Evaluation of deep learning approaches for high-resolution chromatin accessibility prediction from genomic sequence

Aayush Grover[1,2], Till Muser[3], Liine Kasak[1], Lin Zhang[3], Ekaterina Krymova[3*], Valentina Boeva[1,2,4*]

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland.
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland.
[3]Swiss Data Science Center, EPF Lausanne and ETH Zurich, Zurich, Switzerland.
[4]Université Paris Cité, Institut Cochin, INSERM U1016, Paris, France.

*Corresponding author(s). E-mail(s): ekaterina.krymova@sdsc.ethz.ch; valentina.boeva@inf.ethz.ch;

Abstract

Accurately predicting high-resolution chromatin accessibility signals is crucial for precisely identifying regulatory elements and understanding their role in gene expression regulation. In the absence of experimental assays, machine learning prediction provides an alternative data source to quantify the effects of specific non-coding mutations, thereby accelerating cancer research. While several deep learning methods have been developed to predict chromatin accessibility from DNA sequences including genetic variants, most of these methods either do it at low resolutions or treat this problem as a classification task making it difficult to study the variant effects on chromatin accessibility. In this work, we rigorously evaluated existing deep learning approaches on their ability to predict ATAC-seq signal with the 4bp resolution and assessed the robustness of their predictions. We further introduced a new class of deep learning architectures – ConvNextCNNs, ConvNextLSTMs, ConvNextDCNNs, and ConvNextTransformers, that use ConvNeXt stem to effectively extract genomic features from the DNA sequences. These models outperform the existing methods for predicting high-resolution ATAC-seq signals when compared using data from 2 healthy cell lines, 2 cancer cell lines, and 4 cancer patients, resulting in a diverse experimental setup. Moreover, our study utilized patient-specific data from tumor TCGA samples to analyze the methods' ability to capture changes in chromatin accessibility caused by patient-specific single-nucleotide variants. Based on their predictive accuracy, robustness and ability to predict the effects of single-nucleotide mutations, we observe that ConvNextDCNNs perform better than the other methods. This extensive study opens the door for utilizing these patient-specific deep learning approaches to understand the regulatory landscape alterations caused by specific mutations.

Keywords: Chromatin Accessibility, Variant Effects, Deep Learning, Cancer Epigenetics

1 Introduction

The prediction of chromatin accessibility from DNA sequences is an important task in the study of the mechanisms of different cancers. Accurate data on the chromatin landscape for a given cell type is crucial for downstream tasks including predicting activity of regulatory elements, chromatin organization, and gene expression. Various genome-wide assays have been developed to quantitatively measure chromatin accessibility, including FAIRE-seq [1], DNase-seq [2], and ATAC-seq [3]. At the same time, in order to study the effects of novel mutations in the DNA sequence, these technologies can be resource-extensive. In the absence of experimentally obtained assays for specific mutations, utilizing the predictions of neural networks provides an alternative reliable and efficient data source.

Various deep learning approaches have been proposed to predict chromatin accessibility from DNA sequences. Earlier methods [4–8] focused solely on the classification of a chromatin segment as open or closed. Such problem setting makes it challenging to retrieve the cis-regulatory information within this chromatin segment. Some of the later methods, such as Deopen [9] demonstrated the performance of their model in predicting the maximum intensity of chromatin accessibility corresponding to the input DNA segment. Subsequently, Basenji [10] and Enformer [11] showed that it was possible to predict the intensity of the chromatin accessibility signal corresponding to the input DNA sequence with a relatively lower resolution (128bp) by binning the signal tracks. Although these methods are highly accurate, predicting chromatin accessibility profiles at 128bp (or lower) resolutions is still not sufficient to study local changes in chromatin accessibility caused by point mutations. Moreover, it hinders their applicability for downstream tasks that require high-resolution chromatin accessibility tracks as input such as predicting transcription factor binding [12], cell-type-specific chromatin structure [13–15], and gene expression [16]. Therefore, there is a need for the development of deep learning methods that can accurately predict high-resolution chromatin accessibility signals.

Recently introduced ChromBPNet [17] predicts chromatin accessibility at base-pair resolution using dilated convolutional networks. However, it is not extensively compared against other model architectures, like recurrent neural networks and transformers. GOPHER [18] was introduced as an extensive benchmark that compared various deep learning methods such as Basenji [10] and BPNet [19] in their ability to predict chromatin accessibility for different experimental setups. Both of these methods use dilated convolutional neural networks (dCNNs) to capture long-range information from the input DNA sequence. These methods were evaluated against a convolutional neural network (CNN) baseline highlighting the predictive prowess of dCNN-based methods over CNN-based method. However, recurrent neural network-based methods and transformer-based methods are not explored in this benchmark.

In this work, we create a benchmark to compare the different deep learning models on their ability to predict ATAC-seq signals at 4bp resolution using DNA sequences as inputs. We first test the performance of three architectures, namely Convolution Neural Networks (CNNs), Recurrent Neural Networks (specifically, LSTMs), and Dilated CNNs (dCNNs), which are inspired by existing works, by extensively tuning them

for the task of high-resolution chromatin accessibility prediction. We further improve these architectures by introducing a ConvNeXt V2 [20] block for genomic feature extraction while also introducing a transformer-based architecture for this task. We evaluate these four architectures on their predictive capabilities in accessible regions of DNA as well as on the whole genome. Additionally, we compare the robustness of the predictions from these methods to input shifts. Finally, the sensitivity of these methods is tested in making varying predictions in the presence of a single nucleotide variant (SNV) in cancer patients.

2 Results

2.1 An extensive benchmark of genomic models for high-resolution ATAC-seq prediction

In this work, we created an extensive benchmark of four categories of model architectures (CNNs, LSTMs, dCNNs, and transformers) on their ability to predict the ATAC-seq signal at 4bp resolution. To fairly test all the models, we first defined our experimental setup. Each model used 2048bp of one-hot encoded DNA sequence as input and predicted ATAC-seq corresponding to the central 1024bp genomic region at 4bp resolution (**Figure 1**). Since all transcription factors have a binding motif of greater than 4bp [21], having a resolution of 4bp allows the predicted signal to capture variability arising from all transcription factors. Additionally, having an input of 2048bp provided sufficient context information from the neighborhood of region of interest for a model to make accurate predictions.

Initially, we compared architectures based on existing methods which can be broadly categorized into CNNs [4], LSTMs [7, 9], and dilated CNNs [10, 17] (**Figure 1a**). Since these model architectures were proposed for different experimental setups, we did an extensive hyperparameter search for each of these models on our experimental setup as defined above. We then improved these architectures by applying transformer-inspired ConvNeXt layers instead of convolution layers [20] as genomic feature extractors in the stem of all methods (**Figure 1b**). For CNNs, we also replace the CNN layers in the core block with ConvNeXt layers. Transformer models have shown significant improvements over the convolution-based models in effectively capturing the dependencies between genomic regions [11, 22–24]. However, the application of transformer architecture for high-resolution chromatin accessibility prediction task is missing. Therefore, we introduced a new transformer-based architecture with rotary embeddings [25] for this task (**Figure 1b**). This provided an extensive setup for benchmarking of the four classes of deep learning architectures.

2.2 Model performance on new sequences

Each model was extensively trained and tested on data from four human cell lines – GM12878, K562, IMR90, and HepG2, and four cancer patients – two suffering from colon adenocarcinoma (COAD) and two from kidney renal papillary cell carcinoma (KIRP). We trained a cell-type-specific model since different cell types are regulated by different regulatory mechanisms. Each model was then tested on unseen chromosomes

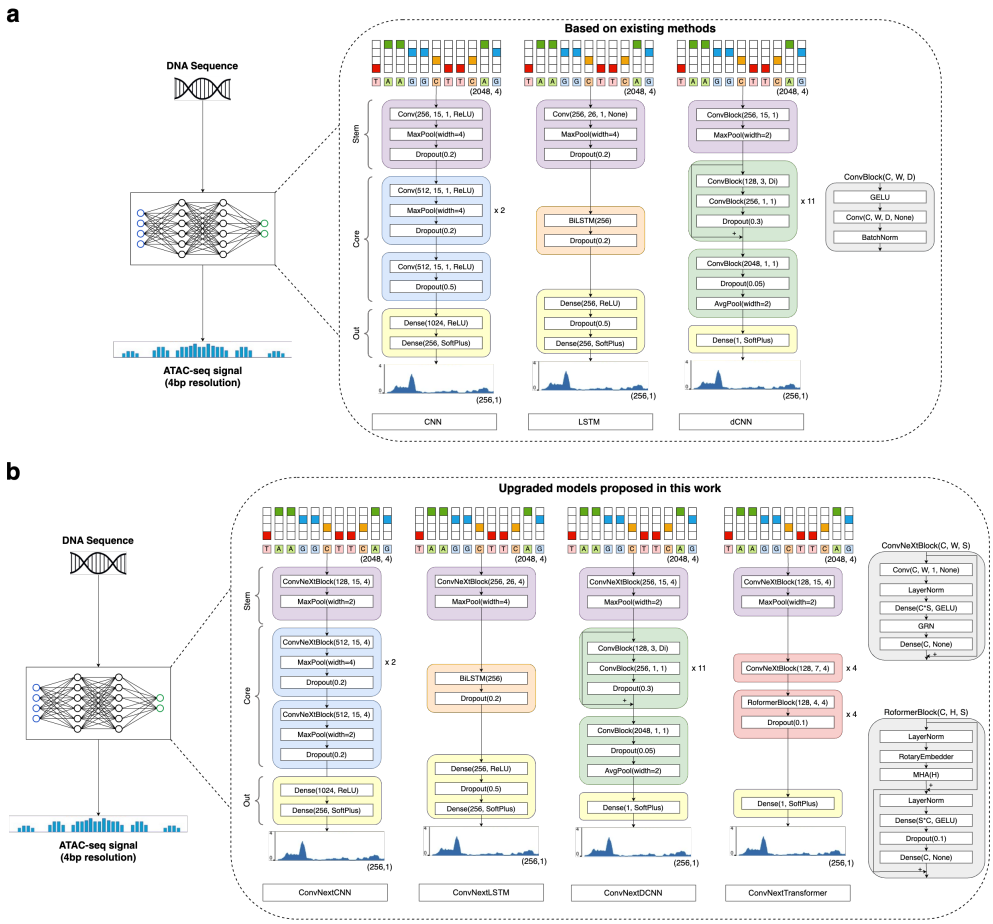

Fig. 1 Overview of the benchmarking setup. Various model architectures are used in this study to predict high-resolution ATAC-seq from DNA sequences. These include CNN, LSTM, and dCNN-based models based on existing works (**a**) as well improved versions of these architectures along with a new transformer-based architecture (**b**).

from the same cell type that was used for training. To provide robust results, we employed 5-fold cross-validation of each model on each cell type/cancer patient where the folds were stratified by chromosomes. The models were compared based on their predictive power on ATAC-seq peaks from each unseen chromosomes as well as on each entire unseen chromosome. The former helped us understand each model's ability to accurately predict ATAC-seq in highly accessible regions whereas the latter also considered all inaccessible regions.

Among the architectures based on existing methods (CNNs, LSTMs, dCNNs), dCNNs outperformed other methods for the task of prediction around accessible regions as well as the whole genome (**Figure 2a**, **Figure S1**). However, upon using the ConvNext in the stem of the models, all models showed high correlations (average

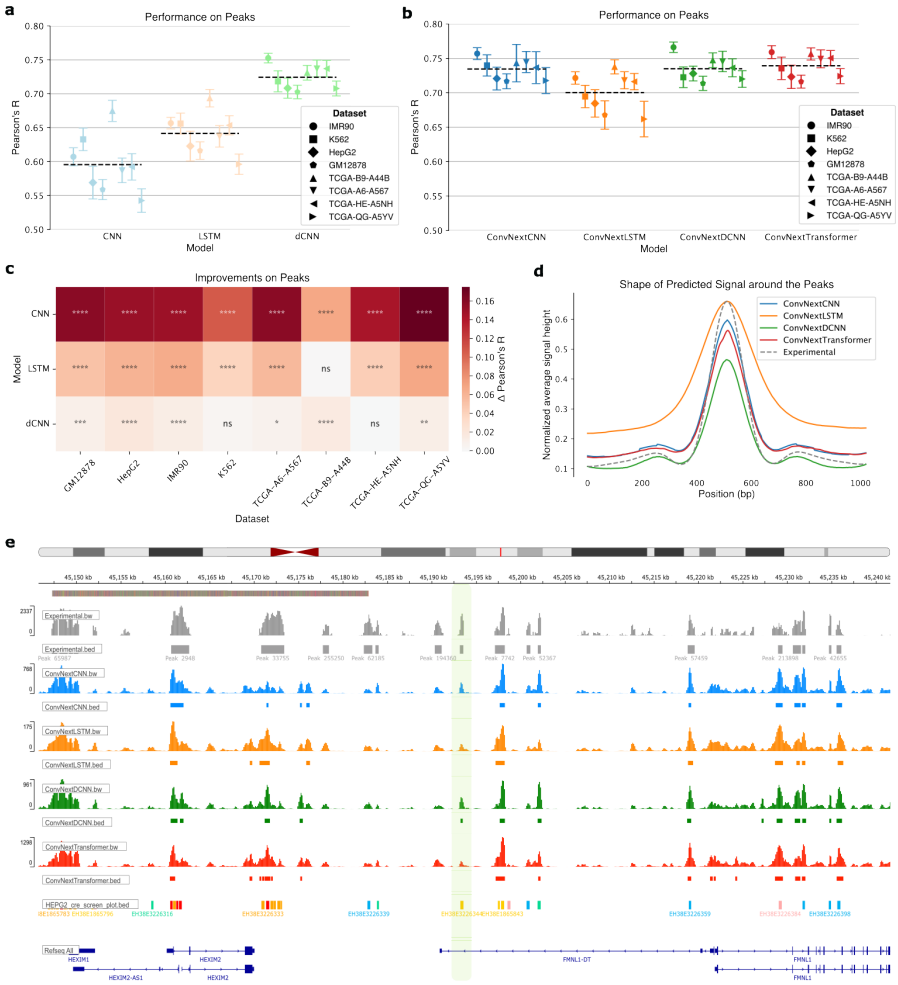

Fig. 2 Model performance comparison across datasets. a, Pearson's correlation between true and predicted ATAC-seq signals in peak regions across eight distinct datasets. Predictions are generated by various existing architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (specifically, LSTMs), and Dilated Convolutional Neural Networks (dCNNs). **b,** Performance comparison among the four proposed methods in this work (ConvNextCNNs, ConvNextLSTMs, ConvNextDCNNs, and ConvNextTransformers) around the peak regions stratified by cell lines and cancer patients. Each point depicts a test chromosome. The black dashed line shows the average performance of a model across all datasets and chromosomes. **c,** Improvements of the new ConvNeXt-based methods proposed in this work around the ATAC-seq peaks as compared to existing methods. The significance is calculated based on two-sided Mann-Whitney U test on Pearson's R calculated for each test chromosome. ****: $P \leq 0.0001$, ***: $P \leq 0.001$, **: $P \leq 0.01$, *: $P \leq 0.05$, ns: $P > 0.05$. The Δ Pearson's R is calculated as the difference between mean Pearson's R across all chromosomes for ConvNeXt-based method and existing method. **d,** The average shape of the predicted ATAC-seq signal is compared against the experimental shape around the peak regions. For each peak, the signal is first min-max scaled to account for different peak heights. **e,** An example of predicted signal vs the experimental signal around the *FMNL1* gene in HepG2 cells. A distal enhancer is highlighted which is only discovered as a peak by the ConvNextDCNN method.

Pearson's R ≥ 0.7) (**Figure 2b**, **Figure S1**). The novel transformer-based architecture, ConvNextTransformer, also performed comparable to the ConvNextCNN and ConvNextDCNN models. Addition of ConvNext to the stem significantly improved the performance of the existing models across all datasets (**Figure 2c**, **Figure S1**) highlighting the better effectiveness of ConvNext layers in capturing genomic features with local and global context.

It is fundamental for these deep learning models to accurately capture the height and the width of accessible regions. We investigated the newly proposed ConvNext models for this property by plotting the predicted min-max normalized ATAC-seq signal around all the peaks across all the datasets and training folds (**Figure 2d**). While ConvNextLSTM showed a tendency to over-predict the signal around the peak and generate smoother signals, other ConvNeXt-based methods accurately captured the nucleosomes around ATAC-seq peak centers closely resembling the experimental ATAC-seq signal. Finally, we export each ConvNeXt-based model's predictions on unseen chromosome as bigwigs. These bigwigs were also used to call peaks for each method. While the produced bigwigs and peaks overlapped with many known promoters and enhancers [26], some enhancers were found accessible only by specific methods. For example, an enhancer in the gene body of *FMNL1-DT* gene in HepG2 cells was only marked as accessible by ConvNextDCNN (**Figure 2e**). Overall, ConvNextCNN, ConvNextDCNN, and ConvNextTransformer were the best performing methods based on correlation with experimental data.

2.3 Robustness of models to input shifts

All models are expected to have robustness to input shifts *i.e.* irrespective of whether the input to the model is peak-centered or not, the model should hold the ability to generalize well and predict accurate ATAC-seq signal. To verify this, we performed the robustness test on all the ConvNeXt-based models, which measures the variation in predicted ATAC-seq signal corresponding to the common region for various shifted inputs [18] (**Figure 3a**). For each ATAC-seq peak across the test chromosomes of all the datasets and training folds, $N = 16$ different input shifts were considered. For the commonly predicted region, we calculated variation score across each 4bp bin. Lower variation score corresponds to higher robustness and vice versa.

ConvNextDCNN was the most robust method across all the datasets followed by ConvNextLSTM (**Figure 3b**, **Figure S2**). Position-stratified variation score confirmed the robustness of ConvNextDCNN which was averaged across all the peaks in different cell types (**Figure 3c**). The periodic behavior of ConvNextCNN can be attributed to the sensitivity of convolutional networks to random input shifts [27]. The robustness of ConvNextDCNN was further confirmed by a low variation score observed using the entire unseen chromosomes of each dataset (**Figure S2**).

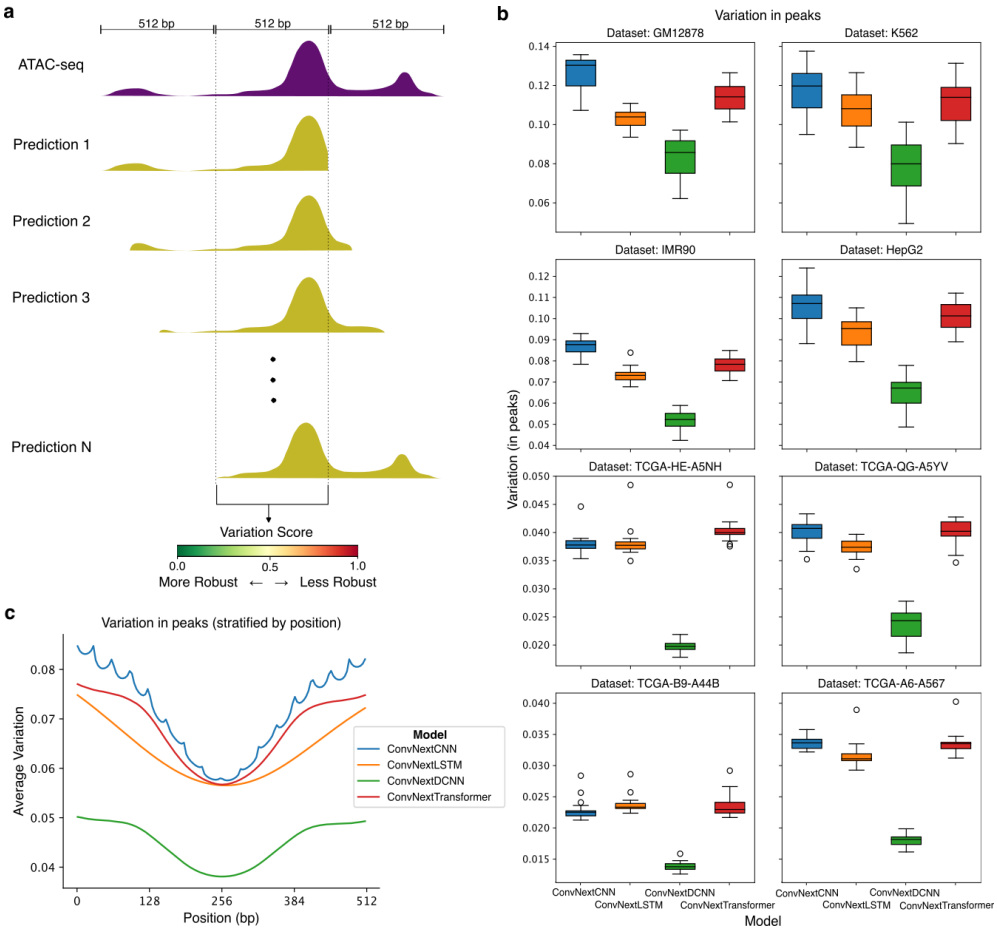

Fig. 3 Evaluation of robustness to input shifts. a, The robustness tests evaluates each method's ability to predict the correct accessibility signal despite small shifts in input. For a given model, the input is shifted by a few base pairs and the common predicted outputs are compared for variation. N such predictions are taken into account. For our experiment, we choose $N = 16$. Variation score is calculated as mean coefficient of variation across all non-overlapping 4bp bins corresponding to the common 512bp genomic region. Lower variation scores suggest high robustness to input shifts. **b,** The variation score measured around peak regions of unseen chromosomes across the 8 datasets and all folds. Four models proposed in this works are compared for robustness. Each point represents variation score for a chromosome unseen by the model during training. **c,** Position-stratified robustness is calculated for all the methods around the peak regions. The variation is averaged across all the peaks in unseen chromosomes of the 8 datasets.

2.4 Sensitivity of models to point mutations

Generating experimental chromatin accessibility profiles for each variant is extremely resource extensive. Therefore, it is pivotal for the ideal model to be able to accurately predict *in-silico* the effect of single-nucleotide variants (SNVs) on chromatin accessibility. We obtained the SNVs for the four cancer patients (two suffering from colon

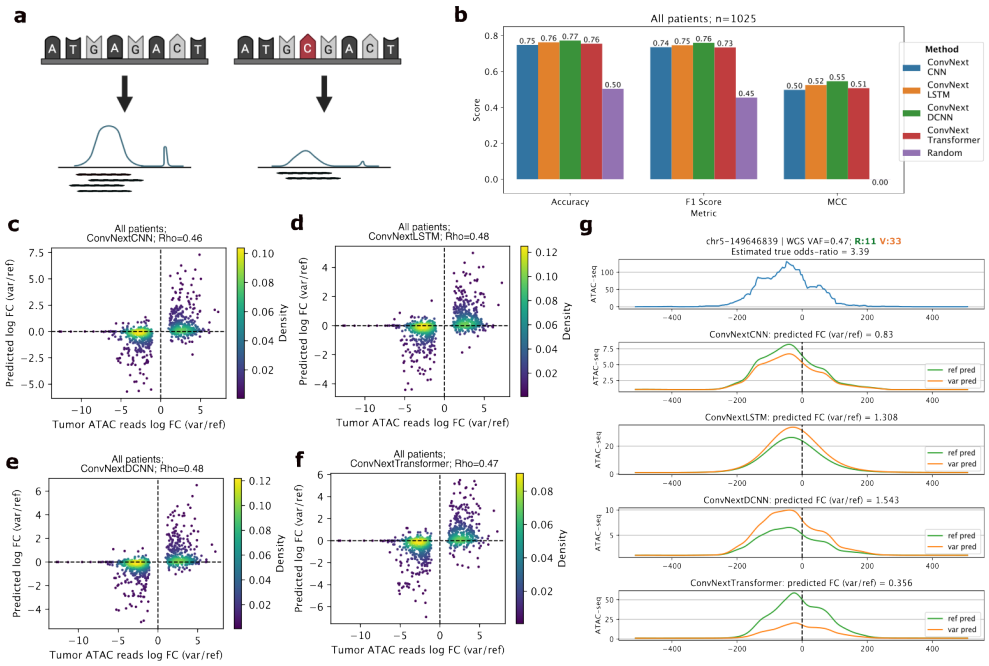

Fig. 4 Sensitivity of models to accurately predict effects of single-nucleotide variants on chromatin accessibility. a, Experimental setup to test each model's ability to predict the change in chromatin accessibility between reference (left) and a variant (right). The image is created using elements from BioRender.com. **b,** ConvNeXt-based models are compared based on their ability to predict the correct directionality of change in accessibility between reference and variant. A total of 1025 variants are considered, across all cancer patients included in this study. Accuracy, F1 score, and Matthew's correlation coefficient are used as metrics. Random baseline is included to show the improvements in metrics by all the models proposed in this work. **c-f,** Scatter plots showing the quantitative change in accessibility by each model. Log-odds-ratio based on ATAC-seq read counts for reference and variants are compared against log-fold-change between model predictions with reference and variant as model inputs and Spearman's correlation (*Rho*) is calculated. **g,** An example of a variant in chromosome 5 is depicted. Whole genome sequencing's variant allele frequency (WGS VAF) is mentioned for the variant. ATAC-seq total reference reads (R) and variant reads (V) are also mentioned. Each model's ATAC-seq prediction for the reference and variant inputs are shown below.

adenocarcinoma and two from kidney renal papillary cell carcinoma) from ICGC [28] and used the ATAC-seq read counts from TCGA for reference and variant nucleotide to determine the change in accessibility (**Figure 4a**). A total of 1025 variants across the four patients were considered for this task. Variants with small change in true odds-ratio (between 0.67 and 1.33) were discarded from this analysis. For this task, all the models were retrained on all chromosomes, omitting only those regions of DNA that contained SNVs. Finally, the predictions were made with SNV-centric 2048bp genomic region as input to each model for reference and variant allele – with the inputs only differing by one nucleotide.

We first assessed each model's ability to correctly predict the directionality of change in accessibility between the reference and variant. All models outperformed

the random baseline. Moreover, ConvNextDCNN achieved the highest performance with an accuracy of 0.77, F1 score of 0.76, and the Matthew's Correlation Coefficient (MCC) of 0.55 (**Figure 4b**). F1 score and MCC take into account the class imbalance and are therefore, are more robust metrics for this task. We then moved to a more challenging task of quantitatively predicting the difference in accessibility for an SNV. We considered the estimated log-odds-ratio using ATAC-seq read counts for the reference and variant as the ground truth for this experiment and this was compared against the log-fold-change between the predicted ATAC-seq signal for the variant vs the reference nucleotide for each model (Methods). ConvNextLSTM and ConvNextDCNN obtained the highest Spearman's correlation with the true odds-ratio ($Rho = 0.48$) (**Figure 4c-f**) with ConvNextTransformer and ConvNextCNN not falling further behind ($Rho = 0.47$ and $Rho = 0.46$ respectively). Finally, we showed an example using an SNV located around 149.6Mb of chromosome 5 where the variant ATAC-seq reads are thrice in number as compared to reference reads (33 vs 11) (**Figure 4g**). Therefore, we expected the models to show higher prediction for the input with variant nucleotide than the one with the reference. Only the ConvNextLSTM and ConvNextDCNN models were able to predict the directionality of change correctly with ConvNextDCNN being closer to the estimated odds-ratio (**Figure 4g**). For several SNVs, while all models predicted the directionality of change in ATAC-seq signal correctly, the predicted signal shape was best predicted by the ConvNextDCNN model (**Figure S3**). In conclusion, while all models performed well on the task of prediction of effect of SNVs on chromatin accessibility, ConvNextDCNN had a slight edge over the other methods.

3 Discussion

In this work, we compared existing deep learning architectures—convolutional neural networks (CNNs), recurrent neural networks (specifically, LSTMs), and dilated CNNs (dCNNs)—on their ability to predict high-resolution chromatin accessibility signals from genome sequences (**Figure 1a**). To improve performance, we proposed new ConvNeXt-based models: ConvNextCNN, ConvNextLSTM, and ConvNextDCNN. Additionally, we introduced a transformer-based model, ConvNextTransformer, to create a more comprehensive set of architectures for this task (**Figure 1b**).

The ConvNeXt-based models outperformed their traditional counterparts in accurately predicting ATAC-seq signals at 4bp resolution across eight different cell types and cancer patients. Among them, ConvNextCNN, ConvNextDCNN, and ConvNextTransformer best captured shapes of the ATAC-seq peaks (**Figure 2**). We further evaluated the robustness of these models to random input shifts. ConvNextDCNN emerged as the most robust, efficiently integrating local and global sequence features through its dilated CNN layers (**Figure 3**).

Finally, we assessed the models' ability to predict changes in chromatin accessibility caused by single-nucleotide variants (SNVs). ConvNextLSTM and ConvNextDCNN achieved the highest accuracy in predicting the log-odds-ratio of ATAC-seq reads between variant and reference alleles. ConvNextDCNN also best captured the directionality of accessibility changes, accurately distinguishing between increases and

decreases in chromatin accessibility upon SNV introduction (**Figure 4**). This highlights its sensitivity to point mutations and suggests its potential utility in studying the functional effects of non-coding genetic variants on chromatin accessibility.

While our ConvNeXt-based models demonstrated strong performance in predicting chromatin accessibility and the effects of SNVs, our models were trained in a cell-type-specific manner. Therefore, they were optimized for individual cellular contexts rather than generalizing across different cell types. While this design ensures high accuracy within a given cell type, it also means that a new model must be trained for each additional cell type of interest. Future work could explore data-efficient fine-tuning approaches to accelerate model adaptation to new cellular contexts without requiring full retraining.

This study highlighted the advantages of ConvNeXt-based models in predicting high-resolution chromatin accessibility and regulatory effects of genetic variants. By combining ConvNeXt layers as genomic feature extractors with CNNs, LSTMs, dilated CNNs, and Transformers, we established a comprehensive framework of sequence-based chromatin modeling. Our results demonstrated that ConvNextDCNN is not only highly accurate but also robust to input shifts and effective at capturing SNV-effects on chromatin accessibility, making it a promising method for genome-wide regulatory analysis.

Methods

ATAC-seq bigwig data

We downloaded the ATAC-seq signal p-value tracks for four cell lines – GM12878, IMR90, HepG2, and K562, from the ENCODE Consortium [29] where the experimentally measured ATAC-seq has been preprocessed with the open source processing pipeline [30]. These cell lines consist of healthy and cancer cells thereby creating a diverse dataset.

To also include datasets from another source, we used ATAC-seq data of two cancer patients suffering from colon adenocarcinoma (COAD) and two patients suffering from kidney renal papillary cell carcinoma (KIRP) in the The Cancer Genome Atlas (TCGA) [31]. These datasets have been made available by the Genome Data Commons (GDC) Data Portal [32]. The choice of these patients was made due to availability of the data on their somatic mutations and single-nucleotide polymorphism (SNP) in the International Cancer Genome Consortium (ICGC) and GDC data portals respectively [32, 33]. Moreover, these patients had high number of combined SNV calls ($> 175,000$) before applying SNV-filtering (described below) and high purity (Consensus Purity Estimate > 0.8) [34].

The ATAC-seq BAM files for the above patient were downloaded from National Cancer Institute's Genomic Data Commons portal [35]. Signal p-value bigwigs were then generated using the ENCODE's ATAC-seq processing pipeline [30]. All used data are listed in Supplementary Table S1. Additionally, ATAC-seq quality measures that were obtained while running the ENCODE pipeline are available for all selected patients in Supplementary Table S2.

Preprocessing

For the ATAC-seq of each cell line, we compose a dataset for each cell line or cancer patient where each item is a pair of one-hot encoded DNA sequence of shape `(2048, 4)` and the corresponding 4 bp resolution ATAC-seq of shape `(512)`. To reduce the ATAC-seq resolution from 1bp to 4bp in the dataset, the values were binned using the maximum value.

Each model is trained on whole genome data, with only samples at genomic locations of unreliable accessibility measurements excluded. Specifically, we exclude any samples that overlap with the blacklisted regions[1] along with samples where at most 35% of the 2048bp input overlaps with unmappable regions[2]. The annotations of blacklisted and unmappable regions for the hg38 genome are obtained from Kelley *et al* [10].

During the evaluation, we impose a tighter constraint on the mappability and only evaluate samples that do not overlap with any unmappable region. Additionally, we use only the predictions corresponding to the central 1024bp genomic section. This ensures that there is at least 512bp of genomic context present on either side of any region that we are trying to predict.

In addition to evaluating performance on the whole genome data, we also compose a dataset of peaks by taking samples centered on the ATAC-seq peaks to compare how models perform on high signal regions.

Model architectures

As a standardized setting, the input and output of all models span 2048 bp, and the resolution of the predicted ATAC-seq is 4bp. Each model is described pictorially in **Figure 1**. The runtime and memory requirements of each model is described in **Figure S1**.

CNN: The convolutional neural network (CNN) architecture is inspired by the DeepSEA [4] model, and modified to accommodate the larger input and output. The model applies three convolutional layers with 512, 640, and 768 filters of size 8. Each convolutional layer is followed by max pooling with a window of 4 and dropout with a rate of 0.2 to reduce the dimensionality and prevent overfitting. The outputs of the convolutional layers are flattened and then fed into two dense layers with 1024 and 512 ReLU-activated units. The CNN architecture is designed to learn and recognize motif patterns of transcription factors, which are indicative of chromatin openness in the input DNA sequence.

LSTM: The architecture of the recurrent model (*i.e.* LSTM) is a modified version of the DanQ [6] model that has been modified similarly to the CNN model. In this model, the data first passes through a convolutional layer with 512 filters of length 26, followed by a max pooling layer and a dropout layer. The core of this network is a bidirectional LSTM [36] layer with 512 units. Similar to the vanilla CNN, this network ends with two dense layers with 1024 and 512 ReLU-activated units. The assumed benefit of this architecture is that the recurrent layers will additionally capture the

[1] https://storage.googleapis.com/basenji_barnyard2/hg38.blacklist.rep.bed
[2] https://storage.googleapis.com/basenji_barnyard2/umap_k36_t10_l32_hg38.bed

longitudinal relations in the DNA sequence, as they can maintain a state or memory over time.

dCNN: Kelly *et al* [10] proposed Basenji, which uses dilated convolutional layers to capture long-range dependencies and complex patterns in the DNA sequence. Pampari *et al* [17] adjusted the dilated CNN (dCNN) for a shorter input sequence of length 2048. With minimal modifications to facilitate the higher output resolution, we have included the model in our study as well. The architecture for this model contains 11 dilated convolutional blocks with residual skip connections, where each block contains two convolutional layers of which the first one uses an increasing dilation rate across successive blocks. The GELU activation function is used in the convolutional blocks. The model ends with a softplus activation function.

We improved the performance of above methods by introducing ConvNeXt V2 [20] blocks as the genomic feature extractor instead of a the traditional convolutional blocks.

ConvNextCNN: The model uses a single ConvNeXt block with 128 filters of size 15 to efficiently capture the genomic information from the one-hot encoded DNA input. This is followed by three additional ConvNeXt blocks with 256, 512, and 512 filters of size 15 where each input channel is convolved using its own set of filters reducing the computational costs [37]. Each ConvNeXt block is followed by a max pooling layer and a dropout layer to reduce the dimensionality and to prevent overfitting. Finally, two dense layers are applied to each channel of the learned feature matrix from the ConvNeXt blocks. These layers contain 512 and 32 neurons respectively with ReLU activation. Softplus activation is added to the end to ensure smooth flow of gradients.

ConvNextLSTM: ConvNextLSTM also uses a ConvNeXt stem to learn efficient genomic features. However, it obtains the best performance when 256 filters of size 26 are used in the stem. The output is passed through a bidirectional LSTM [36] layer with 256 units to learn the effects of distal genomic elements. The LSTM produces a sequences of hidden states, from which only the final state's output is selected as this captures information from the entire input. Two dense layers with 256 and 512 units respectively are finally used to predict ATAC-seq at 4bp resolution.

ConvNextDCNN: The ConvNextDCNN model is very similar to the dCNN model. They only differ in the stem where the ConvNextDCNN benefits from a ConvNeXt stem with 256 filters of size 15.

ConvNextTransformer: Lastly, we constructed a transformer model [38], built around self-attention. This model passes the input through a number of ConvNeXt blocks [20], and pooling layers to extract motifs from the sequence and condense information into meaningful tokens at the output sample rate of 4bp. This sequence of 512 tokens is then passed through a number of transformer blocks. As genomic distances play a major functional role in genomics, the transformer positional encodings need to be designed carefully. Mirroring the behavior seen in current large-language models [39, 40], we observe better performance when using rotary positional encodings (RoPE) [25] compared to conventional sinusoidal encodings proposed in Vaswani *et al* [38]. The `(512, C)` output of the transformer is passed through a $C \times 1$ dense layer, aggregating the information across channels to produce a signal. Finally, we pass this output through a softplus activation function to restrict it to the positive domain.

Training details

We use a Poisson negative log-likelihood loss to train all models with the AdamW [41] optimizer. We begin by linearly increasing the learning rate during the first epoch until we reach a maximum of 1×10^{-3}. We then use exponential decay coupled with cosine annealing [42], restarting at the beginning of every epoch.

Variation score

To calculate the robustness of each model to input shifts, we used the variation score as described in Toneyan *et al* [18]. For each 512bp region in our test dataset, we constructed 16 distinct 2048bp long input sequences such that the output corresponding to each input contains the predictions for the selected 512bp region. Using these 16 predictions, the coefficient of variation is calculated for each 4bp bin. Finally, variation score is obtained by taking the average coefficient of variation across all the bins. For the whole-genome evaluation, this is repeated for all the 512bp regions of the test chromosomes whereas for peak-centered evaluation, only those 512bp regions are selected that contain an ATAC-seq peak.

SNV filtering

We define a single-nucleotide variant (SNV) as either a somatic mutation or a single-nucleotide polymorphism (SNP). For the four cancer patients, we obtained somatic mutation calls from the study of pan-cancer analysis of whole genome (PCAWG) [43]. The SNP calls were obtained from GDC data portal [32].

For each patient, the SNVs were filtered to retain only those SNVs that significantly change the accessibility profiles. First, somatic mutations that did not pass the quality control measures of PCAWG were removed. Similarly, we only selected those SNPs that had a heterozygous genotype and a high confidence score (uncertainty value < 0.1). The filtered files were then combined and total read counts from the whole genome sequencing (WGS) and ATAC-seq experiments were calculated for the reference and variant alleles for each SNV. Using these values, the variant allele frequency (VAF) for each SNV were calculated giving us WGS VAF and ATAC VAF. As we estimate the expected change in ATAC-seq signal for each SNV using the ATAC-seq read counts, we expect the SNVs to be unbiased towards either allele. Therefore, we select SNVs with WGS VAF in the range of $0.4 - 0.6$. Finally, to obtain SNVs that have a significant change in ATAC-seq, we calculate significance of change using Fisher exact test and odds-ratio. Each of these quantities are calculated using the reference and alternate allele read counts in WGS and ATAC-seq experiments. SNVs that had a p-value < 0.01 and odds-ratio $\in [0.67, 1.33]$ were considered to have significant change in ATAC-seq signal and therefore, were selected for SNV analysis.

This filtering led to a total SNV count of 1025 with 80 somatic mutations and 945 SNPs.

Declarations

Funding

This project is partially funded by the Swiss Data Science Center (SDSC) collaborative projects grant (C22-09); Swiss Government Excellence Scholarship (ESKAS-Nr: 2021.0468) to A.G..

Data availability

The input ATAC-seq bigwig or bam files were obtained from the ENCODE data portal [29] or the GDC data portal [32]. Each patient's SNPs were also obtained from the GDC data portal [32] whereas information on somatic mutations were downloaded from the ICGC data portal [33]. All accession codes are provided in Supplementary Table S1.

References

[1] P.G. Giresi, J. Kim, R.M. McDaniell, V.R. Iyer, J.D. Lieb, Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome research **17**(6), 877–885 (2007)

[2] L. Song, G.E. Crawford, Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols **2010**(2), pdb–prot5384 (2010)

[3] J.D. Buenrostro, B. Wu, H.Y. Chang, W.J. Greenleaf, Atac-seq: a method for assaying chromatin accessibility genome-wide. Current protocols in molecular biology **109**(1), 21–29 (2015)

[4] J. Zhou, O.G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods **12**(10), 931–934 (2015)

[5] D.R. Kelley, J. Snoek, J.L. Rinn, Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research **26**(7), 990–999 (2016)

[6] D. Quang, X. Xie, Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. Nucleic acids research **44**(11), e107–e107 (2016)

[7] X. Min, W. Zeng, N. Chen, T. Chen, R. Jiang, Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. Bioinformatics **33**(14), i92–i101 (2017)

[8] Y. Guo, D. Zhou, R. Nie, X. Ruan, W. Li, Deepanf: A deep attentive neural framework with distributed representation for chromatin accessibility prediction. Neurocomputing **379**, 305–318 (2020)

[9] Q. Liu, F. Xia, Q. Yin, R. Jiang, Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics **34**(5), 732–738 (2018)

[10] D.R. Kelley, Y.A. Reshef, M. Bileschi, D. Belanger, C.Y. McLean, J. Snoek, Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research **28**(5), 739–750 (2018)

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

[12] T.A. Cazares, F.W. Rizvi, B. Iyer, X. Chen, M. Kotliar, A.T. Bejjani, J.A. Wayman, O. Donmez, B. Wronowski, S. Parameswaran, et al., maxatac: Genome-scale transcription-factor binding prediction from atac-seq with deep neural networks. PLOS Computational Biology **19**(1), e1010863 (2023)

[13] R. Yang, A. Das, V.R. Gao, A. Karbalayghareh, W.S. Noble, J.A. Bilmes, C.S. Leslie, Epiphany: predicting hi-c contact maps from 1d epigenomic signals. Genome Biology **24**(1), 134 (2023)

[14] J. Tan, N. Shenker-Tauris, J. Rodriguez-Hernaez, E. Wang, T. Sakellaropoulos, F. Boccalatte, P. Thandapani, J. Skok, I. Aifantis, D. Fenyö, et al., Cell-type-specific prediction of 3d chromatin organization enables high-throughput in silico genetic screening. Nature biotechnology **41**(8), 1140–1150 (2023)

[15] V.R. Gao, R. Yang, A. Das, R. Luo, H. Luo, D.R. McNally, I. Karagiannidis, M.A. Rivas, Z.M. Wang, D. Barisic, et al., Chromafold predicts the 3d contact map from single-cell chromatin accessibility. Nature Communications **15**(1), 9432 (2024)

[16] A. Karbalayghareh, M. Sahin, C.S. Leslie, Chromatin interaction–aware gene regulatory modeling with graph attention networks. Genome Research **32**(5), 930–944 (2022)

[17] A. Pampari, A. Shcherbina, E.Z. Kvon, M. Kosicki, S. Nair, S. Kundu, A.S. Kathiria, V.I. Risca, K. Kuningas, K. Alasoo, et al., Chrombpnet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. bioRxiv pp. 2024–12 (2024)

[18] S. Toneyan, Z. Tang, P.K. Koo, Evaluating deep learning for predicting epigenomic profiles. Nature machine intelligence **4**(12), 1088–1100 (2022)

[19] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, et al., Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature genetics **53**(3),

354–366 (2021)

[20] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie, *Convnext v2: Co-designing and scaling convnets with masked autoencoders*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 16133–16142

[21] A.J. Stewart, S. Hannenhalli, J.B. Plotkin, Why transcription factor binding sites are ten nucleotides long. Genetics **192**(3), 973–985 (2012)

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

[23] M. Yang, J. Ma, Unadon: transformer-based model to predict genome-wide chromosome spatial position. Bioinformatics **39**(Supplement_1), i553–i562 (2023)

[24] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A.H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B.P. de Almeida, H. Sirelkhatim, et al., Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods pp. 1–11 (2024)

[25] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

[26] J. Moore, M. Purcaro, H. Pratt, C. Epstein, N. Shoresh, J. Adrian, T. Kawli, C. Davis, A. Dobin, R. Kaul, et al., Expanded encyclopaedias of dna elements in the human and mouse genomes. Nature **583**(7818), 699–710 (2020)

[27] R. Zhang, *Making convolutional networks shift-invariant again*, in *International conference on machine learning* (PMLR, 2019), pp. 7324–7334

[28] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L.D. Stein, V. Ferretti, The international cancer genome consortium data portal. Nature biotechnology **37**(4), 367–369 (2019)

[29] M.S. Kagda, B. Lam, C. Litton, C. Small, C.A. Sloan, E. Spragins, F. Tanaka, I. Whaling, I. Gabdank, I. Youngworth, et al., Data navigation on the encode portal. arXiv preprint arXiv:2305.00006 (2023)

[30] J. Lee, G. Christoforo, C. Foo, C. Probert, A. Kundaje, N. Boley, D. Kim, M. Dacre, et al., kundajelab/atac_dnase_pipelines: 0.3. 3. Zenodo (2016)

[31] M.R. Corces, J.M. Granja, S. Shams, B.H. Louie, J.A. Seoane, W. Zhou, T.C. Silva, C. Groeneveld, C.K. Wong, S.W. Cho, et al., The chromatin accessibility landscape of primary human cancers. Science **362**(6413), eaav1898 (2018)

[32] R.L. Grossman, A.P. Heath, V. Ferretti, H.E. Varmus, D.R. Lowy, W.A. Kibbe, L.M. Staudt, Toward a shared vision for cancer genomic data. New England Journal of Medicine **375**(12), 1109–1112 (2016)

[33] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L.D. Stein, V. Ferretti, The international cancer genome consortium data portal. Nature biotechnology **37**(4), 367–369 (2019)

[34] D. Aran, M. Sirota, A.J. Butte, Systematic pan-cancer analysis of tumour purity. Nature communications **6**(1), 8971 (2015)

[35] A.P. Heath, V. Ferretti, S. Agrawal, M. An, J.C. Angelakos, R. Arya, R. Bajari, B. Baqar, J.H. Barnowski, J. Burt, et al., The nci genomic data commons. Nature genetics **53**(3), 257–262 (2021)

[36] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks. IEEE transactions on Signal Processing **45**(11), 2673–2681 (1997)

[37] F. Chollet, *Xception: Deep learning with depthwise separable convolutions*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 1251–1258

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Advances in neural information processing systems **30** (2017)

[39] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

[40] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[41] I. Loshchilov, F. Hutter, et al., Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 **5** (2017)

[42] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[43] M.J. Goldman, J. Zhang, N.A. Fonseca, I. Cortés-Ciriano, Q. Xiang, B. Craft, E. Piñeiro-Yáñez, B.D. O'Connor, W. Bazant, E. Barrera, et al., A user guide for the online exploration and visualization of pcawg data. Nature communications **11**(1), 3400 (2020)