper, we present BEND, a **Ben**chmark for **D**NA Language Models, a collection of realistic and biologically meaningful downstream tasks. BEND aims to provide a standardized set of tasks that measure the ability of LMs to capture the intricacies of genomic data, and to help advance this nascent field. In summary, BEND contributes:

*Equal contribution

- **Seven curated tasks and datasets**, probing understanding of different DNA functional elements over a variety of length scales.
- **Experiments covering DNA LMs from six different sources.** To our knowledge, this represents first evaluation of all publicly available self-supervised DNA LMs suitable for the human genome together with appropriate baseline methods.
- **An adaptable benchmarking framework** for preparing embeddings and training lightweight supervised models.
- **Result: DNA LMs approach expert method performance on some tasks**. However, no LM consistently outperforms all others, and reasoning over very long contexts, as e.g. required for finding enhancers, is still challenging.
- **Result: DNA LMs can learn distinct features in masked language modeling.** Some LMs' embeddings primarily capture information about gene structure, while others focus on noncoding regions.

2 BACKGROUND

2.1 EUKARYOTIC DNA ORGANIZATION AND TERMINOLOGY

In order to facilitate understanding how different prediction tasks relate to various aspects of the genome, we briefly discuss the fundamental structure and function of eukaryotic genomic DNA (Figure 1). DNA is a linear polymer of four nucleotide bases, which are represented by the four letters `A`, `C`, `G` and `T`. It consists of two complementary *strands* that form a *double helix* by *base pairing* the bases `A`, `T`, and `C`, `G` respectively.

Genomic DNA is physically organized in a hierarchical manner. The DNA polymer is coiled around *histone* proteins, which reduces its physical length and plays a role in regulation. A complex of 8 histone proteins together with coiled DNA is called a *nucleosome*. Nucleosomes further condense to form *chromatin* fibers, which occur in compact (closed) or loose (open) form. This controls the accessibility of the DNA sequence to the transcriptional machinery, a process tightly regulated by chemical modifications of the histones (Bannister & Kouzarides, 2011). Chromatin can form loops, which allows regions distant in the sequence to be close in physical space. DNA appears in independent modules called *chromosomes*, which are typically millions of base pairs (bp) in length.

The genome contains *genes*, segments that are transcribed to RNA molecules and potentially translated to proteins. Protein-coding genes are structured as *introns* and *exons*. For expression, a gene is first transcribed to a pre-mRNA molecule, and introns are removed via *splicing*. This combines the exons to one contiguous sequence that encodes the protein. Flanking nucleotides in the RNA that do not code for the protein are called untranslated regions (UTRs) and can have regulatory function. In addition, genes are associated with regulatory regions such as *promoters*, *enhancers*, *silencers* and *insulators* that modulate their expression. Some elements, such as promoters, may lie in close proximity to the start of the gene, the *transcription start site* (TSS). Others can appear multiple thousands bp away from the gene, but mediate their effect by physical proximity.

2.2 LANGUAGE MODELING FOR BIOMOLECULAR SEQUENCES: FROM PROTEINS TO DNA

Over the last years, language modeling has achieved breakthroughs in representation learning for protein property and structure prediction, with transformer-based pLMs emerging as powerful foundation models, capable of learning long-range interactions fully unsupervised (Rives et al., 2019; Elnaggar et al., 2022; Lin et al., 2023). The development of pLMs benefitted from the availability of standardized, representative benchmarks, such as TAPE (Rao et al., 2019) and PEER (Xu et al., 2022), as well as long-running protein machine learning tasks with an emphasis on fair benchmarking to measure progress (Kryshtafovych et al., 2021; Zhou et al., 2019).

While LMs have been extremely successful for modeling proteins, key differences between the two types of macromolecules hinder their widespread adoption for DNA. A typical protein consists of 400-500 amino acids, which are represented as tokens from an alphabet of size 20. The analogy of amino acid tokens with word tokens in NLP, as well as the fact that size of inputs to pLMs and NLP models are on the same order of magnitude, made methods developed for NLP directly transferable to protein data, with little to no methodological adaption required (Rao et al., 2020; Elnaggar et al., 2022). The alphabet of DNA is significantly smaller (4 tokens), while at the same time sequences,

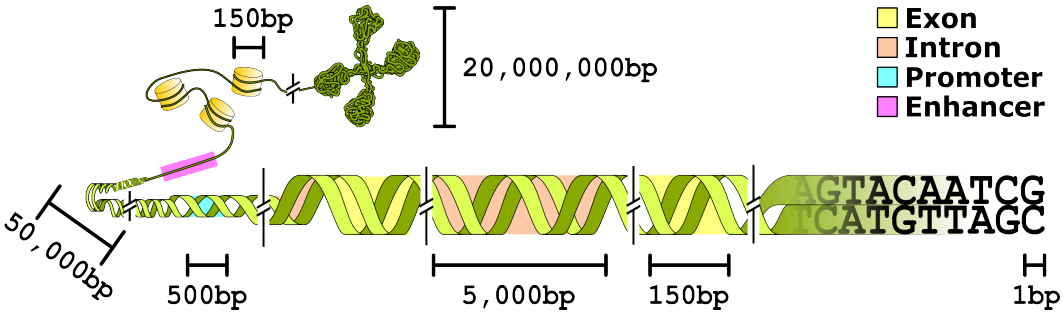

Figure 1: The organization of eukaryotic genomic DNA. Numbers are indicative examples for the human genome. Genes are structured as alternating introns (average: 5,400 bp) and exons (average: 170 bp), and have a promoter regulatory element before their TSS. Enhancers can be thousands of bp away from the gene. DNA is wrapped around histone proteins and densely packed as a chromosome.

such as those of genes, are considerably longer and have no naturally defined border, as e.g. the position of the most distant relevant regulatory element is typically unknown. In contrast, protein sequences are naturally self-contained and, being the final gene product, have a significantly higher information density. Together, sparsity and long sequences pose unique challenges to DNA LMs.

2.3 Related works

2.3.1 DNA language models

The first available DNA LM was DNABERT (Ji et al., 2021), a 12-layer BERT (Devlin et al., 2018) model trained on sequences of length 512 from the human genome. Sequences were tokenized as k-mers using a sliding window. DNABERT was evaluated by fine-tuning on tasks comprising promoter, transcription factor (TF) binding site and splice site (SS) prediction.

A growing number of DNA LMs has been proposed since the release of DNABERT. These include the Genomic Pretrained Network (GPN) (Benegas et al., 2023), FloraBERT (Levy et al., 2022), the Nucleotide Transformer (NT) (Dalla-Torre et al., 2023), Species-aware LM (Gankin et al., 2023), GENA-LM (Fishman et al., 2023), DNABERT-2 (Zhou et al., 2023) and HyenaDNA (Poli et al., 2023). With the exception of HyenaDNA, models were trained using the MLM objective, but differ in their model architectures, tokenization strategies and training data.

GPN uses dilated convolution layers rather than a transformer model. It showed strong performance for zero-shot prediction of variant effects in the *A. thaliana* genome it was trained on. Qualitative results showed that GPN captures information about gene structure and motifs of binding sites.

Nucleotide Transformer introduced the first large-scale transformer-based DNA LMs. All NT models share the same architecture, but differ in their number of training genomes and model parameters. Models were trained on either the human reference genome, 3,202 different genetically diverse human genomes or a selection of 850 genomes from a range of species. To increase the receptive field of the model, sequences were tokenized as 6-mers, allowing for processing sequences of up to 5,994 bp in length. A second generation of multispecies models released later (NT-V2) extended the input length to 12,282 bp. The NT models were evaluated on tasks comprising promoter, SS, histone modification and enhancer prediction with a context length of up to 600 bp.

GENA-LM (Fishman et al., 2023) proposed multiple medium-size LMs trained on human and multi-species genomes based on BERT and the BigBird (Zaheer et al., 2020) architecture for long sequences. Byte-Pair Encoding (BPE) was used for tokenization to further increase the receptive field, enabling an input length of about 36,000 bp. Models were evaluated on tasks comprising promoter, SS, enhancer, chromatin profile and polyadenylation site prediction. While covering the same biological phenomena, tasks were defined differently than in NT. Similarly, DNABERT-2 (Zhou et al., 2023) replaced DNABERT's k-mer tokenizer with BPE and pre-trained on multi-species genomes, while GROVER (Sanabria et al., 2023) adopted BPE for the human genome.

Predating NT and GENA-LM, FloraBERT (Levy et al., 2022) proposed pre-training on 93 different plant genomes to enable transfer learning for predicting gene expression. However, FloraBERT was

trained exclusively on promoter-containing sequences. As this requires features to already be annotated in the genome, it can be considered a departure from the paradigm of fully self-supervised learning. Similarly, Gankin et al. (2023) pre-trained on 3' UTRs from 1,500 fungal genomes. Species information was made explicit by providing a species label with each sequence to the model.

HyenaDNA (Nguyen et al., 2023) introduced a collection of autoregressive LMs, trained using the next token prediction objective at single-nucleotide resolution on the human genome. The Hyena LM architecture (Poli et al., 2023) enabled scaling to input lengths of up to 1 million nucleotides. HyenaDNA models were evaluated by fine-tuning on NT's supervised tasks and the Genomic Benchmarks (Grešová et al., 2023) collection, outperforming NT on the majority of tasks.

A number of DNA LMs were proposed without making trained models available. These comprise the original BigBird (Zaheer et al., 2020), GeneBERT, which includes the prediction of ATAC-seq signals in the pre-training stage, MoDNA (An et al., 2022), with a motif prediction task as an additional objective, the BERT-based LOGO (Yang et al., 2021), and Revolution (Cheng et al., 2023), which adopts convolutions with circular padding.

2.3.2 Supervised learning on DNA

Developing models on genomic DNA sequences for the prediction of properties and understanding of transcriptional regulation has long been a central task of computational genomics research. The availability of large-scale functional genomics data and advancements in deep learning techniques have brought progress in predicting various genomic features directly from DNA sequences. DeepBind (Alipanahi et al., 2015) and DeepSEA (Zhou & Troyanskaya, 2015) were two of the first methods leveraging shallow CNNs for predicting TF binding and chromatin features, respectively. DeepCpG (Angermueller et al., 2017) predicts DNA methylation via a CNN/GRU architecture. Basset (Kelley et al., 2016) and ChromTransfer (Salvatore et al., 2023) model chromatin state in a cell type specific manner by predicting the presence of DNase-I peaks. Using chromatin state as an auxiliary input, DeepChrome (Singh et al., 2016) predicts gene expression via multi-modal learning on DNA sequence and histone mark information.

Recently, methods for predicting gene expression have leveraged information across thousands of functional genomic tracks by training in a large-scale, multi-task fashion. Basenji (Kelley et al., 2018) and Enformer (Avsec et al., 2021) demonstrated state-of-the-art performance for gene expression prediction from DNA sequence alone, by integrating genomic information across up to 200 kilobases and multi-task training across several genome-wide functional tasks, including DNase-I activity and CAGE signal prediction. Similarly, Sei (Chen et al., 2022) models cis-regulatory TF binding, chromatin accessibility and histone modification profiles across a large range of cell types.

2.3.3 Benchmark collections on DNA

Genomic Benchmarks (Grešová et al., 2023) features balanced classification tasks on DNA sequences with median lengths ranging from 200 to 2,381 bp. The benchmark covers the classification of functional elements and the prediction of a sequence's origin. The element classification tasks are defined on human DNA, with one task covering *D. melanogaster* additionally. For each task, only performance of a baseline supervised neural network model was reported.

DNABERT-2 introduced Genome Understanding Evaluation (GUE), a collection of classification tasks ranging from 70 to 1,000 bp. On the human genome, it includes classification of promoter, SS and TF binding sequences. It covers other species with a TF binding task on mouse, a histone modification task on yeast and a Covid variant classification task on viruses. DNABERT, DNABERT-2 and NT were evaluated. No non-LM task-specific baselines are included in GUE.

The authors of NT provide a public leaderboard for their tasks, comprising promoter (human/mouse), enhancer (human), SS (human/multispecies) and histone modification (yeast) prediction with lengths ranging from 300 to 600 bp. NT is compared to DNABERT, DNABERT-2, HyenaDNA and Enformer. No task-specific baselines are included in the leaderboard.

2.3.4 Motivation of BEND

While existing DNA LMs have reported good performance on the tasks on which they were evaluated, evaluation strategies to date have shown limited consistency across individual works, with GUE constituting the most recent attempt at benchmarking on equal terms. Beyond comparability, it is important to ensure that benchmark tasks reflect the complexity and characteristics of real-world genome analysis. In practice, genomes are vast, and functional regions are sparsely distributed

Table 1: Overview of the tasks included in the benchmark. Nucleotide-wise tasks require the prediction of a sequence of labels with the same length as the input. In sequence-wise tasks the whole input sequence is to be classified. In binned tasks, multiple nucleotides share a label.

Task	Type (# labels)	# Samples	Length range	Evaluation (# train/val/test)	Metric	Source
Gene finding	Nucleotide-wise Multiclass (9)	5,976	1,433 - 14,000 bp	4780/597/597	MCC	GENCODE (Frankish et al., 2021)
Enhancer annotation	Binned (128bp) Binary	285	100,096 bp	10-fold CV	AUPRC	Fulco et al. (2019), Gasperini et al. (2019), Enformer (Avsec et al., 2021)
Chromatin accessibility	Sequence-wise Multilabel (125)	2,005,617	512 bp	1,354,042/ 279,422/372,153	AUROC	ENCODE Project Consortium (2012)
Histone modification	Sequence-wise Multilabel (19)	612,081	512 bp	420,713/ 70,801/120,567	AUROC	ENCODE Project Consortium (2012)
CpG methylation	Sequence-wise Multilabel (7)	959,039	512 bp	743,095/ 109,717/106,227	AUROC	ENCODE Project Consortium (2012)
Noncoding variant effects (expression)	Sequence-wise Binary	105,263	512 bp	zero-shot	AUROC	DeepSEA (Zhou & Troyanskaya, 2015)
Noncoding variant effects (disease)	Sequence-wise Binary	295,495	512 bp	zero-shot	AUROC	ClinVar (Landrum et al., 2020)

throughout the genome. While there are tasks on DNA that are inherently local, such as classifying functional regions (e.g. classifying TF binding sites), it needs to be recognized that such tasks do not allow us to evaluate a model's understanding of the genome over longer ranges.

Therefore, focusing solely on tasks on short sequences, such as distinguishing promoter from non-promoter sequences, falls short of evaluating the extent to which a model's representations capture complex features of genomic organization, preventing us from measuring benefits of modeling the genome with larger context windows. For instance, reporting performance on predicting SSs, which can be done on short sequences, does not allow us to evaluate how useful a model would be for gene finding over longer ranges, a common task in genome annotation.

To provide a more comprehensive assessment, BEND proposes genomic tasks that rely less on prior knowledge of feature positions and require reasoning over potentially long contexts. The tasks cover a range of length scales, selected to be both biologically relevant and to cover a variety of DNA properties. The tasks explore representations at different resolutions, requiring modelling of DNA at single bp resolution as well as over longer stretches (Table 1). We establish our benchmark on the human genome, as it offers ample experimental data for the derivation of tasks, has a complex organization, and was the focus of most published DNA LMs.

3 Tasks and Datasets

We introduce the collection of tasks included in BEND. For each task, we additionally provide a datasheet (Gebru et al., 2018) in section A.1. All tasks are provided in `bed` format, listing the genome coordinates of samples (A.2). This makes it convenient to include more flanking context without reprocessing the data, should future works find it useful to take more bp into account.

3.1 Gene finding

Definition Gene finding is a multiclass problem where each nucleotide is either part of an exon ($E_{F/R}$), intron ($I_{F/R}$), a donor ($D_{F/R}$) or acceptor ($A_{F/R}$) splice site or a noncoding region (NC). The F/R subscript denotes whether the gene is located on the forward or reverse strand.

Biological relevance Annotating genes and identifying coding sequences is a key step in genome annotation and protein discovery. It requires a model to use local context to identify correct reading frames and codon structure, while using longer range signals to propagate the location of SS to distant bp between SS, and correctly annotate them as lying in introns or exons. Introns vary in length from a few hundred to several thousand bp, requiring an LM to understand long-range dependencies.

Data GENCODE (Frankish et al., 2021) gene annotations were processed to construct sequences of nucleotide labels $y \in \{E_F, D_F, I_F, A_F, E_R, D_R, I_R, A_R, NC\}$ for each gene. Detailed processing is laid out in A.1.1. Samples were partitioned at 80% identity following AUGUSTUS' recommendations (Stanke & Waack, 2003). It should be noted that there is a large label imbalance as there is only one donor and acceptor site per intron segment.

Metric We compute the multi-class Matthews correlation coefficient (MCC) (Gorodkin, 2004) over all bp. The MCC is used as it is robust to the inherently highly uneven label ratios of this task.

3.2 ENHANCER ANNOTATION

Definition Enhancer annotation is the problem of finding enhancer regions for a given gene. We define enhancer annotation as a binary classification task. Given a sequence of gene-adjacent genomic DNA that contains enhancers, a binary label indicating whether it is part of an enhancer needs to be predicted for each segment of 128bp.
Biological relevance Enhancers are short, noncoding segments that contribute to regulating gene expression. They can be located anywhere from a few thousand to a million bp away from their target gene and work by being brought into physical proximity to the gene's promoter. Their annotation is a highly challenging task that requires detection of long-range interactions.
Data Experimentally validated enhancer-gene pairs were taken from CRISPR interference experiments (Fulco et al. (2019); Gasperini et al. (2019) and paired with the main TSS of each gene from Avsec et al. (2021). We extracted a sequence of 100,096 bp centered on the TSS for each gene. Each 128bp were annotated with a binary label $y \in \{0, 1\}$ indicating whether the bin is part of an enhancer, yielding a label sequence of length 782. Detailed processing is laid out in A.1.2. Samples were partitioned based on chromosomes.
Metric The AUPRC is computed over all labels. As the number of samples is too limited for measuring performance robustly on a single test split, we perform 10-fold cross-validation in order to evaluate performance over all samples.

3.3 CHROMATIN ACCESSIBILITY PREDICTION

Definition Chromatin accessibility prediction is a multilabel task where sequences are classified as being in open or closed chromatin across a range of cell types.
Biological relevance Dynamically modulating chromatin accessibility is a key mechanism for the cell type specific regulation of gene expression, as binding of the transcription machinery is highly dependent on the accessibility of DNA elements, including promoters, enhancers and TSS.
Data DNase I hypersensitive sites were obtained from ENCODE (ENCODE Project Consortium, 2012) for 125 cell types. Following the preprocessing of Kelley et al. (2016), segments of length 512 bp were labeled with binary vectors $\mathbf{y} \in \{0, 1\}^{125}$, with $y_i = 1$ if the chromatin is open for the i'th cell type. Detailed processing is laid out in A.1.3. Samples were partitioned based on chromosomes.
Metric The AUROC is computed for each label and averaged.

3.4 HISTONE MODIFICATION PREDICTION

Definition Histone modification prediction is a multilabel task, where the histones which are part of the nucleosomes of a given DNA sequence are labeled with one or more histone marks.
Biological relevance Histone proteins are key to the organisation of DNA into chromatin. Modifications of histones modulate chromatin structure and thus contribute to regulating chromatin accessibility and gene expression. Histone modification prediction requires modeling local binding of TFs as well as long-range regulation, such as by distant enhancers.
Data Histone ChIP-seq data for 11 histone marks and 19 replicates in the K562 cell line was obtained from ENCODE. Detailed processing is laid out in A.1.4 and follows the methodology of 3.3. Each sample is a sequence of length 512 bp with a label vector $\mathbf{y} \in \{0, 1\}^{19}$, such that $y_i = 1$ if a histone bound to this sequence carries mark i. Samples were partitioned based on chromosomes.
Metric The AUROC is computed for each label and averaged.

3.5 CPG METHYLATION PREDICTION

Definition CpG methylation prediction is a multilabel classification task, where a given CpG site is either methylated or unmethylated in different cell lines.
Biological relevance Methylation of cytosine nucleotides in CpG sites is a prominent form of epigenetic modification and plays a key role in the repression of gene expression.
Data Bisulfite sequencing data for 7 human cell lines was obtained from ENCODE. Detailed processing is laid out in A.1.5. Each sample is a sequence of length 512 bp centered on the CpG site with a label vector $\mathbf{y} \in \{0, 1\}^{7}$, such that $y_i = 1$ if the C is methylated. Samples were partitioned based on chromosomes.
Metric The AUROC is computed for each label and averaged.

3.6 NONCODING VARIANT EFFECTS (EXPRESSION AND DISEASE)

Definition Predicting variant effects is a binary problem, where single-bp mutations are classified as either having an effect or not. We treat classification as a zero-shot task, using the cosine distance

Table 2: Overview of the LMs applicable to the human genome included in the benchmark.

Model	Seq length	Trained on	Architecture	Source
AWD-LSTM	Infinite[a]	Multispecies	RNN	This work
Dilated ResNet	10,000	Human Ref[d]	CNN	This work
Nucleotide Transformer	5,994	Human Ref[d]	BERT	Dalla-Torre et al. (2023)
Nucleotide Transformer	5,994	Multispecies	BERT	Dalla-Torre et al. (2023)
Nucleotide Transformer	5,994	1000 Genomes project[d]	BERT	Dalla-Torre et al. (2023)
Nucleotide Transformer V2	12,282	Multispecies	BERT	Dalla-Torre et al. (2023)
DNABERT	512	Human Ref[d]	BERT	Ji et al. (2021)
DNABERT-2	Infinite[b]	Multispecies	BERT	Zhou et al. (2023)
GENA-LM	4,500	1000 Genomes project[e]	BERT	Fishman et al. (2023)
GENA-LM	36,000	1000 Genomes project[e]	BigBird	Fishman et al. (2023)
HyenaDNA	1,000,000	Human Ref[d]	Hyena	Nguyen et al. (2023)
HyenaDNA	1,000	Human Ref[d]	Hyena	Nguyen et al. (2023)
GROVER	8,160[c]	Human Ref[d]	BERT	Sanabria et al. (2023)

[a] As the LSTM compresses all preceding tokens into a single hidden state, it can technically process infinite sequences, even though it was trained at finite lengths and might not have learnt to exploit such long contexts.
[b] DNABERT-2 uses ALiBi (Press et al., 2022) to encode position, which can technically scale to any sequence length. In practice, the model was trained on finite lengths and the authors recommend embedding sequences below 10,000 bp.
[c] No explicit length was reported in bp. The indicated number was derived by considering 510 BPE tokens of size 16.
[d] Schneider et al. (2017), [e] McVean et al. (2012)

in embedding space between a variant nucleotide and its reference nucleotide as the prediction score.

Biological relevance Single-bp variants in noncoding regions can have functional consequences by altering gene expression levels or causing disease. This task probes the LM's understanding of local context and potentially the structure of regulatory motifs. We focus on noncoding regions, as coding variant effects can be predicted with high accuracy by modeling the mutation in the resulting protein sequence (Frazer et al., 2021).

Data For expression variants, we adapt the DeepSEA dataset (Zhou & Troyanskaya, 2015). For disease-associated variants, we process ClinVar (Landrum et al., 2020). We apply Ensembl VEP (McLaren et al., 2016) to categorize variants by genomic regions into consequence types. Detailed processing is laid out in A.1.6 and A.1.7. Each variant is a genomic position with a mutation $x \in \{\text{A, C, G, T}\}$ and a label $y \in \{0, 1\}$. The adjacent 512 bp serve as embedding context.

Metric We compute the AUROC. Additionally, we report separate AUROCs for the variant consequence types to gain further insight into what genomic features are driving performance.

4 Modeling

Language Models We benchmark available LMs suitable for the human genome (Table 2). Checkpoint selection criteria are laid out in A.6.2. Additionally, we train two simple baseline DNA LMs: An AWD-LSTM (Merity et al., 2017) model trained on three species, and a dilated CNN similar to GPN (Benegas et al., 2023), trained on the human genome. The model differs from GPN in the parameter count and the length of training sequences (A.6.1).

Downstream model We train a lightweight supervised two-layer CNN model with 64 channels on top of the LM embeddings for each task. LM weights are kept frozen and are not fine-tuned. For LMs with reduced output length due to tokenization, embeddings are upsampled to the original sequence length (A.6.3). For sequence-level tasks, we apply average pooling after the last CNN layer. For the enhancer annotation task, the number of channels was reduced to prevent overfitting. No downstream model is trained for variant effect prediction, as the cosine distance of the LM embeddings directly serves as the zero-shot predictor.

Supervised baselines For each task, we train two supervised models without pre-training. For a direct comparison of raw and embedded DNA, we train the two-layer CNN on one-hot encoded sequences. For chromatin accessibility, histone modificaton and CpG methylation prediction, we train the Basset model (Kelley et al., 2016), which was specifically designed for modeling genome-wide functional genomics data. For gene finding and enhancer annotation, we train the ResNet CNN model on one-hot encoded sequences. For variant effect prediction, no supervised models are trained. For all tasks where Basset is not applicable, we report the performance of a previously published task-specific expert method on the benchmark dataset to put LM performance into context.

Table 3: Results on all tasks. The best performing DNA LM for each task is highlighted in bold.

		Gene finding	Enhancer annotation	Chromatin accessibility	Histone modification	CpG Methylation	Variant effects (expression)	Variant effects (disease)
Expert method		0.80 AUGUSTUS	0.07 ENFORMER	0.85 BASSET	0.74 BASSET	0.93 BASSET	0.70 DEEPSEA	0.56 DEEPSEA
Fully supervised	**ResNet**	0.46	0.06	-	-	-	-	-
	CNN	0.00	0.03	0.75	0.76	0.84	-	-
Pre-trained	**ResNet-LM**	0.36	0.02	0.82	0.77	0.87	0.55	0.55
	AWD-LSTM	0.05	0.03	0.69	0.74	0.81	0.53	0.45
	NT-H	0.41	0.05	0.74	0.76	0.88	0.55	0.48
	NT-MS	**0.68**	**0.06**	0.79	0.78	**0.92**	0.54	**0.77**
	NT-1000G	0.49	0.04	0.77	0.77	0.89	0.45	0.49
	NT-V2	0.64	0.05	0.80	0.76	0.91	0.48	0.48
	DNABERT	0.20	0.03	**0.85**	**0.79**	0.91	**0.60**	0.56
	DNABERT-2	0.43	0.03	0.81	0.78	0.90	0.49	0.51
	GENA-LM BERT	0.52	0.03	0.76	0.78	0.91	0.49	0.55
	GENA-LM BigBird	0.39	0.04	0.82	0.78	0.91	0.49	0.52
	HyenaDNA large	0.35	0.03	0.84	0.76	0.91	0.51	0.45
	HyenaDNA tiny	0.10	0.02	0.78	0.76	0.86	0.47	0.44
	GROVER	0.28	0.03	0.82	0.77	0.89	0.56	0.51

5 RESULTS

Gene finding DNA LMs show promising performance for gene finding (Table 3). The two-layer CNN baseline fails to learn, possibly due to its inherent limitation to local context. However, the same CNN is able to achieve varying levels of performance when using LM embeddings, suggesting that embeddings capture some long-range information. NT-MS and NT-V2 outperform all other models by a wide margin, but still do not approach the highly specialized AUGUSTUS (Stanke & Waack, 2003) gene finding model. This highlights that while more specialized downstream models are still needed to accurately predict gene structure, using pre-trained DNA LM embeddings presents a promising avenue to attain good performance. Computing individual performance metrics across all classes (Table A8) reveals that although there is high variance in the performance across all classes, some embeddings capture splice sites fairly considering their low frequency. HyenaDNA-large, although being the only LM whose context length fully covers the input length of the task, only shows modest performance.

Enhancer annotation All investigated models perform poorly on this task. Enhancer annotation is an extremely difficult task due to the length scale, sparsity of the signal, and small dataset, which pose challenges for all investigated models. Although the supervised baseline has a large enough receptive field to detect the long-range interaction, the size of the dataset is prohibitive for performance. The performance of Enformer (Avsec et al., 2021) (A.7) is comparable on this task, but it must be noted that this is an unsupervised method that was not trained directly on enhancer data. Rather, it infers their locations from learning to predict other genome annotations. Predicting gene-specific enhancers from sequence alone without considering supporting experimental data as input therefore remains a highly difficult problem. While this task already proves to be highly challenging for current models at the given length scales, we note that biology is even more complex, with enhancers potentially being millions of bp away.

Chromatin accessibility DNABERT shows the highest performance, on par with the specialized Basset model (0.85). All other LMs perform worse on this task, offering no advantage over Basset.

Histone modification NT-MS and DNABERT show the highest performance (0.74), outperforming Basset (0.72). This suggests that LM embeddings can improve performance for histone modification prediction, albeit at marginal levels.

CpG methylation NT-MS performs best (0.92) on all included cell lines (Table A11), but is outperformed by Basset (0.93). DNABERT, GENA-LM and HyenaDNA-large also perform competitively, indicating that embeddings capture information about CpG island methylation patterns.

Variant effect prediction DNA LMs show some signal for unsupervised prediction of noncoding variant effects. As the two datasets focus on different genomic regions, we only see limited consistency between the expression and disease variant tasks, with DNABERT and NT-MS performing best respectively. While being worse than the supervised DeepSEA method, DNABERT matches DeepSEA's unsupervised performance on the expression dataset (AUROC 0.6, A.7). On the disease

dataset, multiple LMs approach DeepSEA's Disease Impact Score, with NT-MS outperforming it. When dissecting performance by variant types, we find that the performance of NT-MS is driven by variants affecting splice sites and introns (Table A13). While splice sites can be considered noncoding DNA, they are not the focus of DeepSEA, which models chromatin features, and shows stronger performance in UTRs and up- or downstream regions. Similar to the results on the expression dataset, we find that DNABERT outperforms NT-MS in such regions, suggesting that the two LMs learned distinct sequence features during pre-training. As all other NT models show weaker performance on variants affecting gene structure, this could be a consequence of the model's large size and multi-species pre-training. However, we do not see similarly strong performance in the multi-species DNABERT-2 and NT-V2.

6 Discussion

We find that currently available DNA LMs already show promising performance on some tasks over fully supervised baselines, but do not offer consistent improvements over all included tasks and can fall short of surpassing specialized existing prediction methods. Overall, we find that NT-MS is a strong default LM, but is in some tasks inferior to the much smaller DNABERT. Interestingly, while both models trained using the MLM objective, we find that they learned distinct genomic features during pre-training. With the pre-training data and the tokenization strategy being the key architectural difference, these choices may deserve more attention in future DNA LMs.

For modeling functional genomics data, DNA LMs only show limited utility. In direct comparison to the Basset model trained on the same data, LM embeddings fail to yield consistent improvements in performance when only using a two-layer CNN.

On the gene finding task, we observe that NT-MS with a simple two-layer CNN shows promising performance compared to the specialized AUGUSTUS, which was found to be the state of the art in a recent benchmark (Scalzitti et al., 2020). This suggests that future more sophisticated LM-based gene finders might become a method of choice for this problem. The result also indicates that current DNA LMs are capable of modeling long-range dependencies to some extent.

Probing LMs at even longer ranges in the enhancer annotation task reveals that long-range understanding still needs improvement for sparse problems with limited data. This highlights a key issue facing DNA LMs: Not only is there a need for long-range modeling to improve our understanding of the genome, as demonstrated by Avsec et al. (2021), but it also raises a fundamental question as to whether current LM training objectives will lead to the incorporation of such distant, sparse signals, or whether the local sequence context is all that is required for sequence reconstruction and some level of supervision is needed. Since BEND is not inherently tied to an LM objective, our standardized benchmark may also prove useful for evaluating eventual DNA representation models that follow a different paradigm.

7 Limitations and Outlook

As the curation of a comprehensive benchmark task collection requires experimental ground-truth data to be available, and most published models are trained on human data, we focused BEND on the human genome. BEND aims at comparing the effectiveness of different model architectures and training strategies for learning representations from genomic data, under the assumption that other, similarly structured genomes should behave comparably under self-supervision. However, an important question that remains unanswered is whether DNA LMs can aid with generalization across different organisms. In the future, we hope to extend the benchmark to other, diverse organisms, so that generalization power can be tested in a transfer-learning setting, i.e. by training a task on a given organism, and evaluating performance on another.

In BEND, we benchmarked to what extent embeddings capture features that can be leveraged by downstream models for prediction. This approach is fully agnostic regarding the underlying LM's methodology and scales to models of any size. Other works proposed to fine-tune LMs on tasks directly. While this potentially conflates a representation's content with the inductive bias of a model architecture for a given task, fine-tuning may yield performance gains beyond the results observed in this work (Nguyen et al., 2023; Zhou et al., 2023). Another aspect to be investigated in the future is to dive deeper into how LMs learn features during pre-training, as done previously for protein LMs (Vig et al., 2021).

REFERENCES

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, August 2015. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3300. URL https://www.nature.com/articles/nbt.3300.

Weizhi An, Yuzhi Guo, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li, and Junzhou Huang. MoDNA: motif-oriented pre-training for DNA language model. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–5, Northbrook Illinois, August 2022. ACM. ISBN 978-1-4503-9386-7. doi: 10.1145/3535508.3545512. URL https://dl.acm.org/doi/10.1145/3535508.3545512.

Christof Angermueller, Heather J. Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1):67, December 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1189-z. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1189-z.

Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa

Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Bioinformatics Institute European Molecular Biology Laboratory, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and NIH National Eye Institute. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015. ISSN 1476-4687. doi: 10.1038/nature15393. URL https://www.nature.com/articles/nature15393. Number: 7571 Publisher: Nature Publishing Group.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x. Number: 10 Publisher: Nature Publishing Group.

Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, March 2011. ISSN 1748-7838. doi: 10.1038/cr.2011.22. URL https://doi.org/10.1038/cr.2011.22.

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. DNA language models are powerful zero-shot predictors of genome-wide variant effects. *bioRxiv*, pp. 2022.08.22.504706, January 2023. doi: 10.1101/2022.08.22.504706. URL http://biorxiv.org/content/early/2023/04/12/2022.08.22.504706.abstract.

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, June 2021. ISSN 24054712. doi: 10.1016/j.cels.2021.05.017. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471221002039.

Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7):940–949, July 2022. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-022-01102-2. URL https://www.nature.com/articles/s41588-022-01102-2.

Lei Cheng, Tong Yu, Tero Aittokallio, Jukka Corander, Ruslan Khalitov, and Zhirong Yang. Self-supervised learning for DNA sequences with circular dilated convolutional networks. preprint, Bioinformatics, February 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.01.30.526193.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, pp. 2023.01.11.523679, January 2023. doi: 10.1101/2023.01.11.523679. URL http://biorxiv.org/content/early/2023/01/15/2023.01.11.523679.abstract.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. doi: 10.48550/ARXIV.1810.04805. URL https://arxiv.org/abs/1810.04805. Publisher: arXiv Version Number: 2.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3095381. URL https://ieeexplore.ieee.org/document/9477085/.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 1476-4687. doi: 10.1038/nature11247.

Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences, June 2023. URL https://www.biorxiv.org/content/10.1101/2023.06.12.544594v1. Pages: 2023.06.12.544594 Section: New Results.

Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L Howe, Toby Hunt, Osagie G Izuogu, Rory Johnson, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y Wolf, Jinuri Xu, Yucheng T Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Michael L Tress, and Paul Flicek. GENCODE 2021. *Nucleic Acids Research*, 49(D1):D916–D923, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1087. URL https://doi.org/10.1093/nar/gkaa1087.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04043-8. URL https://doi.org/10.1038/s41586-021-04043-8.

Charles P. Fulco, Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, Rockwell Anyoha, Benjamin R. Doughty, Tejal A. Patwardhan, Tung H. Nguyen, Michael Kane, Elizabeth M. Perez, Neva C. Durand, Caleb A. Lareau, Elena K. Stamenova, Erez Lieberman Aiden, Eric S. Lander, and Jesse M. Engreitz. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669, December 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0538-0. URL https://www.nature.com/articles/s41588-019-0538-0. Number: 12 Publisher: Nature Publishing Group.

Dennis Gankin, Alexander Karollus, Martin Grosshauser, Kristian Klemon, Johannes Hingerl, and Julien Gagneur. Species-aware DNA language modeling. *bioRxiv*, pp. 2023.01.26.525670, January 2023. doi: 10.1101/2023.01.26.525670. URL http://biorxiv.org/content/early/2023/01/27/2023.01.26.525670.abstract.

Molly Gasperini, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S. Noble, Cole Trapnell, Nadav Ahituv, and Jay Shendure. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1):377–390.e19, January 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2018.11.029. URL https://www.sciencedirect.com/science/article/pii/S009286741831554X.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for Datasets. 2018. doi: 10.48550/ARXIV.1803.09010. URL https://arxiv.org/abs/1803.09010. Publisher: arXiv Version Number: 8.

J. Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367–374, December 2004. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2004.09.006. URL https://www.sciencedirect.com/science/article/pii/S1476927104000799.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, May 2023. ISSN 2730-6844. doi: 10.1186/s12863-023-01123-8. URL https://doi.org/10.1186/s12863-023-01123-8.

Benjamin C. Hitz, Jin-Wook Lee, Otto Jolanki, Meenakshi S. Kagda, Keenan Graham, Paul Sud, Idan Gabdank, J. Seth Strattan, Cricket A. Sloan, Timothy Dreszer, Laurence D. Rowe, Nikhil R. Podduturi, Venkat S. Malladi, Esther T. Chan, Jean M. Davidson, Marcus Ho, Stuart Miyasato, Matt Simison, Forrest Tanaka, Yunhai Luo, Ian Whaling, Eurie L. Hong, Brian T. Lee, Richard Sandstrom, Eric Rynes, Jemma Nelson, Andrew Nishida, Alyssa Ingersoll, Michael Buckley, Mark Frerker, Daniel S Kim, Nathan Boley, Diane Trout, Alex Dobin, Sorena Rahmanian, Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, Neva C. Durand, Olga Dudchenko, David Weisz, Suhas S. P. Rao, Alyssa Blackburn, Dimos Gkountaroulis, Mahdi Sadr, Moshe Olshansky, Yossi Eliaz, Dat Nguyen, Ivan Bochkov, Muhammad Saad Shamim, Ragini Mahajan, Erez Aiden, Tom Gingeras, Simon Heath, Martin Hirst, W. James Kent, Anshul Kundaje, Ali Mortazavi, Barbara Wold, and J. Michael Cherry. The ENCODE Uniform Analysis Pipelines. preprint, Bioinformatics, April 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.04.04.535623.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, August 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

Meenakshi S. Kagda, Bonita Lam, Casey Litton, Corinn Small, Cricket A. Sloan, Emma Spragins, Forrest Tanaka, Ian Whaling, Idan Gabdank, Ingrid Youngworth, J. Seth Strattan, Jason Hilton, Jennifer Jou, Jessica Au, Jin-Wook Lee, Kalina Andreeva, Keenan Graham, Khine Lin, Matt Simison, Otto Jolanki, Paul Sud, Pedro Assis, Philip Adenekan, Eric Douglas, Mingjie Li, Pedro Assis, Keenan Graham, Paul Sud, Stuart Miyasato, Weiwei Zhong, Yunhai Luo, Zachary Myers, J. Michael Cherry, and Benjamin C. Hitz. Data navigation on the ENCODE portal. 2023. doi: 10.48550/ARXIV.2305.00006. URL https://arxiv.org/abs/2305.00006. Publisher: arXiv Version Number: 2.

David R. Kelley, Jasper Snoek, and John L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, July 2016. ISSN 1549-5469. doi: 10.1101/gr.200535.115.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, May 2018. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.227819.117. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.227819.117.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021. ISSN 1097-0134. doi: 10.1002/prot.26237. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26237. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26237.

Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, Vitaly Lyoshin, Zenith Maddipatla, Rama Maiti, Joseph Mitchell, Nuala O'Leary, George R Riley, Wenyao Shi, George Zhou, Valerie Schneider, Donna Maglott, J Bradley Holmes, and Brandi L Kattman. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz972. URL https://doi.org/10.1093/nar/gkz972.

Richard Leslie, Christopher J. O'Donnell, and Andrew D. Johnson. GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, 30(12):i185–i194, June 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu273. URL https://doi.org/10.1093/bioinformatics/btu273.

Benjamin Levy, Zihao Xu, Liyang Zhao, Karl Kremling, Ross Altman, Phoebe Wong, and Chris Tanner. FloraBERT: cross-species transfer learning withattention-based neural networks for geneexpression prediction. preprint, In Review, August 2022. URL https://www.researchsquare.com/article/rs-1927200/v1.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Yunhai Luo, Benjamin C. Hitz, Idan Gabdank, Jason A. Hilton, Meenakshi S. Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, Ulugbek K. Baymuradov, Keenan Graham, Casey Litton, Stuart R. Miyasato, J. Seth Strattan, Otto Jolanki, Jin-Wook Lee, Forrest Y. Tanaka, Philip Adenekan, Emma O'Neill, and J. Michael Cherry. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48(D1):D882–D889, January 2020. ISSN 1362-4962. doi: 10.1093/nar/gkz1062.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, January 2023. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2.

William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, June 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4. URL https://doi.org/10.1186/s13059-016-0974-4.

Gil A. McVean, David M. Altshuler (Co-Chair), Richard M. Durbin (Co-Chair), Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs (Principal Investigator), Huyen Dinh, Christie Kovar, Sandra Lee, Lora Lewis, Donna Muzny, Jeff Reid, Min Wang, Jun Wang (Principal Investigator), Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Guoqing Li, Jingxiang Li, Yingrui Li, Zhuo Li, Xiao Liu, Yao Lu, Xuedi Ma, Zhe Su, Shuaishuai Tai, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Ye Yin, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Yan Zhou, Eric S. Lander (Principal Investigator), David M. Altshuler, Stacey B. Gabriel (Co-Chair), Namrata Gupta, Paul Flicek (Principal Investigator), Laura Clarke, Rasko Leinonen, Richard E. Smith, Xiangqun Zheng-Bradley, David R. Bentley (Principal Investigator), Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach (Principal Investigator), Ralf Sudbrak (Project Leader), Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Stephen T. Sherry (Principal Investigator), Gil A. McVean (Principal Investigator), Elaine R. Mardis (Co-Principal Investigator) (Co-Chair), Richard K. Wilson (Co-Principal Investigator), Lucinda Fulton, Robert Fulton, George M. Weinstock, Richard M. Durbin (Principal Investigator), Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt (Principal Investigator), Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton (Principal Investigator), Richard A. Gibbs (Principal Investigator), Fuli Yu (Project Leader), Matthew Bainbridge, Danny Challis, Uday S. Evani, James Lu, Donna

Muzny, Uma Nagaswamy, Jeff Reid, Aniko Sabo, Yi Wang, Jin Yu, Jun Wang (Principal Investigator), Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Nan Qin, Haojing Shao, Bingqiang Wang, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Gabor T. Marth (Principal Investigator), Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Charles Lee (Principal Investigator), Lauren Griffin, Chih-Heng Hsieh, Ryan E. Mills, Xinghua Shi, Marcin von Grotthuss, Chengsheng Zhang, Mark J. Daly (Principal Investigator), Mark A. DePristo (Project Leader), David M. Altshuler, Eric Banks, Gaurav Bhatia, Mauricio O. Carneiro, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Robert E. Handsaker, Chris Hartl, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Stephen F. Schaffner, Khalid Shakir, Seungtai C. Yoon (Principal Investigator), Jayon Lihm, Vladimir Makarov, Hanjun Jin (Principal Investigator), Wook Kim, Ki Cheol Kim, Jan O. Korbel (Principal Investigator), Tobias Rausch, Paul Flicek (Principal Investigator), Kathryn Beal, Laura Clarke, Fiona Cunningham, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Xiangqun Zheng-Bradley, Andrew G. Clark (Principal Investigator), Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Pardis C. Sabeti (Principal Investigator), Sharon R. Grossman, Shervin Tabrizi, Ridhi Tariyal, David N. Cooper (Principal Investigator), Edward V. Ball, Peter D. Stenson, David R. Bentley (Principal Investigator), Bret Barnes, Markus Bauer, R. Keira Cheetham, Tony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Kai Ye (Principal Investigator), Mark A. Batzer (Principal Investigator), Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur (Principal Investigator), Monkol Lek, Sudbrak (Project Leader), Vyacheslav S. Amstislavskiy, Ralf Herwig, Mark D. Shriver (Principal Investigator), Carlos D. Bustamante (Principal Investigator), Jake K. Byrnes, Francisco M. De La Vega, Simon Gravel, Eimear E. Kenny, Jeffrey M. Kidd, Phil Lacroute, Brian K. Maples, Andres Moreno-Estrada, Fouad Zakharia, Eran Halperin (Principal Investigator), Yael Baran, David W. Craig (Principal Investigator), Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Stephen T. Sherry (Principal Investigator), Chunlin Xiao, Jonathan Sebat (Principal Investigator), Vineet Bafna, Kenny Ye, Esteban G. Burchard (Principal Investigator), Ryan D. Hernandez (Principal Investigator), Christopher R. Gignoux, David Haussler (Principal Investigator), Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares (Principal Investigator), The 1000 Genomes Project Consortium, Corresponding Author, Steering committee, Production group:, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, US National Institutes of Health, University of Oxford, Washington University in St Louis, Wellcome Trust Sanger Institute, Analysis group:, Affymetrix, Albert Einstein College of Medicine, Boston College, Brigham and Women's Hospital, Cold Spring Harbor Laboratory, Dankook University, European Molecular Biology Laboratory, Cornell University, Harvard University, Human Gene Mutation Database, Leiden University Medical Center, Louisiana State University, Massachusetts General Hospital, Pennsylvania State University, Stanford University, Tel-Aviv University, Translational Genomics Research Institute, San Diego University of California, San Francisco University of California, Santa Cruz University of California, University of Chicago, University College London, and University of Geneva. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012. ISSN 1476-4687. doi: 10.1038/nature11632. URL https://doi.org/10.1038/nature11632.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. 2017. doi: 10.48550/ARXIV.1708.02182. URL https://arxiv.org/abs/1708.02182. Publisher: arXiv Version Number: 1.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. 2023. doi: 10.48550/ARXIV.2306.15794. URL https://arxiv.org/abs/2306.15794. Publisher: arXiv Version Number: 1.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena Hierarchy: Towards Larger Convolutional Language Models, April 2023. URL http://arxiv.org/abs/2302.10866. arXiv:2302.10866 [cs].

Ofir Press, Noah A. Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation, April 2022. URL http://arxiv.org/abs/2108.12409. arXiv:2108.12409 [cs].

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE. *Advances in neural information processing systems*, 32:9689–9701, December 2019. ISSN 1049-5258. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7774645/.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. preprint, Synthetic Biology, December 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.15.422761.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. preprint, Synthetic Biology, April 2019. URL http://biorxiv.org/lookup/doi/10.1101/622803.

Marco Salvatore, Marc Horlacher, Annalisa Marsico, Ole Winther, and Robin Andersson. Transfer learning identifies sequence determinants of cell-type specific regulatory element accessibility. *NAR Genomics and Bioinformatics*, 5(2):lqad026, March 2023. ISSN 2631-9268. doi: 10.1093/nargab/lqad026. URL https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqad026/7092956.

Melissa Sanabria, Jonas Hirsch, and Anna R. Poetsch. The human genome's vocabulary as proposed by the DNA language model GROVER, September 2023. URL https://www.biorxiv.org/content/10.1101/2023.07.19.549677v2. Pages: 2023.07.19.549677 Section: New Results.

Nicolas Scalzitti, Anne Jeannin-Girardon, Pierre Collet, Olivier Poch, and Julie D. Thompson. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics*, 21(1):293, December 2020. ISSN 1471-2164. doi: 10.1186/s12864-020-6707-9. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6707-9.

Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan M. Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, Evan E. Eichler, and Deanna M. Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, May 2017. ISSN 1549-5469. doi: 10.1101/gr.213611.116.

Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17): i639–i648, September 2016. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btw427. URL https://academic.oup.com/bioinformatics/article/32/17/i639/2450757.

Mario Stanke and Stephan Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics (Oxford, England)*, 19 Suppl 2:ii215–225, October 2003. ISSN 1367-4811. doi: 10.1093/bioinformatics/btg1080.

Felix Teufel, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. GraphPart: homology partitioning for biological sequence analysis. *NAR Genomics and Bioinformatics*, 5(4):lqad088, October 2023. ISSN 2631-9268. doi: 10.1093/nargab/lqad088. URL https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqad088/7318077.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models, March 2021. URL http://arxiv.org/abs/2006.15222. arXiv:2006.15222 [cs, q-bio].

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e467582d42d9c13fa9603df16f31de6d-Abstract-Datasets_and_Benchmarks.html.

Meng Yang, Haiping Huang, Lichao Huang, Nan Zhang, Jihong Wu, Huanming Yang, and Feng Mu. LOGO, a contextualized pre-trained language model of human genome flexibly adapts to various downstream tasks by fine-tuning. preprint, In Review, August 2021. URL https://www.researchsquare.com/article/rs-448927/v1.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, October 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3547. URL https://www.nature.com/articles/nmeth.3547.

Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S. Rifaioglu, Alperen Dalkıran, Rengul Cetin Atalay, Chengxin Zhang, Rebecca L. Hurto, Peter L. Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M. Fernández, Branislava Gemovic, Vladimir R. Perovic, Radoslav S. Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad R.K. Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Heiko Schoof, Indika Kahanda, Natalie Thurlby, Alice C. McHardy, Alexandre Renaux, Rabie Saidi, Julian Gough, Alex A. Freitas, Magdalena Antczak, Fabio Fabris, Mark N. Wass, Jie Hou, Jianlin Cheng, Zheng Wang, Alfonso E. Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J. Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po-Han Chi, Wei-Cheng Tseng, Michal Linial, Peter W. Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T. Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Warwick Vesztrocy, Jose Manuel Rodriguez, Michael L. Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B. Roche, Jonas Reeb, David W. Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang, Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio C.E. Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S. Black, Dane Jo, Erica Suh, Jonathan B. Dayton, Dallas J. Larsen, Ashton R. Omdahl, Liam J. McGuffin, Danielle A. Brackenridge, Patricia C. Babbitt, Jeffrey M. Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia-Ming Chang, Wen-Hung Liao, Yi-Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E.E. Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E. Brenner, Christine A. Orengo, Constance J. Jeffery, Giovanni Bosco, Deborah A. Hogan, Maria J. Martin, Claire O'Donovan, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, and Iddo Friedberg. The CAFA challenge reports improved protein

function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):244, November 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1835-8. URL https://doi.org/10.1186/s13059-019-1835-8.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome, June 2023. URL http://arxiv.org/abs/2306.15006. arXiv:2306.15006 [cs, q-bio].