GUANinE v0.9: 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 v0.9 primarily focuses on functional genomics tasks such as functional element annotation and gene expression prediction, but 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 [59, 5] to neural network-based sequence-to-function maps (i.e. $f : X \to y$) for [epi]genomic annotation, imputation, and understanding [88, 7, 89, 39, 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 [63] 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

Preprint. Under review.

B) ensure statistical significance while having limited or quantifiable confounders (e.g. batch effects or socioeconomic factors [79]), a requirement of any evaluatory dataset [25].

GUANinE v0.9 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 [39, 21, 37]. Although GUANinE does not cover every domain of genomic AI, e.g. taxonomic classification or comparative biology [49, 73], 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 [69] to Natural Language Understanding (NLU) [13, 82].

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 [15] 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 [19], which inspired AlexNet [42] and ResNets [28] in vision, to the bilingual parallel corpora that led to the transformer architecture [80, 20] and modern benchmarks of question-answering [62] and comprehensive evaluation [82, 83, 68] 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 [12], perpetuate cultural biases [12, 10], limit model expressivity [53], or yield inaccurate metrics due to duplication errors [9]. Given present and historical biases in genomics [57, 1, 79] and medicine [50], 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 [58] for promoter annotation [52, 34, 85], gene expression data from the Roadmap [75], GTEx [74], and Geuvadis [44] consortia for promoter understanding [65, 32, 3, 89, 39], and GWAS or eQTL variant association datasets [89, 32, 21] or curated clinical variant annotations from HGMD or ClinVar [31, 81, 41, 32, 39] for variant interpretation. Recent small-scale experimental validations make use of techniques such as CRISPRi [6], among other approaches. 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 [30], in contrast, involve benchmarks created for specific multi-submission challenges, but are typically limited in scope, with submitted models tailored specifically to each individual challenge. The recent GenomicBenchmarks paper [27] 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 G/C-content balancing), and comprehensive baselining.

2 GUANinE Tasks

The GUANinE benchmark is designed to be supervised, human (eukaryote)-centric, and well-controlled, with our formulation focusing on large training and evaluation sets. Compared to other AI disciplines, the relative infeasibility of manual human-labelling for genomic annotation is a clear limitation. For GUANinE, great care and consideration were placed in data cleaning, limiting obvious confounders, and (where applicable) negative-sampling or negative set selection. Our suite of tasks is meant to broadly characterize human-relevant genomic complexity and span several, but not all, possible functional genomic task domains.

2.1 Functional Elements

Endogenous functional element annotation is commonly used for the training and evaluation of genomic AI methods [52, 27]. 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 [88, 89, 39, 6] into a concise, interpretable metric of cell type specificity.

Task	Gene Expression		Conservation		Functional Elements	
	gpra-c	**gpra-d**	**cons30**	**cons100**	**dnase-prop**	**ccre-prop**
Train set	23 M	16 M	5.4 M	5.0 M	2.8 M	7.1 M
Dev set	230 k	170 k	55 k	51 k	35 k	91 k
Test set	230 k	170 k	55 k	51 k	44 k	101 k
Split	Random	Random	Random	Random	Dev21/Test22	Dev21/Test22
Organism	Yeast	Yeast	Human	Human	Human	Human
Min. Length	80 (110)	80 (110)	512	512	511	509
Target	Scalar	Scalar	Scalar	Scalar	Scalar	Multi-task
Output range	0 - 17	0 - 17	0 - 24	0 - 24	0 - 4	0 - 4

Table 1: Benchmark summary statistics.

Compared to existing functional element datasets, these tasks have a relatively low ($< 20\%$) number of negatives (zeros) and stricter downsampling in repeat-masked regions.

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 [71], a collection of several hundred cell type tracks from ENCODE [72]. 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 G/C 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 [71]. 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 [66].

2.2 Conservation

Sequence conservation across evolutionarily related organisms suggests the presence of negative selection against deleterious variation and thus biological function [41, 81]. Many per-base or per-element conservation scores can be directly computed from multiple sequence alignments across related organisms [5, 56] (though these alignments may induce bias due to evolutionary divergence). Human Accelerated Regions (HARs [55]) 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. [56]. 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,

[2]The 2 bp decrease allows for length-512 models to pass a 'task-code' token as input [60]

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.

2.3 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 [89, 3, 65, 2]. However, the space of naturally occurring promoter sequences in any one organismal context is limited [78, 18], both in sample size relative to complexity (∼30,000 in humans [58]) and in diversity of sequences (e.g. phylogenetically related or constrained promoters, and similar G/C gradient patterning across genes [73]). Experimental techniques such as MPRAs or oligonucleotide assembly offer the means to perturb regulatory element motif grammars and add sequence diversity [78, 17] 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 [18]. 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. [78]. 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 [24, 51], due to the presence of confounding sequence (e.g. dinucleotide frequency [8]) 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 G/C content, a surprisingly strong predictor due to its correlation with sequence function in vertebrates and other species [73], to k-mer frequency-based models, which have an established record of success in genomics [26].

G/C content From its original prominence in isochore theory [11] to the identification of conserved G/C-rich patterns in large-scale evolutionary genome databases [5, 73], G/C content (relative to sequence length) 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 [26]. 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 [54, 23] 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 [35] to permit the kNN 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 corresponds to a prediction of an experimental measurement of activity (e.g. DNase hypersensitivity, TF binding) at that sequence for one cell type (or line). We pass these features into an L2-regularized layer, as in linear evaluation [86, 14]. Since we do not fine-tune the models themselves, this performance metric is meant

Model	dnase-propensity	ccre-propensity	conservation 30	conservation 100	gpra-c	gpra-d
	ρ	ρ	ρ	ρ	ρ	ρ
G/C 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	**59.1229**	**71.0417**	**56.4568**	69.5484	72.9872

Table 2: Performance of non-neural and supervised baseline models on GUANinE's test set. The best score per task along with close scores (if any) are **bolded**.

to evaluate GUANinE's tasks and the few-shot performance of these models, 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.

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 benchmarking task, providing a contextless comparison to the longer length supervised 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 via output tokenization [61]. 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 [20, 47], 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 [43]. 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 [20, 15], we stipulate tokenization [43] 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

[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).

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

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 [88, 89, 39, 6], our propensity score metrics allow for tissue-agnostic predictions from sequence and face few issues of variable class-balance across tracks[5] [40].

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.

Model	**dnase -propensity** ρ	**DHS -subtask** ρ
G/C 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 the dnase-propensity task with the DHS-subtask of ccre-propensity. The best *and* best contextless scores are **bolded**.

We can 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 that the issue is not with the T5 architecture, but rather a performance 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 [21]). 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 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

[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.

Model	dnase-propensity	ccre-propensity	conservation 30	conservation 100	gpra-c	gpra-d
	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. The best reported score per task along with close scores (if any) are **bolded**.

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. Further optimizations to k-mer frequency SVRs would likely boost performance slightly higher [26], and we encourage inclusion of this non-neural baseline in future benchmarking efforts. The 5-mer frequency kNN 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 the conservation and ccre-propensity tasks. However, Basenji2 is highly dependent on this broad context, and mildly underperforms Beluga on many tasks when the context is ablated. Basenji2 still retains moderate performance on the conservation tasks without context, possibly due to its heavy supervision or bi-organismal training. In contrast, Beluga, while not quite as performant across tasks, is less sensitive to input ablation on most tasks (Table 5), and it maintains a faster runtime on its moderate context size (2 kbp). DeepSEA is the oldest and shallowest of the pretrained models, and itunderperforms relative to the Linear SVR on dnase-propensity and to both the Linear SVR and the T5 on ccre-propensity, highlighting the need for more rigorous non-neural baselining in the field. Conversely, it performs decently on the GPRA tasks, which do not benefit from increased 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 [36, 61], 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 model 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 — most notably, this improvement is obtained without any additional context, in contrast to the other pretrained models. This suggests that

[6] Vaishnav et al. [78] included a transformer-esque architecture in their evaluation that after *extensive* training time scored particularly highly on this task.

combining self-supervised language modeling with supervised pretraining or fine-tuning is likely to yield even greater performance gains.

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 [67] versus other genera [73] 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 [49, 16, 34], our hg38 pretraining corpus is repeat-downsampled [45] and contains significant human variation [70], which may improve its utility. See Appendix D for implementation details.

7 Future Work

Functional elements We envision a plethora of possible 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 for specific tissues. 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 within functional elements in the development of future benchmarks.

Conservation The use of alternative conservation-related scores such as background selection [48] or explicitly controlling for non-local drivers of conservation (e.g. chromosome size [84], conserved closed chromatin marks, etc) may make this type of task more tractable. The combination of evolutionary data with regions of mutational constraint [81, 67], or observed selection [41], may yield richer notions of conservation, particularly if our dial of evolutionary scale was directly tunable (i.e. a hyperparameter).

Gene expression The GPRA tasks, even after data cleaning, were by far the largest in `GUANinE` , and strategic subsetting of 'difficult' or 'informative' sequences [64] 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 [33], taxonomic classification and comparative (meta)genomics [49, 73], mRNA degradation [2], oncogenic or loss-of-function potential, phylogenetic or evolutionary distance estimation [29], distal effects comprehension [37], CpG methylation [4], 3D conformation modeling [87], and promoter expression plasticity [22], among numerous others, are all readily or near-readily available tasks that may prove valuable for future benchmarks.

8 Conclusion

Machine learning methods in genomics have greatly advanced since the early days of ORF detection via Fourier transforms [76], and with the increasing demand for 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 evaluation. Here we present a prototypical benchmark, `GUANinE`, for the genomics field, with future expansions and refinement in mind. We offer concise evaluation metrics for learned DNA sequence representations across multiple distinct categories of tasks, and our curation of evaluation sets with large-scale training data offers ample opportunity for rigorous baselining, fine-tuning, and architecture selection. We expect that benchmarking, alongside efforts in model training, interpretation, and auditing, and with the vital critique of both bioethicists and AI ethicists, will continue to shape the field, enabling the development of powerful models for genome interpretation and medical applications.

Benchmark and Model Availability

`GUANinE`'s training and development sets, as well as code to train our baselines, are available at https://github.com/ni-lab/guanine. Training data are downloadable in 1% (few-shot) and full versions. Testing data is available upon request. Our hg38 corpus and many of the T5 and hgT5 models can also be found there for auditing and fine-tuning.

Acknowledgments

Thanks to Pooja Kathail, Ruchir Rastogi, Ryan Keivanfar, Aniketh Reddy, and other members of the lab for their invaluable recommendations and feedback. Special thanks to Ryan Chung for initial guidance on

pretraining in the genome. Compute resources generously supported by Google's TPU Research Cloud (TRC), formerly TensorFlow Research Cloud (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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Z. Li, E. Wallace, S. Shen, K. Lin, K. Keutzer, D. Klein, and J. Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. 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 5958–5968. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/li20m.html.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

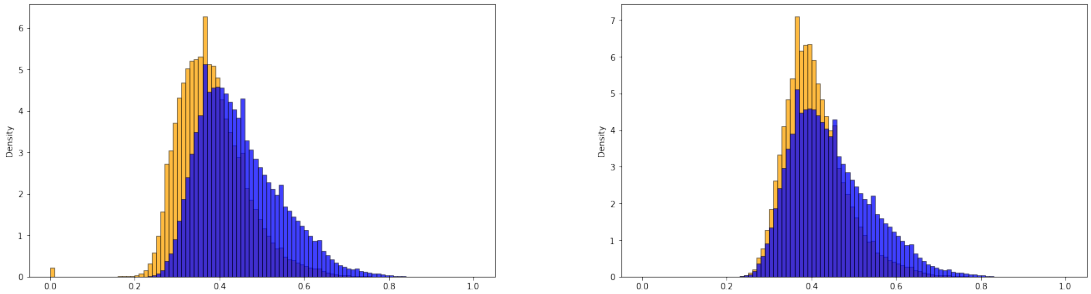
Appendices

A Extended Performance Metrics

Model	dnase-propensity	ccre-propensity	ccre-subtask dnase	ccre-subtask ctcf	ccre-subtask h3k27ac	ccre-subtask h3k4me3	conservation 30	conservation 100	gpra-c	gpra-d
	ρ	ρ	ρ	ρ	ρ	ρ	ρ	ρ	ρ	ρ
G/C 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 kNN	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) and padded with zeroes, which were seen during training as Ns. The best reported scores per column, along with the best low-context scores if different, are **bolded**.

Figure 1: G/C distribution of sequences before and after downsampling negatives (orange) to reduce the difference G/C signal versus positives (blue).



(a) Before downsampling.

(b) After downsampling to reduce G/C signal.

B Additional preprocessing information

B.1 dnase-propensity

We downloaded SCREEN v2 DHS locations from Encode file ENCFF503GCK. We remove 6 legacy cell lines corresponding to HeLa and A549 cell lines due to age at time of experiment, leaving 727 DNase hypersensitivity assays.[7] We extract 3.1 M non-zero signal DHSes, then downsample sequences with more than 25% repeat elements in proportion to their percentage of nucleotides masked by RepeatMasker (higher repeat percentage = more heavily downsampled), which left us with 2.3 M sequences. We augment these 'positives' with ≈ 400 k 'negative' sequences from the rest of the genome, downsampled to reduce the G/C signal. The pre- and post- G/C balancing distribution of positives and negatives can be seen in Figure 1.

We next construct our propensity score and label each DHS sequence by it. Specifically, we count the number of cell types where each consensus DHS was peak-called, then sum this count to a propensity score. We downweigh 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 specific) through 4 (near-ubiquitous). To maintain some semblance of class balancing, given the rarity of ubiquitous DHSes, we do not attempt finely-grained classes, aside from class 4. 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 task, we begin with our 2.3 M repeat-downsampled DHS 'positive' sequences from the dnase-propensity task. We then annotate each sequences with a vector its raw experiment inputs for the SCREEN v2 cCRE registry (available from ENCODE), before constructing propensity scores per epigenetic signal by summing these weighted tracks as in the dnase-propensity task. These propensities for the 2.3 M sequences are based on 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 epigentic signal. We target 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 is achieved by downsampling those with low G/C signal as in the dnase-propensity negatives.

B.3 gpra-c and gpra-d

After downloading the data from Vaishnav et al. [78], we analyzed and decide to further preprocess the data. Most pertinently, we found 20-25% of sequences had a highly variable length — this length was then steeply 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 the inflated performance measure from length in our benchmark. These sequences may in fact contain biological meaning, but our goal for this benchmark was reducing spurious factors [25], rather maximizing biological

[7] K562 remain due to their ubiquity in current analyses, but future benchmarking work may consider sunsetting the line due to its age as well.

meaning. Additionally, while Vaishnav et al. [78] reported floating point y-values, the average per DNA barcode across multiple observations, we found integerizing y-values (gene expression bin) preserved ∼99.8% of the observed variance, possibly indicative of technical biases during data collection [77]. Future work on statistical correction may be invaluable for finer-grained experiments.

C Baseline methods

Baseline model performance (without hgT5 language models) are presented in Table 2. Non-neural models use the maximum dataset size on a moderate machine (51 GB). The cons30 and cons100 tasks have relatively dense 5-mer frequency tables, so we make use of a 512-dimensional PCA before fitting the kNN or Linear SVR on those tasks. The GPRA task representations are sparser, although for memory constraints we only use half of the sequences during training (11.5 M for gpra-c and 8 M for gpra-d).

Features from 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 from performance on the dev set with validation choices of $\alpha = [0.1, 0.3, 1, 3, 10, 30]$. Basenji2 features are the average of three consecutive bins, centered on inputs, which slightly improved performance on `GUANinE`over predictions from a single bin. For the four human tasks, the full-context of hg38 was inputted to each model. Apple-to-apple model performance on ablated inputs is presented in omnibus Table 5. For the yeast (GPRA) tasks, whose exogenous sequences do not come with a natural context, we cross-validate 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 GPRA tasks). The best performing scaffold is used on our development set and during testing.

We use 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 make use of the 5313 outputs from the human output head as `GUANinE`is 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.

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

D hg38 Pretraining Corpus

From the offset of our language modeling, we intended to

1. use of tokenization to increase our context length within 512-token segments [43, 15]
2. reduce reference bias induced by training on the reference genome
3. reduce the presence of LINE1 and other repeat elements (and phylogenetically proximal genes) as we believe our models should prioritize functional genomics rather than mapping or assembly-related tasks[8]
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 eliminating N-rich sequences (2 or more contiguous Ns), then further dividing these subchromosomal segments into subsegments by splicing out repeat-rich regions (> 50% over a 768-bp sliding window) from our corpus. We then removed sequences less than 1024 bp in length as these would not meet our augmentation procedures (below) as well as had suspect lingering repeat features. We finally employed a combination of 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 downloaded AF 1e-3 and greater variants from unrelated individuals in the 1000 Genomes Project Phase 3 [70]. We then created 4x and 64x upsampled versions of our 1.64 Gbp corpus via offset sweep through segments for each tokenization size, and we augmented each sampling by upsampled alleles at random. For example, a 10% frequency minor allele at a locus had an approximately 30% chance of appearing, independent of other loci, with similar upsampling to increase entropy (50% for one minor allele, 33% for two, etc). The 4x versions

[8] additionally, repeated words or phrases in language corpora are known to reduce training efficacy and model quality [61, 45]

were used for training the ULM tokenizer [43]. After tokenization, we removed any sequences less than 128 tokens in length, which concomitantly made our higher token corpora slightly smaller and more prone to overfitting.

In the future, we would suggest including additional naturally occuring human variation, perhaps via training on pangenomes with the intent to reduce possible algorithmic bias of pretrained models in genomic AI [12, 57].

E T5 and hgT5 (pre)training

All T5 models and pretrained hgT5 variants 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 use 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 tried multiple distinct configurations for hgT5 self-supervised pretraining, and will briefly describe our arrived upon language modeling approach. Models trained on our augmented (though repetitive) corpora were prone to overfitting, so to simplify hyperparameter search on our compute budget we opted for the higher parameter T5-large architectures [46] rather than smaller, possibly more portable models. We converged on a batch size of 2^{17} tokens for 2^{17} steps with the standard learning rate schedule, although we do observe some overfitting in later steps of higher tokenization (likely due to seeing the corpus a second time, something uncommon in language applications [61]). 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 [38], while Zaheer et al. [85] 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 [43] instead of the more popular byte-pair encoding was based in conservativeness about benign mutation; 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).