GUANinE v1.0: Benchmark Datasets for Genomic AI Sequence-to-Function Models

eyes s. robson
Center for Computational Biology

UC Berkeley
Berkeley, CA 94720
eyes.robson@berkeley.edu

Nilah M. Ioannidis
Department of Electrical Engineering
and Computer Sciences
UC Berkeley
Berkeley, CA 94720
nilah@berkeley.edu

Abstract

Computational genomics increasingly relies on machine learning methods for genome interpretation, and the recent adoption of neural sequence-to-function models highlights the need for rigorous model specification and controlled evaluation, problems familiar to other fields of AI. Research strategies that have greatly benefited other fields — including benchmarking, auditing, and algorithmic fairness — are also needed to advance the field of genomic AI and to facilitate model development. Here we propose a genomic AI benchmark, `GUANinE`, for evaluating model generalization across a number of distinct genomic tasks. Compared to existing task formulations in computational genomics, `GUANinE` is large-scale, de-noised, and suitable for evaluating pretrained models. `GUANinE` v1.0 primarily focuses on functional genomics tasks such as functional element annotation and gene expression prediction, and it also draws upon connections to evolutionary biology through sequence conservation tasks. The current `GUANinE` tasks provide insight into the performance of existing genomic AI models and non-neural baselines, with opportunities to be refined, revisited, and broadened as the field matures. Finally, the `GUANinE` benchmark allows us to evaluate new self-supervised T5 models and explore the tradeoffs between tokenization and model performance, while showcasing the potential for self-supervision to complement existing pretraining procedures.

1 Introduction

Genomes are fundamental characteristics of organisms that encode the molecular machinery and regulatory functions that define cellular organization and response to stimuli. Modern statistical machinery to analyze genomes has grown in complexity, from sophisticated tree models to reconstruct demographic and evolutionary histories [57, 5] to neural network-based sequence-to-function maps (i.e. $f : X \rightarrow y$) for [epi]genomic annotation, imputation, and understanding [86, 7, 87, 38, 6]. Due to this increased reliance on high-complexity, difficult-to-interpret models, there is a need for centralized benchmarks and benchmark development tools to maximize research efficacy. Benchmarks offer new perspectives on model evaluation, assess the progress of a field over time, and provide standardized comparability between new and existing models.

Here, we curate large-scale ($N > 10^6$) preprocessed genome interpretation tasks for establishing the **G**enome **U**nderstanding and **AN**notation **in** silico **E**valuation, or `GUANinE`[1], benchmark. While ideally, predictions from genome interpretation models would be confirmed by comprehensive wetlab experiments, such experiments present a significant and costly research bottleneck for model development. Because gold-standard human evaluation [61] for genomic tasks is infeasible, the design of large-N benchmarking tasks is necessary to develop competitive baselines. Our goal is to provide a set of benchmarking tasks for genomic AI that

A) allow for direct, supervised training of high-complexity models from scratch (*tabula rasa*) for comparison to pretraining regimes for transfer learning, and

[1] `GUANinE` data and evaluation tools are available at https://github.com/ni-lab/guanine

18th Conference on Machine Learning in Computational Biology (MLCB 2023).

B) ensure statistical significance while having limited or quantifiable confounders (e.g. batch effects or socioeconomic factors [77]), a requirement of any evaluatory dataset [24].

GUANinE v1.0 prioritizes functional genomic annotation and understanding on short-to-moderate length sequences (between 80 and 512 nucleotides), rather than exploring long sequences inputs or distal effects [38, 20, 36]. Although GUANinE does not cover every domain of genomic AI, e.g. taxonomic classification or comparative biology [47, 71], it has an emphasis on tasks that require a complex understanding of genome regulation, and we hope it will encourage further task design and benchmarking in the field, similar to the diversity of tasks in NLP from cloze completion [67] to Natural Language Understanding (NLU) [12, 80].

Additionally, we make use of the standardized performance metrics of GUANinE to evaluate a variety of non-neural baselines and existing genomic AI models across diverse tasks, and we demonstrate the power of self-supervised pretraining in the genome while exploring key hyperparameter implications with numerous and varied T5 models. These experiments confirm the perplexity benefits of longer sequences [14] while demonstrating the tradeoff of reduced representation sizes for fixed-length tasks at higher levels of tokenization.

1.1 Background — Benchmarking

Benchmarking has a long history across the AI fields of natural language processing (NLP, "language") and computer vision ("vision"), from ImageNet [18], which inspired AlexNet [41] and ResNets [27] in vision, to the bilingual parallel corpora that led to the transformer architecture [78, 19] and modern benchmarks of question-answering [60] and comprehensive evaluation [80, 81, 66] in language. The potential for model development, designing new tasks, and evaluating models enabled by benchmarking is difficult to overstate. On the other hand, reliance on benchmarks is not without risks — benchmarking is an intrinsically normative process that can entrench suboptimal priorities [11], perpetuate cultural biases [11, 9], limit model expressivity [51], or yield inaccurate metrics due to duplication errors [8]. Given present and historical biases in genomics [55, 1, 77] and medicine [48], benchmarking biomedical tasks based on clinical or volunteer-based data is challenging. GUANinE relies on experimentally determined and evolutionary data for its tasks to reduce socioeconomic confounders, although *in vitro* and evolutionary data come with their own biases as we discuss.

1.2 Related Work

Previous evaluation efforts for genomic AI models and noncoding variation have utilized the Eukaryotic Promoter Database [56] for promoter annotation [50, 33, 83], gene expression data from the Roadmap [73], GTEx [72], and Geuvadis [43] consortia for promoter understanding [63, 31, 3, 87, 38], and GWAS or eQTL variant association datasets [87, 31, 20] or curated clinical variant annotations from HGMD or ClinVar [30, 79, 40, 31, 38] for variant interpretation. Recent small-scale experimental validations make use of wetlab techniques such as CRISPRi [6]. Most of these self-designed evaluations by authors are heavily tailored to the model or task of interest, rather than being explicitly intended or designed as benchmarking tasks for followup comparisons with other models. The Critical Assessment of Genome Interpretation (CAGI) challenges [29], in contrast, involve benchmarks created for specific multi-submission challenges, but are typically limited in scope, with submissions tailored specifically to each individual challenge. The recent GenomicBenchmarks paper [26] is notably distinct from other work and is the most comparable to GUANinE, although GUANinE involves a wider scope of tasks with over 60M (∼70x) training examples, a rigorous approach to task construction (e.g. repeat-downsampling and GC-content balancing), and comprehensive baselining.

2 GUANinE Tasks

The GUANinE benchmark is designed to be supervised, human (eukaryote)-centric, and well-controlled, with a focus on large training sets. Compared to other AI disciplines, the relative infeasibility of manual human-labelling for genome annotation is a clear limitation. For GUANinE, great consideration was placed into cleaning, limiting obvious confounders, and (where applicable) selecting negatives. Our suite of tasks is meant to broadly characterize human genomic complexity and span several domains of functional genomics.

2.1 Functional Elements

Endogenous functional element annotation is commonly used for the training and evaluation of genomic AI methods [50, 26]. In GUANinE , we label finite spans of nucleotides (centered at an experimentally called peak) with a scalar output corresponding to a functional 'propensity' across a catalogue of cell types. This propensity is based on a weighted sum of the number of cell types displaying signal for the consensus peak in the experiment of interest; see Appendix B for details. This scalar target subsumes the canonical vector output across multiple cell types [86, 87, 38, 6] into a concise, interpretable metric of cell type specificity. Compared to existing functional element datasets, these tasks have a relatively low ($< 20\%$) number of negatives (zeros) and stricter downsampling in repeat-masked regions.

Task	Functional Elements		Conservation		Gene Expression	
	dnase-prop	**ccre-prop**	**cons30**	**cons100**	**gpra-c**	**gpra-d**
Train set	2.8 M	7.1 M	5.4 M	5.0 M	23 M	16 M
Dev set	35 k	91 k	55 k	51 k	230 k	170 k
Test set	44 k	101 k	55 k	51 k	230 k	170 k
Split	Dev21/Test22	Dev21/Test22	Random	Random	Random	Random
Organism	Human	Human	Human	Human	Yeast	Yeast
Target	Scalar	Multi-task	Scalar	Scalar	Scalar	Scalar
Output range	0 - 4	0 - 4	0 - 24	0 - 24	0 - 17	0 - 17

Table 1: Benchmark summary statistics.

dnase-propensity This task asks a model to estimate the ubiquity of DNase Hypersensitive Site (DHS) across cell types for sequences with some non-zero DHS signal alongside negative sequences from the rest of the genome. It is constructed from the DNase hypersensitivity subset of the SCREEN v2 database [69], a collection of several hundred cell type tracks from ENCODE [70]. We label a 511 bp hg38 reference sequence with an aggregate propensity score, where a 'positive-class' score of 1 through 4 represents the relative number of cell type tracks with DNase hypersensitivity signal at the peak locus (4 being nearly ubiquitous), while a 'negative-class' score of 0 represents a partially GC-balanced negative randomly sampled from the genome. In effect, the y-value is a low-dimensional summary of the binarized accessibility across the 727 cell types. Compared to the ccre-propensity task below, this is a simpler, annotative task of DHS ubiquity. The y-values are integers ranging from 0 to 4, and we use Spearman rho in evaluations.

ccre-propensity This task asks a model to estimate DHS functionality across cell types among the subset of sequences annotated as candidate Cis-Regulatory Elements (cCREs) in ENCODE's Candidate Registry of cis-Regulatory Elements, as used in the SCREEN v2 cCRE database [69]. We start with the 'positive' DHSes from the dnase-propensity task, and we label them with the corresponding signal from each of four peak-called epigenetic markers: H3K4me3, H3K27ac, CTCF, and DNase hypersensitivity. As before, this propensity corresponds to a weighted sum of binarized signal over different cell types for each experiment type; see Appendix B for details. Each example in the ccre-propensity task has 509 bp of hg38 context centered at the middle of a DHS[2]. Compared to the dnase-propensity task, this is a more complex, understanding-based task of DHS function. The y-values are integers ranging from 0 (non-marker DHS sites from the 'positives' of the dnase-propensity task above) to 4, indicating the highest number of cell types in which a signal was detected (e.g. 100 of the 198 cell types with a CTCF experiment).

Although our dnase- and ccre- propensity tasks both reflect patterns of ubiquity versus cell type specificity, neither explicitly asks for the specific cell types in which a DHS or cCRE is active. This choice to bin (or quantize) our scores allows for clearly defined negatives without worrying about zero-inflation in the loss function, provides universal post-hoc groupings (e.g. 0 vs all, 4 vs all, etc) that reflect different standardized constructions of ubiquity and activity, and reduces the risk of inflated performance due to correlation structures or from prioritizing certain tracks to the detriment of others [64].

2.2 Conservation

Sequence conservation across evolutionarily related organisms suggests the presence of negative selection against deleterious variation and thus biological function [40, 79]. Many per-base or per-element conservation scores can be directly computed from multiple sequence alignments across related organisms [5, 54] (though these alignments may induce bias due to evolutionary divergence). Human Accelerated Regions (HARs [53]) are distinct for their markedly unconserved nature yet high impact on human physiology, as evidenced by their recent positive selection since our last common ancestor with chimpanzees and bonobos.

cons30 and cons100 These tasks are constructed by labeling 512 bp contiguous segments of the hg38 reference genome with the mean phyloP30 or phyloP100 multiple sequence alignment conservation score, respectively, as reported in Pollard et al. [54]. The 30-way alignment roughly corresponds to primates/mammals, while the 100-way corresponds to vertebrates. We minimize the inclusion of HARs by removing high-variance, lowly conserved segments, as these have undergone sharp positive selection in contrast to (ultra)conserved elements. All sequences used for task construction have 95+% coverage across the species in their respective MSAs. The y-values correspond to binned quantiles (integer) of the mean conservation score between 0 (least conserved) and 24 (most conserved), and we use Spearman rho in evaluations.

[2] The 2 bp decrease allows for length-512 models to pass a 'task-code' token as input [58]

2.3 Gene Expression

Gene expression is central to cellular identity and function, and genomic AI represents a promising advance for understanding regulatory genomics and the sequence determinants of mRNA abundance and decay [87, 3, 63, 2]. However, the space of naturally occurring promoter sequences in any one organismal context is limited [76, 17], both in sample size relative to complexity (~30,000 in humans [56]) and in diversity of sequences (e.g. phylogenetically related or constrained promoters, and similar GC-gradient patterning across genes [71]). Experimental techniques such as MPRAs or oligonucleotide assembly offer the means to perturb regulatory element motif grammars and add sequence diversity [76, 16] to ensure that genomic AI models learn causal determinants of gene expression rather than simple sequence features or correlates. The tasks below benefit from substantial size and sequence diversity, though we note that because the experiments are conducted in yeast with exogenous sequences, the performance of models designed for the human genome will be affected by the distribution shift between human and yeast.

gpra-c and gpra-d The Gigantic Parallel Reporter Assay (GPRA) tasks are large corpora of short, 80 random (+ 30 scaffold) bp promoter sequences in yeast labeled by a gene expression value measured via dual reporter single-cell fluorescence [17]. We re-process and sanitize datasets collected in both the 'complex' (gpra-c) and 'defined' (gpra-d) growth media as originally reported by Vaishnav et al. [76]. The y-values correspond to the (experimentally) binned fluorescence value of the promoter sequence expression between 0 (lowest) and 17 (highest), and we use Spearman rho in evaluations.

3 Baselines

Rigorous baselining for genomic AI in the context of benchmarks is critical [23, 49], due to the presence of confounding sequence (e.g. dinucleotide frequency [7]) and measurement (e.g. batch effect) factors. We evaluate a selection of neural and non-neural baselines in `GUANinE` to provide insight into performance, and we encourage subsequent work to include similar baselines. Our neural baselines include few-shot performance of a handful of commonly-used existing convolutional models in genomic AI, as well as a specialized transformer architecture that we evaluate both pretrained and *tabula rasa* to explore the benefits of pretraining.

3.1 Non-neural Baselines

Our non-neural baselines range from simple GC-content, a strong predictor due to its correlation with sequence function in vertebrates [71], to *k*-mer frequency models, which have an established record in genomics [25].

GC-content From its original prominence in isochore theory [10] to the identification of conserved GC-rich patterns in large-scale evolutionary genome databases [5, 71], GC-content (relative to sequence length, also known as G+C%) is highly indicative of function, in part because of its prevalence in and around coding regions. We compute it as the summed percentage of guanine and cytosine bases in the input sequences.

***k*-mer frequency SVR** We adopt a straightforward linear-kernel SVR on 5-mer frequencies computed from the input sequences [25]. As support vector machines are a special case of perceptrons, this model can be interpreted as a one-layer CNN with a kernel size of five. We use the liblinear sci-kit learn implementation [52, 22] with the maximum training sample size supported on a moderate machine.

***k*-mer frequency *k*NN** We also evaluate a more complex albeit less statistically efficient class of models, nearest neighbor graphs, on the same 5-mer frequency representations of sequences as above. We use the GPU-accelerated FAISS library [34] to permit the *k*NN model to scale to millions of training points per task.

3.2 Existing Neural Baselines

As `GUANinE` is a supervised benchmark, we include a handful of pretrained multi-task genomic AI models in our baselines. We perform feature extraction on the output of each model, where each element of that output is a predicted experimental measurement (e.g. DNase hypersensitivity, TF binding) for that sequence in one cell type (or line). We pass these features into an L2-regularized layer, as in linear evaluation [84, 13]. Since we do not fine-tune the models, this performance metric is meant to evaluate `GUANinE`'s tasks and the models' few-shot performance, rather than to comprehensively score their architectures or fine-tuning potential. [3]

DeepSEA DeepSEA takes in a 1000-bp sequence and maps it to a vector of 919 output tracks. The model is fast and shallow, with the learned representation from the convolutional layers unraveled into a 50880-dimensional vector and fed into a large, dense layer with over 89% of DeepSEA's 52.8 M parameters.

[3] Because DeepSEA, Beluga, and Basenji2 all had chromosome 22 present in their training data, while we use it for evaluation of the dnase- and ccre-propensity tasks, their performance on these tasks may be exaggerated (as our targets are roughly summary statistics of their training data across cell types).

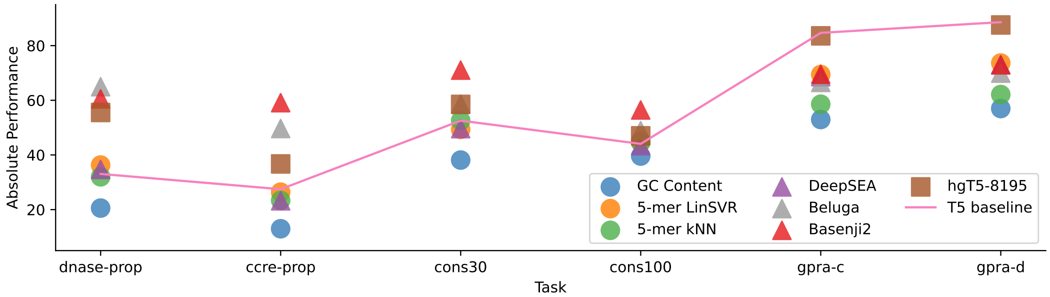


Figure 1: Performance of different methods. Note the T5 and hgT5-8195 use ≤ 512 bp of input sequence.

Beluga Beluga is an enhanced version of DeepSEA. It has more layers, an increased input size of 2000 bp, and uses 2002 experimental tracks for training. Similar to DeepSEA, the learned representation from the convolutional layers is fed into a dense layer that contains over 90% of Beluga's 149.5 M parameters.

Basenji2 Basenji2 is a deeper model prioritizing longer sequence contexts than DeepSEA or Beluga, and it relies heavily on wide, dilated convolutions to reduce its parameter and FLOP counts. Basenji2 was trained on 5313 human *and* 1643 mouse tracks, including many gene expression (CAGE) tasks unlike the purely epigenomic tracks of DeepSEA and Beluga. We calculate Basenji2's effective input size to be 54,784 bp[4]. The dense output layer contains over 26% of Basenji2's 30.1 M parameters.

3.3 Novel Neural Baselines

We also train large (770 M parameter) transformers on each task, providing a low-context comparison to the longer length baselines above. We use the text-to-text transfer transformer (T5) architecture, an encoder-decoder transformer designed for efficient self-supervised pretraining and transfer learning [59]. We evaluate the impact of tokenization and language modeling in T5 variants as outlined below.

T5 Compared to the more common encoder-only BERT and RoBERTa [19, 45], the T5 utilizes a more efficient span-corruption task for language modeling enabled by decoupling the input and output length via decoding. We use single nucleotide token inputs for all results marked 'T5.'

T5-515, -2051, -8195 These models correspond to a T5 architecture with tokenized inputs via a unigram language model [42]. The vocabulary sizes evaluated are 512+3, 2048+3, and 8192+3 tokens, where the +3 represents the <PAD>, <EOS>, and <UNK> tokens.

hgT5-515, -2051, -8195 These models are identical to the T5-515, -2051, and -8195 above except that they undergo self-supervised pretraining on a repeat-downsampled subset of hg38 (1.64 Gbp) before fine-tuning. Because context size is vital factor in language model perplexity [19, 14], we stipulate tokenization [42] for pretraining. See Appendix E for details.

4 Results and Key Findings

We briefly provide key findings from and discussion of our supervised baselines in Tables 2 and 3. A comprehensive table of all model performance metrics, including context ablation studies, is included in Table 5 of the Appendix. We have loosely ordered results in the tables by model complexity/input length, sectioned by pretraining regime, if any. DeepSEA, Beluga, and Basenji2 are progressively more heavily supervised (in terms of using more tracks during training) and have successively larger input context sizes.

4.1 Functional element tasks

Our dnase-propensity and ccre-propensity tasks demonstrate gradated performance with increasing model complexity and context size. In Table 2 we see that Beluga (most parameters) and Basenji2 (largest context size) have the strongest performance for both the dnase- and ccre-propensity tasks. The middling performance of DeepSEA, the Linear k-mer frequency SVR, and the T5 suggest a high degree of intrinsic difficulty for these tasks, with adequate complexity for benchmarking pretrained models. Compared to common metrics such as average auPRC or auROC [86, 87, 38, 6], our propensity score metrics allow for tissue-agnostic predictions from sequence and face few issues of variable class-balance across tracks[5] [39].

[4] During runtime, Basenji2 parallelizes its receptive field over a much longer, 131 kbp sequence.

[5] auROC is insensitive to class balance and can be misleading for highly imbalanced datasets (such as epigenomic tracks), while auPRC is sensitive but incomparable across different class balances.

	dnase-	ccre-	conservation		gpra-c	gpra-d
	propensity		30	100		
Model	ρ	ρ	ρ	ρ	ρ	ρ
GC-content	20.5533	12.9891	38.1307	39.5828	52.9688	56.9824
5-mer LinSVR	36.3022	26.3708	49.3504	45.0548	69.4348	73.6469
5-mer kNN	31.9876	23.3579	52.5776	44.5351	58.5608	62.1082
⋆ T5	33.0548	27.4643	52.5576	44.0542	**84.6738**	**88.5912**
DeepSEA	34.6927	23.2158	49.6163	43.4004	68.6607	72.9729
Beluga	**64.9577**	49.6688	58.6143	48.8816	66.5479	69.9363
Basenji2	60.5149	**56.4568**	**71.0417**	**59.1229**	69.5484	72.9872

Table 2: Performance of non-neural and supervised baselines on `GUANinE`'s test set. Best score per task along with close scores (if any) are **bolded**. These results are also presented visually in Figure 1.

Annotation vs understanding Digging deeper into our functional element tasks (Table 3), we note an interesting distinction between the dnase-propensity task, for which Beluga has the top performance despite its relatively limited 2 kbp input length, and the DHS-subtask of our ccre-propensity task, for which Basenji2 outperforms due to its additional context. Interestingly, the dnase-propensity task performance is less dependent on context size than the DHS-subtask. Recall that our dnase-propensity task asks a model to predict relative accessibility for an *arbitrary* genomic sequence, while our ccre-propensity task asks relative accessibility for candidate-cis regulatory elements (a more difficult task focused on *understanding* the cell type specificity of sequences that all have some propensity for accessibility by virtue of being cCREs). Beluga's model complexity is better able to accurately recognize and annotate local accessibility, but compared to Basenji2, it lacks distal regulatory context that informs potential cell type specificity.

	dnase	DHS
	-propensity	-subtask
Model	ρ	ρ
GC-content	20.5533	13.0236
DeepSEA	34.6927	22.7526
- low-context	34.1037	19.5812
Beluga	**64.9577**	44.5226
- low-context	**62.4819**	**39.9648**
Basenji2	60.5149	**56.4814**
- low-context	58.3123	36.1738
T5-2051	33.8244	28.5074
hgT5-2051	56.2698	39.2990

Table 3: Comparison of dnase-propensity with the DHS-subtask of ccre-propensity. Best *and* best low-context (512 bp) scores are **bolded**.

We also see the benefits of additional supervision (i.e. multi-task training) when comparing DeepSEA, Beluga, and Basenji2 to the T5-2051 model on these tasks. Despite its much larger parameter count (770 M), the T5 is on par with the smaller DeepSEA (50 M) for dnase-propensity, although the additional parameters aid in the DHS-subtask. The significant jump in performance for both tasks with self-supervised pretraining (hgT5-2051) suggests the issue is not with the T5 architecture, but rather a data bottleneck on the original task without additional (self-)supervision. In fact, with self-supervised pretraining, hgT5-2051 nearly competes on these functional element tasks with Beluga's highly informative 2002-track multi-task training.

4.2 Conservation tasks

We see limited improvement with additional supervision or model complexity on the cons100 task, where Basenji2 (with the largest sequence context) has the best (modest) performance, suggesting a performance bottleneck, or possibly even a hard ceiling, on this task. It may be that limited conservation information at the vertebrate scale can be inferred from human genome sequence alone (and thus, if desired for downstream tasks, it should be passed as ancillary input [20]). The cons30 task appears more tractable, with both Beluga and Basenji2 outperforming less complex (non-neural and DeepSEA) or less supervised (T5) models. Basenji2, in particular, appears to have a strong implicit representation of primate sequence conservation, perhaps from its joint training on the mouse genome, which may boost its overall performance. We view conservation estimation as a promising area for benchmarking and model development in genomic AI; however, additional formulations of conservation or an emphasis on conserved element comprehension are needed.

4.3 Gene expression GPRA tasks

DeepSEA, Beluga, and Basenji2 have underwhelming performance relative to baselines, which may be a consequence of organismal or technical transfer and distribution shift (short, exogenous yeast sequences rather than long, endogenous human ones). These models were also trained for inter-sequence annotation

	dnase-	ccre-	conservation		gpra-c	gpra-d
	propensity		30	100		
Model	ρ	ρ	ρ	ρ	ρ	ρ
⋆ T5	33.0548	27.4643	52.5576	44.0542	84.6738	**88.5912**
T5-515	34.3998	26.5567	49.7183	41.6013	83.1336	87.4443
T5-2051	33.8244	30.2823	49.8113	39.5893	84.4885	86.9079
T5-8195	32.3782	30.0196	50.2905	40.1706	81.8273	85.7500
hgT5-515	**57.0310**	36.1465	**58.8254**	45.6393	84.4092	**88.5267**
hgT5-2051	56.2698	**39.3459**	**58.6516**	46.7293	**85.1381**	88.1346
hgT5-8195	55.5411	36.7083	58.4973	**46.9948**	83.5787	87.5641

Table 4: T5 and self-supervised hgT5 variant performance on GUANinE's test set. Best reported score per task along with close scores (if any) are **bolded**. The hgT5-8195 is also presented visually in Figure 1.

rather than intra-sequence (variation) understanding, perhaps limiting their maximum performance. The T5's success here confirms that these tasks are tractable despite the esoteric data distribution. [6]

5 Reflections on Models

5.1 Non-Neural Baselines

The 5-mer frequency Linear SVR performs remarkably well on several tasks, outperforming even the T5 on cons100 and dnase-propensity. We believe its success on dnase-propensity is due to its statistical efficiency[7]. Optimizations to k-mer frequency SVRs would likely boost performance slightly higher [25], and we encourage inclusion of similar non-neural baselines in future benchmarks. The 5-mer frequency *k*NN performs comparably on the cons30 and cons100 tasks, with moderate but generally lower performance on the other tasks.

5.2 DeepSEA, Beluga, and Basenji2

Unsurprisingly, the ∼50 kbp of additional context in Basenji2 gives it a strong advantage over DeepSEA, Beluga, and the T5/hgT5 models on GUANinE, especially for ccre-propensity and the conservation tasks. However, Basenji2 is dependent on this context, and underperforms Beluga on several tasks when input size is ablated. Basenji2 retains moderate performance on the conservation tasks without context, possibly due to its bi-organismal training. In contrast, Beluga, while less performant, is less sensitive to input size ablation (Table 5), and it maintains a faster runtime on its moderate context size (2 kbp). DeepSEA is the shallowest of the pretrained models and underperforms relative to the Linear SVR on dnase-propensity and to both the Linear SVR and the T5 on ccre-propensity, highlighting the need for rigorous non-neural baselining in the field. It does, however, perform decently on the GPRA tasks, which do not benefit from large context sizes.

5.3 T5

The T5 models are the largest tested models by parameter count, which may explain their strong performance on the GPRA tasks. However, we note that performance could be further optimized through a more extensive hyperparameter search, or adjustments to our output tokenization (designed for transfer learning). The issue of hyperparameter search for large models is a known problem in AI [35, 59], and until genomic AI progresses it may be difficult to have prior information about the efficacy of different hyperparameters for large models.

6 hgT5 and the impact of self-supervised pretraining

Pretraining with self-supervised language modeling (span corruption) has a strong, positive impact on T5 performance in our benchmark, as seen in the hgT5 scores of Table 4. On GPRA tasks, the loss incurred by tokenization (i.e. T5 vs T5-515, etc) is largely offset by pretraining, while on conservation tasks, particularly the 30-way (primate/mammal) alignment, self-supervision is able to overcome the performance bottleneck seen in the non-pretrained models of Table 2. Language modeling may have an implicit connection to sequence conservation, if conserved elements or motifs tend to be most imputable. On the dnase- and ccre-propensity tasks, pretraining strongly improves performance as well — notably, this improvement is obtained without additional sequence context, in contrast to other pretrained models. This suggests that combining self-supervision with supervised pretraining or fine-tuning may yield greater performance gains.

[6] Vaishnav et al. [76] included a transformer-esque model that after extensive training time scored more highly.

[7] SVRs are much more data efficient than neural nets and perform well on smaller effective sample sizes for tasks

Finally, our pretraining corpus made use of only the human genome; however, the relative identicalness of the human and primate reference genomes and the relative uniqueness of primates [65] versus other genera [71] may yield diminishing returns from pretraining on distally related organisms. For human- (and mouse-)specific tasks, the significant amount of supervised annotations may also limit the benefits of self-supervision, but significant future work on language models in genomic AI is still warranted. Compared to other DNA language models [47, 15, 33], our hg38 pretraining corpus is repeat-downsampled [44] and contains significant human variation [68], which may improve its utility. See Appendix D for implementation details.

7 Future Work

Functional elements We envision a plethora of follow-ups and revisions to our dnase- and ccre-propensity tasks, ranging from more informative summaries of experiments (e.g. SVD) to 'metatissue' propensity scores. We do not anticipate functional element understanding to be 'solved' in the near future, and we plan to include the more difficult task of variant interpretation in functional elements in the creation of future benchmarks.

Conservation The use of alternative metrics such as background selection [46] or explicitly controlling for non-local drivers of conservation (e.g. chromosome size [82], etc) may make this type of task more tractable. The combination of evolutionary data with regions of mutational constraint [79, 65], or observed selection [40], may yield richer notions of conservation, particularly if our dial of evolutionary scale was directly tunable.

Gene expression The GPRA tasks, even after data cleaning, were by far the largest in GUANinE , and strategic subsetting of 'difficult' or 'informative' sequences [62] or multi-task training on gpra-c and -d may make for more efficient tasks. The development of high throughput exogenous promoter expression methods for mammalian cells will also provide benchmarks for transfer learning from human-centric models.

Potential future tasks As genomic AI and our *in silico* evaluation capabilities advance, so too will our ability to rely on smaller scale ($N < 10^4$) benchmarks or finer-grained tasks, or on tasks subject to significant gene-environment interaction or environmental confounding, e.g. GWAS lead SNPs or eQTLs [1]. Splicing [32], taxonomic classification and comparative (meta)genomics [47, 71], mRNA degradation [2], oncogenic or loss-of-function potential, phylogenetic or evolutionary distance estimation [28], distal effects comprehension [36], CpG methylation [4], 3D conformation modeling [85], and promoter expression plasticity [21], among numerous others, are all readily or near-readily available tasks that may prove valuable for future benchmarks.

8 Conclusion

Machine learning in genomics has greatly advanced since the days of ORF detection via Fourier transform [74], and with the increasing use of genomic AI models, we face the need for rigorous model selection and standardized evaluation, as in other AI fields. Centralized, well-documented benchmarks are essential for such practice, and here we present such a prototypical benchmark for genomic AI, GUANinE, with future expansions and refinement in mind. GUANinE offers concise evaluation across multiple tasks for learned DNA sequence representations, and our curation of tasks with large-scale training sets offers ample opportunity for rigorous baselining, fine-tuning, and model selection. We expect benchmarking, alongside efforts in model training, interpretation, and auditing, with the vital critique of both bioethicists and AI ethicists, will continue to shape the field and enable the development of models for biomedical applications and genome interpretation.

Benchmark and Model Availability

GUANinE, as well as baseline models, are available at https://github.com/ni-lab/guanine. Training sets are in both full and few-shot (1%) versions. To reduce the risk of overfitting and ensure identical evaluation, the test set is provided without labels and final scores can be calculated via prediction server, see the repository for instructions. Our hg38 corpus and the T5 and hgT5 models can be found there for fine-tuning or auditing.

Acknowledgments

Thanks to Pooja Kathail, Ruchir Rastogi, Ryan Keivanfar, Aniketh Reddy, and other members of the Ioannidis lab for invaluable recommendations and feedback. Special thanks to Ryan Chung for guidance on genomic pretraining. Compute was generously supported by Google's TPU Research Cloud (TRC, formerly TFRC). This work was partially supported by an NIH Training Grant (T32HG000047), an Okawa Foundation Research Grant, and a grant from the UC Noyce Initiative for Computational Transformation.

References

[1] A. Abdellaoui, C. V. Dolan, K. J. H. Verweij, and M. G. Nivard. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics*, 54(9):1345–1354, Sep 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01158-0. URL https://doi.org/10.1038/s41588-022-01158-0.

[2] V. Agarwal and D. R. Kelley. The genetic and biochemical determinants of mrna degradation rates in mammals. *Genome Biology*, 23(1):245, Nov 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02811-x. URL https://doi.org/10.1186/s13059-022-02811-x.

[3] V. Agarwal and J. Shendure. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep*, 31(7):107663, May 2020.

[4] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome Biology*, 18(1):67, Apr 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1189-z. URL https://doi.org/10.1186/s13059-017-1189-z.

[5] J. Armstrong, G. Hickey, M. Diekhans, I. T. Fiddes, A. M. Novak, A. Deran, Q. Fang, D. Xie, S. Feng, J. Stiller, D. Genereux, J. Johnson, V. D. Marinescu, J. Alföldi, R. S. Harris, K. Lindblad-Toh, D. Haussler, E. Karlsson, E. D. Jarvis, G. Zhang, and B. Paten. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, Nov 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2871-y. URL https://doi.org/10.1038/s41586-020-2871-y.

[6] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, Oct 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://doi.org/10.1038/s41592-021-01252-x.

[7] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, and J. Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, Mar 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00782-6. URL https://doi.org/10.1038/s41588-021-00782-6.

[8] B. Barz and J. Denzler. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6), 2020. ISSN 2313-433X. doi: 10.3390/jimaging6060041. URL https://www.mdpi.com/2313-433X/6/6/41.

[9] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[10] G. Bernardi. Misunderstandings about isochores. part 1. *Gene*, 276(1):3–13, 2001. ISSN 0378-1119. doi: https://doi.org/10.1016/S0378-1119(01)00644-8. URL https://www.sciencedirect.com/science/article/pii/S0378111901006448. Papers presented at the ISME Symposium on Chromosomes: Structure, Function and Evolution, Guananacaste (Costa Rica), 8-12 January 2001.

[11] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

[12] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

[14] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

[15] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. L. Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679. URL https://www.biorxiv.org/content/early/2023/09/19/2023.01.11.523679.

[16] C. G. de Boer and J. Taipale. Hold out the genome: A roadmap to solving the cis-regulatory code. *bioRxiv*, 2023. doi: 10.1101/2023.04.20.537701. URL https://www.biorxiv.org/content/early/2023/04/20/2023.04.20.537701.

[17] C. G. de Boer, E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, and A. Regev. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol*, 38(1):56–65, Jan 2020.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[20] K. K. Dey, B. van de Geijn, S. S. Kim, F. Hormozdiari, D. R. Kelley, and A. L. Price. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nature Communications*, 11(1):4703, Sep 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18515-4. URL `https://doi.org/10.1038/s41467-020-18515-4`.

[21] H. Einarsson, M. Salvatore, C. Vaagensø, N. Alcaraz, J. Bornholdt, S. Rennie, and R. Andersson. Promoter sequence and architecture determine expression variability and confer robustness to genetic variants. *eLife*, 11: e80943, nov 2022. ISSN 2050-084X. doi: 10.7554/eLife.80943. URL `https://doi.org/10.7554/eLife.80943`.

[22] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[23] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347058. URL `https://doi.org/10.1145/3298689.3347058`.

[24] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov 2021. ISSN 0001-0782. doi: 10.1145/3458723. URL `https://doi.org/10.1145/3458723`.

[25] M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLOS Computational Biology*, 10(7):1–15, 07 2014. doi: 10.1371/journal.pcbi.1003711. URL `https://doi.org/10.1371/journal.pcbi.1003711`.

[26] K. Grešová, V. Martinek, D. Čechák, P. Šimeček, and P. 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`.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[28] I. H. Holmes. Historian: accurate reconstruction of ancestral sequences and evolutionary rates. *Bioinformatics*, 33(8):1227–1229, 01 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw791. URL `https://doi.org/10.1093/bioinformatics/btw791`.

[29] R. A. Hoskins, S. Repo, D. Barsky, G. Andreoletti, J. Moult, and S. E. Brenner. Reports from CAGI: The Critical Assessment of Genome Interpretation. *Hum Mutat*, 38(9):1039–1041, Sep 2017.

[30] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C. L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, and W. Sieh. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet*, 99(4):877–885, Oct 2016.

[31] N. M. Ioannidis, J. R. Davis, M. K. DeGorter, N. B. Larson, S. K. McDonnell, A. J. French, A. J. Battle, T. J. Hastie, S. N. Thibodeau, S. B. Montgomery, C. D. Bustamante, W. Sieh, and A. S. Whittemore. FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics*, 33(24):3895–3901, 08 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx534. URL `https://doi.org/10.1093/bioinformatics/btx534`.

[32] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K.-H. Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176 (3):535–548.e24, 2019. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.12.015. URL `https://www.sciencedirect.com/science/article/pii/S0092867418316295`.

[33] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL `https://doi.org/10.1093/bioinformatics/btab083`.

[34] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[35] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL `https://arxiv.org/abs/2001.08361`.

[36] A. Karollus, T. Mauermeier, and J. Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1):56, Mar 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02899-9. URL `https://doi.org/10.1186/s13059-023-02899-9`.

[37] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, 1987. doi: 10.1109/TASSP.1987.1165125.

[38] D. R. Kelley. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology*, 16(7):1–27, 07 2020. doi: 10.1371/journal.pcbi.1008050. URL `https://doi.org/10.1371/journal.pcbi.1008050`.

[39] M. Kim and K.-B. Hwang. An empirical evaluation of sampling methods for the classification of imbalanced data. *PLOS ONE*, 17(7):1–22, 07 2022. doi: 10.1371/journal.pone.0271260. URL `https://doi.org/10.1371/journal.pone.0271260`.

[40] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–315, Mar 2014.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[42] T. Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL `https://aclanthology.org/P18-1007`.

[43] T. Lappalainen, M. Sammeth, M. R. nder, P. A. 't Hoen, J. Monlong, M. A. Rivas, M. lez Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. f, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. sler, A. C. nen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. ó, I. G. Gut, X. Estivill, E. T. Dermitzakis, X. Estivill, R. Guigo, E. Dermitzakis, S. Antonarakis, T. Meitinger, T. M. Strom, A. Palotie, J. F. Deleuze, R. Sudbrak, H. Lerach, I. Gut, A. C. nen, U. Gyllensten, S. Schreiber, P. Rosenstiel, H. Brunner, J. Veltman, P. A. Hoen, G. J. van Ommen, A. Carracedo, A. Brazma, P. Flicek, A. Cambon-Thomsen, J. Mangion, D. Bentley, and A. Hamosh. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sep 2013.

[44] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL `https://aclanthology.org/2022.acl-long.577`.

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

[46] G. McVicker, D. Gordon, C. Davis, and P. Green. Widespread genomic signatures of natural selection in hominid evolution. *PLOS Genetics*, 5(5):1–16, 05 2009. doi: 10.1371/journal.pgen.1000471. URL `https://doi.org/10.1371/journal.pgen.1000471`.

[47] F. Mock, F. Kretschmer, A. Kriese, S. Böcker, and M. Marz. Taxonomic classification of dna sequences beyond sequence similarity using deep neural networks. *Proceedings of the National Academy of Sciences*, 119(35): e2122636119, 2022. doi: 10.1073/pnas.2122636119. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2122636119`.

[48] M. Mukwende. Mind the gap: A clinical handbook of signs and symptoms in black and brown skin. *Wounds UK*, pages 16–16, 2020.

[49] K. Musgrave, S. Belongie, and S.-N. Lim. A metric learning reality check. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV*, page 681–699, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58594-5. doi: 10.1007/978-3-030-58595-2_41. URL `https://doi.org/10.1007/978-3-030-58595-2_41`.

[50] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong. Deepromoter: Robust promoter predictor using deep learning. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00286. URL `https://www.frontiersin.org/articles/10.3389/fgene.2019.00286`.

[51] C. Paik, S. Aroca-Ouellette, A. Roncone, and K. Kann. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.63. URL `https://aclanthology.org/2021.emnlp-main.63`.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[53] K. S. Pollard, S. R. Salama, N. Lambert, M.-A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, A. D. Kern, C. Dehay, H. Igel, M. Ares, P. Vanderhaeghen, and D. Haussler. An rna gene expressed during cortical development evolved rapidly in humans. *Nature*, 443(7108):167–172, Sep 2006. ISSN 1476-4687. doi: 10.1038/nature05113. URL `https://doi.org/10.1038/nature05113`.

[54] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–121, Jan 2010.

[55] A. B. Popejoy and S. M. Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, Oct 2016. ISSN 1476-4687. doi: 10.1038/538161a. URL `https://doi.org/10.1038/538161a`.

[56] R. C. Périer, V. Praz, T. Junier, C. Bonnard, and P. Bucher. The Eukaryotic Promoter Database (EPD). *Nucleic Acids Research*, 28(1):302–303, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.302. URL `https://doi.org/10.1093/nar/28.1.302`.

[57] F. Racimo, S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, Jun 2015. ISSN 1471-0064. doi: 10.1038/nrg3936. URL `https://doi.org/10.1038/nrg3936`.

[58] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[59] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

[60] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

[61] E. Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, Sept. 2018. doi: 10.1162/coli_a_00322. URL `https://aclanthology.org/J18-3002`.

[62] J. Schreiber, J. Bilmes, and W. S. Noble. apricot: Submodular selection for data summarization in python. *Journal of Machine Learning Research*, 21(161):1–6, 2020. URL `http://jmlr.org/papers/v21/19-467.html`.

[63] R. Singh, J. Lanchantin, G. Robins, and Y. Qi. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 08 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw427. URL `https://doi.org/10.1093/bioinformatics/btw427`.

[64] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[65] L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, and K. K. Farh. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet*, 50(8):1161–1170, Aug 2018.

[66] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=qVyeW-grC2k`.

[67] W. L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953. doi: 10.1177/107769905303000401. URL `https://doi.org/10.1177/107769905303000401`.

[68] The 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, Oct 2015. ISSN 1476-4687. doi: 10.1038/nature15393. URL `https://doi.org/10.1038/nature15393`.

[69] The ENCODE Project Consortium. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, Jul 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2493-4. URL `https://doi.org/10.1038/s41586-020-2493-4`.

[70] The ENCODE Project Consortium, J. Moore, and M. e. a. Purcaro. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, Jul 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2493-4. URL `https://doi.org/10.1038/s41586-020-2493-4`.

[71] The G10K Consortium et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, Apr 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03451-0. URL https://doi.org/10.1038/s41586-021-03451-0.

[72] The GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, Oct 2017. ISSN 1476-4687. doi: 10.1038/nature24277. URL https://doi.org/10.1038/nature24277.

[73] The Roadmap Epigenomics Consorstium, et al. . Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb 2015. ISSN 1476-4687. doi: 10.1038/nature14248. URL https://doi.org/10.1038/nature14248.

[74] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics*, 13(3):263–270, 06 1997. ISSN 1367-4803. doi: 10.1093/bioinformatics/13.3.263. URL https://doi.org/10.1093/bioinformatics/13.3.263.

[75] B. L. Trippe, B. Huang, E. A. DeBenedictis, B. Coventry, N. Bhattacharya, K. K. Yang, D. Baker, and L. Crawford. Randomized gates eliminate bias in sort-seq assays. *Protein Science*, 31(9):e4401, 2022. doi: https://doi.org/10.1002/pro.4401. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4401.

[76] E. D. Vaishnav, C. G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D. A. Thompson, J. Z. Levin, F. A. Cubillos, and A. Regev. The evolution, evolvability and engineering of gene regulatory dna. *Nature*, 603(7901): 455–463, Mar 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04506-6. URL https://doi.org/10.1038/s41586-022-04506-6.

[77] S. van Alten, B. W. Domingue, J. Faul, T. Galama, and A. T. Marees. Correcting for volunteer bias in gwas uncovers novel genetic variants and increases heritability estimates. In *medRxiv*, 2022. URL https://api.semanticscholar.org/CorpusID:262032876.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[79] D. Vitsios, R. S. Dhindsa, L. Middleton, A. B. Gussow, and S. Petrovski. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nature Communications*, 12(1):1504, Mar 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21790-4. URL https://doi.org/10.1038/s41467-021-21790-4.

[80] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

[81] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[82] P. D. Waters, H. R. Patel, A. Ruiz-Herrera, L. Álvarez González, N. C. Lister, O. Simakov, T. Ezaz, P. Kaur, C. Frere, F. Grützner, A. Georges, and J. A. M. Graves. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proceedings of the National Academy of Sciences*, 118(45):e2112494118, 2021. doi: 10.1073/pnas.2112494118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2112494118.

[83] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[84] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9.

[85] Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue. Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nature Communications*, 9(1):750, Feb 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03113-2. URL https://doi.org/10.1038/s41467-018-03113-2.

[86] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, Oct 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3547. URL https://doi.org/10.1038/nmeth.3547.

[87] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*, 50(8):1171–1179, Aug 2018.

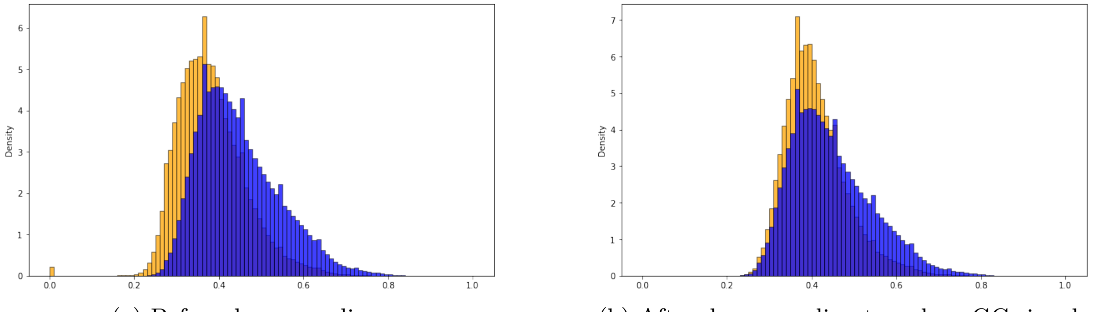
Appendices

A Extended Performance Metrics

	dnase-	**ccre-**		**ccre-subtask**			**conservation**		**gpra-c**	**gpra-d**
	propensity		dnase	ctcf	h3k27ac	h3k4me3	**30**	**100**		
Model	ρ	ρ	ρ	ρ	ρ	ρ	ρ	ρ	ρ	ρ
GC-content	20.5533	12.9891	13.0236	12.6525	13.2115	13.0690	38.1307	39.5828	52.9688	56.9824
5-mer LinSVR	36.3022	26.3708	23.3577	27.6775	27.7758	26.6722	49.3504	45.0548	69.4348	73.6469
5-mer *k*NN	31.9876	23.3579	22.0112	24.5605	23.8530	23.0069	52.5776	44.5351	58.5608	62.1082
⋆ T5	33.0548	27.4643	26.8736	28.2986	27.1943	27.4908	52.5576	44.0542	84.6738	**88.5912**
DeepSEA	34.6927	23.2158	22.7526	24.0649	23.3712	22.6745	49.6163	43.4004	68.6607	72.9729
- low-context	34.1037	20.4241	19.5812	21.2250	20.9838	19.9063	48.1403	42.7819	67.8070	71.9632
Beluga	**64.9577**	49.6688	44.5226	51.6006	52.5047	50.0473	58.6143	48.8816	66.5479	69.9363
- low-context	**62.4819**	**39.8753**	**39.9648**	**40.5294**	**39.6200**	**39.3870**	54.9345	45.6777	65.6781	69.0738
Basenji2	60.5149	**56.4568**	**56.4814**	**56.3126**	**57.3374**	**56.3651**	**71.0417**	**59.1229**	69.5484	72.9872
- low-context	58.3123	35.8789	36.1738	36.6209	35.7707	34.9501	55.4415	46.8741	68.6921	72.3155
T5-515	34.3998	26.5567	23.8884	28.957	26.7359	26.6455	49.7183	41.6013	83.1336	87.4443
T5-2051	33.8244	30.2823	28.5074	31.263	30.9719	30.3869	49.8113	39.5893	84.4885	86.9079
T5-8195	32.3782	30.0196	28.9806	32.3107	27.7989	30.9883	50.2905	40.1706	81.8273	85.7500
hgT5-515	57.0310	36.1465	33.9715	36.9684	37.0397	36.6064	**58.8254**	45.6393	84.4092	**88.5267**
hgT5-2051	56.2698	39.3459	39.2990	**40.6559**	39.3374	38.0914	58.6516	46.7293	**85.1381**	88.1346
hgT5-8195	55.5411	36.7083	35.5757	37.7378	35.9764	37.5434	58.4973	**46.9948**	83.5787	87.5641

Table 5: Omnibus test set scores, including subtask breakout metrics for the ccre-propensity multi-task, and including low-context versions of DeepSEA, Beluga, and Basenji2. Low-context sequences were centered (to a bin, if necessary), truncated to 512 bp of input sequence, and padded with zeroes, which were typically seen during training as Ns. Best reported scores per column, along with the best low-context scores if different, are **bolded**.

Figure 2: GC-content distribution of sequences before and after downsampling negatives (orange) to reduce the difference in GC-content versus positives (blue).



(a) Before downsampling.

(b) After downsampling to reduce GC-signal.

B Additional preprocessing information

B.1 dnase-propensity

We downloaded SCREEN v2 DHS locations from Encode file ENCFF503GCK. We removed 6 assays corresponding to legacy HeLa and A549 cell lines due to age at time of experiment, leaving 727 DNase hypersensitivity assays.[8] We extracted 3.1 M non-zero signal DHSes, then downsampled sequences with more than 25% repeat elements in proportion to their percentage of nucleotides masked by RepeatMasker (higher repeat percentages were more heavily downsampled), which left us with 2.3 M sequences. We augmented these 'positive' sequences with ≈ 400 k 'negative' sequences from the rest of the genome, which were downsampled to reduce the difference in GC-content between positive and negative sequences. The pre- and post-downsampling GC-content distributions of positives and negatives can be seen in Figure 2.

We next constructed our propensity score for each DHS sequence. Specifically, we counted the number of cell types where each consensus DHS was peak-called, then summed this count to a propensity score. We downweighted cancer and immortalized cell lines by ½ to prioritize primary tissues and other biosamples. Negative class sequences were given a propensity of 0 as these regions did not report DNase hypersensitivity. Positive class propensity scores were binned into levels 1 (highly cell-type-specific) through 4 (near-ubiquitous). Our end distribution of classes is approximately 18.8% 0 (negatives), 31.2% class 1, 25.0% class 2, 20.0% class 3, and 5.0% for class 4.

B.2 ccre-propensity

To construct our ccre-propensity tasks, we began with our 2.3 M repeat-downsampled DHS 'positive' sequences from the dnase-propensity task and constructed propensity scores for additional epigenetic signals as well. In particular, we annotated each sequence with a vector of its raw experimental measurements from the SCREEN v2 cCRE registry (available from ENCODE) and constructed propensity scores per epigenetic signal by summing these weighted tracks (as in the dnase-propensity task) for 527 DHS, 198 CTCF, 314 H3K4Me3, and 224 H3K27ac experiments. Compared to the dnase-propensity task, a smaller number of cell types (those for which a second epigenetic experiment was available), are present in the cCRE dataset, so many of the DHSes have zero (0) representative cell lines for each epigenetic signal. We targeted a similar class balance as the dnase-propensity task for classes 2, 3, and 4 for each of the epigenetic signals, with up to 20% of the data for each subtask being negative, 0-labeled DHSes without any signal. This up to 20% negative set was obtained by downsampling those with low GC-content as in the dnase-propensity negatives.

B.3 gpra-c and gpra-d

We downloaded the data from Vaishnav et al. [76] and preprocessed it as follows. Most pertinently, we found that 20-25% of sequences had variable length (non-80bp), and that length was strongly associated with differences in observed gene expression. As the T5 and other architectures can facilely detect sequence length, we chose to prune the non-80 (+30 scaffold) length oligonucleotides to reduce inflated performance due to length in our benchmark. These sequences may in fact contain biological meaning, but our goal for this benchmark was to reduce spurious factors [24]. Additionally, while Vaishnav et al. [76] reported floating point y-values (the average per DNA barcode across multiple observations), we found that integerizing the

[8] Future benchmarking work may consider removing the K562 line due to its age as well.

y-values into gene expression bins preserved ∼99.8% of the observed variance, possibly indicative of technical biases during data collection [75]. Future work on statistical correction may be invaluable for finer-grained experiments.

C Baseline methods

Baseline model performances (not including hgT5 language models) are presented in Table 2. Non-neural models were trained using the maximum dataset size on a moderate machine (51 GB). Note that the cons30 and cons100 tasks have relatively dense 5-mer frequency tables, so we used a 512-dimensional PCA projection before fitting the *k*NN or Linear SVR on those tasks. The GPRA task representations are sparser, although for memory constraints we only used half of the sequences during training (11.5 M for gpra-c and 8 M for gpra-d).

Features from the pretrained supervised models (DeepSEA, Beluga, Basenji2) were extracted from 1% of the training data and the final L2-regularized linear model for each was selected based on performance on the development set with validation choices of $\alpha = [0.1, 0.3, 1, 3, 10, 30]$. Basenji2 features were computed as the average of predictions from three consecutive bins centered on the inputs, which led to slightly improved performance on GUANinEcompared to predictions from a single bin. For the four human tasks (non-GPRA), full-context sequence from hg38 was inputted to each model. Apple-to-apple model performance on ablated input sequence lengths (512 bp) is presented in omnibus Table 5. For the yeast (GPRA) tasks, whose exogenous sequences do not come with a natural context, we cross-validated 10 model-length-appropriate fixed scaffolds chosen from promoter contexts in s288c (yeast) and hg38 (human) genomes with 1% of the training data (0.5% for Basenji2). The best performing scaffold was used on our development set and during testing. Finally, we used the experimental track outputs after standardization for each of the DeepSEA, Beluga, and Basenji2 models, although other combinations of layers or intermediate representations may yield higher performance. For the Basenji2 model, we only used the 5313 outputs from the human output head, as GUANinEis meant to primarily evaluate model performance on human sequences, but the combined use of the human and mouse heads of the model may result in higher performance, particularly on the conservation tasks.

D hg38 pretraining corpus

The impact of language modeling on model performance is displayed in Table 4. All T5 scores reported in Tables 3 and 4 are the average of two independent runs to reduce the impact of stochasticity.

For our language modeling, we aimed to:

1. use tokenization to increase our context length within 512-token segments [42, 14];

2. reduce reference bias introduced by training on the reference genome;

3. reduce the frequency of LINE1 and other repeat elements (and phylogenetically proximal genes), as our models prioritize functional genomics rather than mapping or assembly-related tasks, and repeated words or phrases in language corpora are known to reduce training efficacy and model quality [59, 44];

4. augment the limited size (relative to language modeling) of the human genome.

As such, we took the approach of narrowing down the human genome by first removing N-rich sequences (2 or more contiguous Ns), and then further removing repeat-rich regions (> 50% over a 768-bp sliding window) from our corpus. We then removed resulting sequences that were less than 1024 bp in length, as these would not meet our augmentation procedures (below). We finally employed a combination of locality-sensitive hashing (LSH) and blastn deduplication to remove segments with greater than 90% identity (dropping the shorter of the two sequences).

These procedures left us with 1.64 Gbp of genomic sequence before augmentation. Separately, we obtained genetic variants from unrelated individuals in the 1000 Genomes Project Phase 3 [68] with an allele frequency of 1e-3 or greater. We then created 4x and 64x upsampled versions of our 1.64 Gbp corpus (via offset sweeps through segments for each tokenization size), and we augmented each sampling by upsampling minor alleles at random. For a single minor allele of allele frequency $x \leq 0.5$, the resulting upsampled frequency was $f(x) = \sqrt{x(1-x)}$. Any given 10% frequency minor allele had an approximately 30% chance of appearing, independent of other variants, with similar upsampling when multiple minor alleles were present to increase entropy (maximum entropy for minor alleles would be 50% frequency for one minor allele, 33% each for two minor alleles, etc, such that the major allele is nearly on par in terms of frequency). The 4x versions were used for training the ULM tokenizer [42]. After tokenization, we removed any sequences less than 128 tokens in length, which made our higher token corpora slightly smaller and more prone to overfitting. In the future, we suggest including additional naturally occurring human variation, perhaps via training on pangenomes, with the intent to reduce possible algorithmic bias of pretrained models in genomic AI [11, 55].

E T5 and hgT5 (pre)training

All T5 models and pretrained hgT5 models were finetuned on each task separately, with a batch size of 2^{16} tokens and a learning rate of 1e-4 for 2^{18} steps. We used the checkpoint corresponding to the best development set performance for testing. Given the interrelatedness of human genome sequences, larger batches and/or diversity-based learning rates per minibatch may be warranted.

Self-supervised pretraining was done with a 15% noise density and mean span length of 3 tokens. We tested multiple distinct configurations for hgT5 self-supervised pretraining, and we converged on a batch size of 2^{17} tokens for 2^{17} steps with the standard learning rate schedule, although we did observe some overfitting in later steps with higher tokenization (likely due to seeing the corpus a second time, something uncommon in language applications [59]). During pretraining, we created three replicates of each hgT5 model, but to reduce the risk of overfitting, we discarded the lowest perplexity model for each tokenization size.

Our hgT5-515, -2051, -8195 models achieve bits per character (BPC) scores of 1.55, 1.51, and 1.48, respectively, on our development set. At the full length of 512-tokens of input, the hgT5-8195 model achieves a BPC of 1.46. For comparison, a tetragram model with backoff achieved a BPC of 1.76 [37], while Zaheer et al. [83] achieves better BPC albeit on unreleased, large-context language models with greater tokenization and numerous repeats in their corpus.

Worth noting is that our choice of Unigram Language Modeling [42] instead of the more popular byte-pair encoding was based in conservativeness about benign variation; the ULM tokenizers feature slight redundancy as a form of regularization. We see this is the case for our corpus, as tokenization increases the BPC for input DNA sequences from 2.0 to ≈ 2.09 across lengths (the starting point for our BPC during pretraining).