Natural language processing strategies for discovery of cell type-specific DNA regulatory elements

Buzatu R., Wang Y., Kenna K.

Abstract

Understanding the gene transcription rules present in non-coding DNA is essential for unraveling the genetic code that establishes cellular fate. In this study, we aim to narrow down on regulatory regions and motifs within the central nervous system (CNS) that determine cell specificity. While the use of ATAC-seq data has been proven efficient in defining relevant regions of open chromatin, further analysis is required in order to obtain insights into specific regulatory elements. To that end, we propose a strategy involving natural language processing techniques to identify DNA transcription factor (TF) binding sites relevant to each cell type. We employ topic modelling for co-clustering of ATAC-seq peak sequences and cell types; as a result, we can retrieve 'topics' consisting of functionally related non-coding DNA regions, that provide a starting point for further analysis and identification of cell-specific feature combinations. Furthermore, we finetune a BigBird language model, pre-trained on the human genome, to distinguish between GABAergic, glutamatergic, and non-neuronal cells. The Byte-Pair Encoding tokenization method allows us to extract the most important DNA motifs for making the class predictions, as well as their corresponding attention scores, which can be mapped back to the peak sequences to identify TF binding sites. We show that this method allows identification of known regulatory elements and propose new strategies to extract more meaningful and specific information from the language models.

Layman's Abstract

A large portion of our DNA does not code for genes, but instead contains a set of instructions that specify which of these genes need to be active to grant a cell its specific type. This study aims to understand how and which of these DNA elements work together to determine the types of cells in the brain. Using artificial intelligence techniques that have been developed to process natural language, we look for specific DNA patterns that are distinct between these types of cells. To do this, we first employ a method called 'topic modelling' to group together similar DNA sequences and cell types, providing us with a starting set of regions to explore. In order to obtain more specific motifs, we leverage the power of transformer language models which can be efficiently used for transfer learning. In simple terms, a model that has been trained for a very general task on a large amount of data can serve as the starting model and be re-trained for a more specific task. Following this method, we fine-tune a BigBird model to distinguish between different types of brain cells, such as GABAergic, glutamatergic, and non-neuronal cells. By analyzing the internal attention mechanism of this model, we can identify specific patterns that the transformer found to be important for determining the cell type. We further show how information can be used to recognize known regions where proteins called transcription factors bind to the DNA and control gene activity.

Introduction

Cellular fate is a result of the expression of particular gene patterns, which are in turn regulated by the binding of transcription factors (TFs) to DNA regulatory regions. Networks of TFs work together to coordinate cell division, differentiation and death. These DNA regulatory regions are present in the form of specific sequences that can be recognized by the DNA-binding domain of the TFs, and aid in initiating RNA transcription (promoters) or increase the transcription rate (enhancers) of a certain gene. In disease, these networks can be misregulated due to mutations, therefore, understanding the 'code' behind these pathways is a key step in interpreting DNA variants in such context (Minnoye et al., 2020).

Modern technologies used in screening for disease-related genomic mutations, in the form of genome-wide association studies (GWAS) (Uffelmann et al., 2021), or rare variant association studies (RVAS) (Auer & Lettre, 2015), have seen great success in detecting common and rare genetic variants. However, functional interpretation of non-coding mutations is still challenging, because determining their effect requires understanding of the disruption they cause in the gene regulatory pathways. Therefore, knowledge of the regulatory DNA 'code' present in the non-coding genome and the interactions of these regulatory elements can facilitate the interpretation of the effect of such mutations. In this project, we aim to develop strategies that can help us discover regulatory DNA elements and how they relate to cell types in the central nervous system (CNS). The focus of this research will be using supervised and unsupervised natural language processing (NLP) strategies to determine whether DNA language models are able to capture the underlying structure of the non-coding genome, and highlight relevant motifs for cell differentiation.

The central nervous system carries a great variety of cell types with different functional tasks, which is possible due to the activity of alternative gene regulatory programmes. In order to obtain insights into CNS cellular evolution, an ideal study region is the primary motor cortex (M1) due to its functional conservation across mammals. In their study, Bakken et al., 2021 use single-nucleus chromatin accessibility and messenger RNA expression sequencing (SNARE-seq) to human, marmoset and mouse M1 samples; this technique combines ATAC-seq and RNA analysis in order to profile chromatin accessibility and gene expression at once (Chen et al., 2019). The Assay of Transposable Accessible Chromatin sequencing (ATAC-seq) technique employs mutated Tn5 transposases to identify open chromatin regions and cut it to ligate adaptors that allow for later sequencing (Buenrostro et al., 2015). While annotating these regions allows us to determine sequences of open chromatin, it does not provide further insight into the actual mechanisms behind the regulatory networks (Yan et al., 2020), nor does it allow identification of specific motifs to which transcription factors might bind. In order to facilitate this process, we investigate whether natural language tools, in the form of topic modelling and classification using transformers, are able to pinpoint cell-type-relevant regulatory sites in the ATAC-seq data.

Firstly, for topic modelling, the cisTopic (Bravo González-Blas et al., 2019) framework, built on latent Dirichlet allocation (LDA) (Blei et al., 2003) and Gibbs sampling, can co-cluster DNA regulatory regions and cell types. This grouping is based on patterns of opening and

closing between functionally related regions. As a language tool, topic modelling is generally used for document clustering. In this context, it works by retrieving representative words for each pre-defined topic, and consequently assigning topics to input documents. In biological terms, the input peaks are assigned into regulatory topics based on activity patterns, followed by cell type classification based on these findings. Thus, the topics that are found through this analysis provide a starting point to investigate combinations of regulatory elements that are responsible for gene transcription management in particular cell types.

Secondly, we harness the power of supervised language models to determine whether they are able learn the 'language' of DNA non-coding elements. Due to the inherent similarities and translatability between biological sequences and natural language, the latest research in the field of bioinformatics has been focused on identifying ways in which these language models can be applied to answer biological questions (Zhang et al., 2023). In this project, we are exploring the use of transformers, which revolutionized the field of artificial intelligence when introduced in 2017 by Vaswani et al. due to their attention mechanism. Previously, Recurrent Neural Networks (RNNs) and Long-Short Term Memory (LSTM) networks were used to tackle text processing tasks; however, these methods have proven less effective when dealing with long input sequences, as they struggle to process the dependencies between long-distance word groups. This issue was solved by the introduction of neural attention mechanism, which processes the inputs in pairs of tokens instead of sequentially, meaning that an attention score is calculated for each pair of words in the input sequence. This score allows the model to understand the importance of each interaction of possible word pairs, thus being able to more efficiently model the relationships between the data.

While these algorithms have been used for a variety of tasks, from translation to question answering, when it comes to input classification, an encoder is preferred, such as the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). This type of model is used to generate numerical embeddings for the input sequences which capture contextual information, and can be then passed on to a classifier for predictive tasks. This concept is called transfer learning, where a model is pre-trained on a large amount of data to learn general relevant features, and can then be fine-tuned to answer a specific classification task.

In the field of bioinformatics, DNA sequence processing is usually done using DNABERT (Ji et al., 2021), which was trained to specifically encode DNA sequences into numerical representations. BigBird (Zaheer et al., 2020), a transformer model with a highly similar architecture to BERT, has been developed in order to tackle the need for parsing longer input sequences; it does so by computing a limited number of the paired attention scores, thus resulting in a sparse attention matrix. In this project, we employ GENA-LM (Fishman et al., 2023), a transformer model using the BigBird architecture that has been pre-trained on the entire genome, and we fine-tune it to distinguish between CNS cell types given the corresponding ATAC-seq peaks as input.

Overall, we wish to determine if NLP algorithms are able to learn the gene regulatory rules encoded in the genome, and whether the decisions behind these models can be interpreted to narrow down on cell type-specific regulatory motifs.

Methods

1. Dataset

The data used in this project has been collected by Bakken et al., 2021 using SNARE-seq to profile the chromatin accessibility of M1 cells. The resulting dataset consists of a sparse matrix of mapped ATAC-seq reads across 273103 genome regions for 84178 cells, along with their corresponding cell type from high-level (neuronal vs non-neuronal), to more specific characterizations. Furthermore, an RNA count matrix is available, providing the read counts of the 84178 cells across 31741 genes.

2. Topic modelling of open chromatin regions

2.1. Co-clustering into topics

In order to perform topic modelling on the available ATAC-seq dataset, the software cisTopic (Bravo González-Blas et al., 2019) was implemented in an R script. The sparse matrix containing the reads for each cell and DNA region pair was transformed into a cisTopic object, and the corresponding metadata was added to each cell record. The model training was performed for 2-20, 30, 40 and 50 topics; finally, 14 topics were chosen as the optimal number using the value of the second derivative in each point of the likelihood curve. For each DNA region present in the input data, a score was obtained representing the contribution to each of the 14 topics, as well as 14 scores defining the topic contributions in each cell.

2.2. Pathway analysis

To determine whether the topics that were found had biological relevance, the ATAC-seq peaks were linked to the closest gene using the LinkPeaks() function available from the Signac package (Stuart et al., 2021), which follows the method described in Ma et al., 2020. A Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was used to determine pathways that were significantly over-represented in each topic, based on all DNA region-topic scores and the linked genes. A linear regression model was used to determine the pathways that are up- and down-regulated in each separate topic, by using those particular topic scores as the function output. The dataset used for this analysis corresponds to the C5 gene ontology set from version 7.4 of the Molecular Signatures Database (Liberzon et al., 2011). Only the resulting pathways with adjusted p-values smaller than 0.05 were considered significantly enriched. Using information found in literature, these results were compared with the cell type most representative of several topics in order to draw conclusions about the gene pathways driving them.

2.3. Cell type prediction from topic scores

A feed-forward neural network implemented for a multi-class classification was used to predict the highest-level type (‘GABAergic’, ‘Glutamatergic’ or ‘Non-Neuronal’) of each input cell, given the 14 topic scores as inputs (**Figure S2**). The model was implemented using the TensorFlow for R package, version 2.11.0.

The raw dataset consisted of 24006 non-neuronal, 22217 GABAergic and 37955 glutamatergic cells. To achieve more balance between the classes, the glutamatergic cells were randomly subset to 24006. During splitting, 80% of the records of each class were used as a training set and 20% were set aside for testing. During training, 20% of the training set was used for validation after each epoch.

The architecture of the neural network was formed by an input layer of 14 nodes, followed by a fully connected hidden layer with 7 nodes, a dropout layer with a dropout rate of 25%, and finally an output layer of three nodes. For the hidden layer, ReLu was employed as an activation function, while softmax was used in the output in order to obtain probability scores for each class. The model was trained for 40 epochs as a result of the implemented early stopping mechanism, using TensorFlow's categorical cross entropy as a loss function and accuracy as a performance metric.

3. Transformer models used for cell type prediction

Fishman et al.'s GENA-LM-BigBird-base-T2T pre-trained model was finetuned to a multi-class classification problem. To predict the type ('GABAergic', 'Glutamatergic' or 'Non-Neuronal') of each input cell, a set of DNA sequences was used as input for each data point, corresponding to ATAC-seq peaks. For tokenization of the input sequences, GENA-LM-BigBird-base-T2T's pre-trained Byte-Pair-Encoding (BPE) tokenizer was used. **Figure 1** shows an overview of the model fine-tuning process.

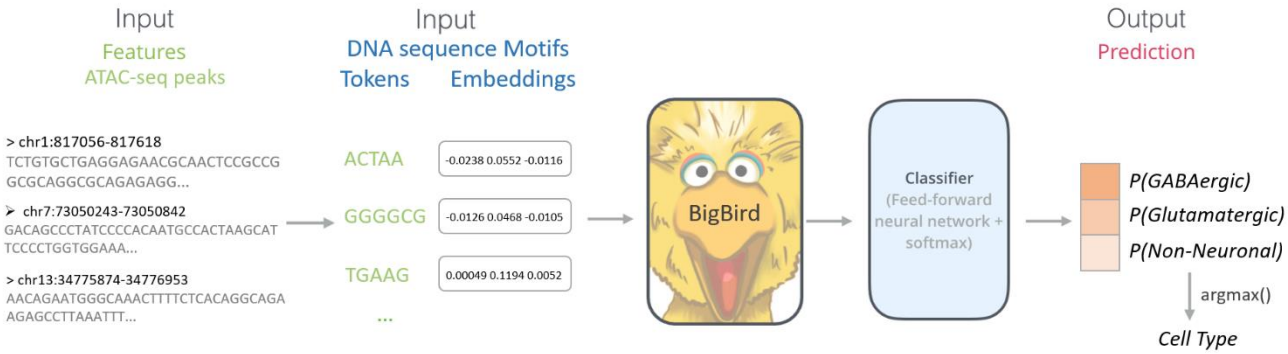

Figure 1: Illustration showing the use of a BigBird transformer model for cell type classification. The input peak sequences are tokenized using BPE tokenization and are encoded using GENA-LM-BigBird's pre-trained tokenizer. They are passed onto the pre-trained BigBird model, and the resulting encodings are classified into one of the three cell-types: GABAergic, Glutamatergic and Non-Neuronal.

3.1. Input pre-processing

To obtain a set of representative ATAC-seq peaks for each data point, the selection was performed in two steps. First, the read counts were normalized by the length of each peak region to avoid bias towards longer peaks. Afterwards, for each cell, only the regions that had more than the 90^{th} percentile number of reads were considered in order to remove noise.

Secondly, a differential accessibility analysis was conducted using Seurat's (Hao et al., 2021) FindMarkers() function between each pair of cell types, where only differentially accessible regions (DARs) that were detected in at least 5% of either population were considered The results were concatenated per cell type, and only the peaks with a positive log-fold change (logfc) of expression and an adjusted *p*-value below 0.05 were considered. We obtained 2602 glutamatergic, 333 GABAergic and 44 non-neuronal DARs.

3.2. Creating training and testing sets

To create the initial input dataset, for each cell, the remaining regions per cell after the first step were intersected with the peaks that had passed the differential accessibility test for that particular cell type. Cells with duplicate inputs were removed. The resulting dataset consisted of 22202 GABAergic, 37955 glutamatergic and 10572 non-neuronal cells.

This number was further subset to balance the classes. Finally, an equal number of 6767 cells (the maximum available from the least frequent cell type) from each class were randomly selected for the *training set 1* (80%), 2114 for the *test set 1* (20%) and 1691 for the *validation set 1* (20% of training). The inputs were tokenized using Fishman et al.'s pretrained Byte Pair Encoding (BPE) tokenizer, with a maximum of 4096 tokens.

A second test set (*test set 2*) was created using the same protocol, but with the difference that the input peaks were no longer intersected with the differentially accessible regions, only subset based on the read count as described previously. This set, containing 1805 glutamatergic, 2905 GABAergic and 1632 non-neuronal cell samples was also used to test the first model.

The pre-trained GENA-LM-BigBird transformer model was fine-tuned again (*model 2*) on a differently processed dataset. In this case, the differentially accessible peaks were separated between the training (80%) and test set (20%) within each cell type, and only those were used for the intersection in the final input processing for the corresponding dataset. The duplicated entries were removed in both. The training set was further balanced and split, so that an equal number of 5748 cells of each type were used for training, 1437 for validation, while the test set remained uneven with 37952 glutamatergic, 16145 GABAergic and 186 non-neuronal cells.

An overview of all the datasets used for model training and testing is available in the **Supplementary Material (Table S2).**

3.3. Model training and testing

The pre-trained model was finetuned two separate times (*model 1, model 2*), each time for 10 epochs on the multi-classification tasks of predicting the type of the input cell (glutamatergic, GABAergic or non-neuronal) using Pytorch's (version 1.9.0) Transformers package. Torch's cross-entropy function was used to estimate the training and validation losses, and the accuracy was recorded for each epoch to estimate the performance. For all three test sets, accuracy, precision, recall and F1 score were recorded in order to determine the classifier performance.

3.4. Interpretation of attention scores

After training of the transformer model was completed, we passed the DNA sequence of a randomly chosen peak through the model and extracted the computed attention matrices between the input tokens at each of the 12 layers and 12 heads. To be able to determine the effect of each token on the class prediction, we considered the attention score between each input token and the [CLS] token, which represents an encoding of the entire input 'sentence'. Following a similar method to DNABERT-viz (Ji et al., 2021), the attention scores of the token pairs were summed over all the layers and heads, thus resulting in numerical values that quantify the importance of each input token.

In order to visualize these directly onto the input sequences, we mapped the scores to the peaks of interest. The corresponding attention score was added at each position in the DNA sequence when a token was encountered. The number of overlapping base-pairs between previously annotated regulatory sites (Pachkov et al., 2013) and the positions with scores higher than the 90^{th} percentile in the given region were then counted. These were compared with the overlap resulting after random shuffling of the attention scores over each sequence 100 times. This analysis was both performed on three hand-picked regions, as well as separately on a larger dataset of 18 peak sequences which had high topic scores. The exact chromosomal locations that were used are provided in the **Supplementary Material (Table S1).** For the latter dataset, a one-sided Wilcoxon rank-sum test was performed to determine whether the difference between the attention-generated overlap and the average shuffled overlap was significant.

The source code for this project is available in two repositories:

1. https://github.com/rafaella-buzatu/regElem (Methods 2)
2. https://github.com/rafaella-buzatu/GENALM_Finetune_regElem (Methods 3)

Results

In this study, we aimed to determine whether natural language models are able to pick up on the gene regulatory rules encoded in the open chromatin regions of the non-coding genome and how they can be used to narrow down on CNS cell-type specific regulatory regions and motifs.

1. Topic modelling of open chromatin regions

1.1. Co-clustering into topics

To achieve this, we began by using cisTopic (Bravo González-Blas et al., 2019) to group the open chromatin regions obtained through ATAC-seq into clusters referred to as 'topics', as well as to determine the contribution of each of these topics to the different cell types. By examining the log likelihood curve, we determined that 14 topics were the number that allowed for the best optimization of the two probability distributions. Thus, we obtained two output sets: first, a contribution score of each topic to every cell in the dataset, and second, a probability score for each topic-region pair.

In order to visualize these results, we performed a tSNE analysis on the cell-topic probabilities through the same software and overlaid the scores of each topic on this mapping. **Figure 2** illustrates the clusters, colored by the cell type at the highest (**Figure 2A)** and lowest level (**Figure 2B)** available in the dataset, while showcasing several examples of the topic specificity.

There is a clear differentiation between the three broad cell types based on the topic modelling results, and even further separation when looking into the more specific subclasses. Astrocytes, micro perivascular macrophages (Micro-PVM), oligodendrocyte progenitor cells (OPC) and oligodendrocytes (Oligo) are examples of cell types that form their own, well-separated clusters, while also showing overlap with specific topics. The peaks assigned to topic 12, for instance, appear to be highly relevant to astrocytes **(Figure 2F)**, while topic 5 is almost exclusively active in Micro-PVM cells **(Figure 2E),** and topic 2 is highly expressed in OPCs (**Figure 2D)**. However, certain topics appear to be active across all cell types, such as topic 1 (**Figure 2C**), thus providing less information about specificity. Figures showcasing the same analysis for each topic are available in the **Supplementary Materials**.

Topic modelling in itself, however, is an unsupervised method, meaning that there is no simple way to determine the quality of this clustering or its level of informativeness. As a result, we decided to examine whether the topic-cell assignments held any biological significance.

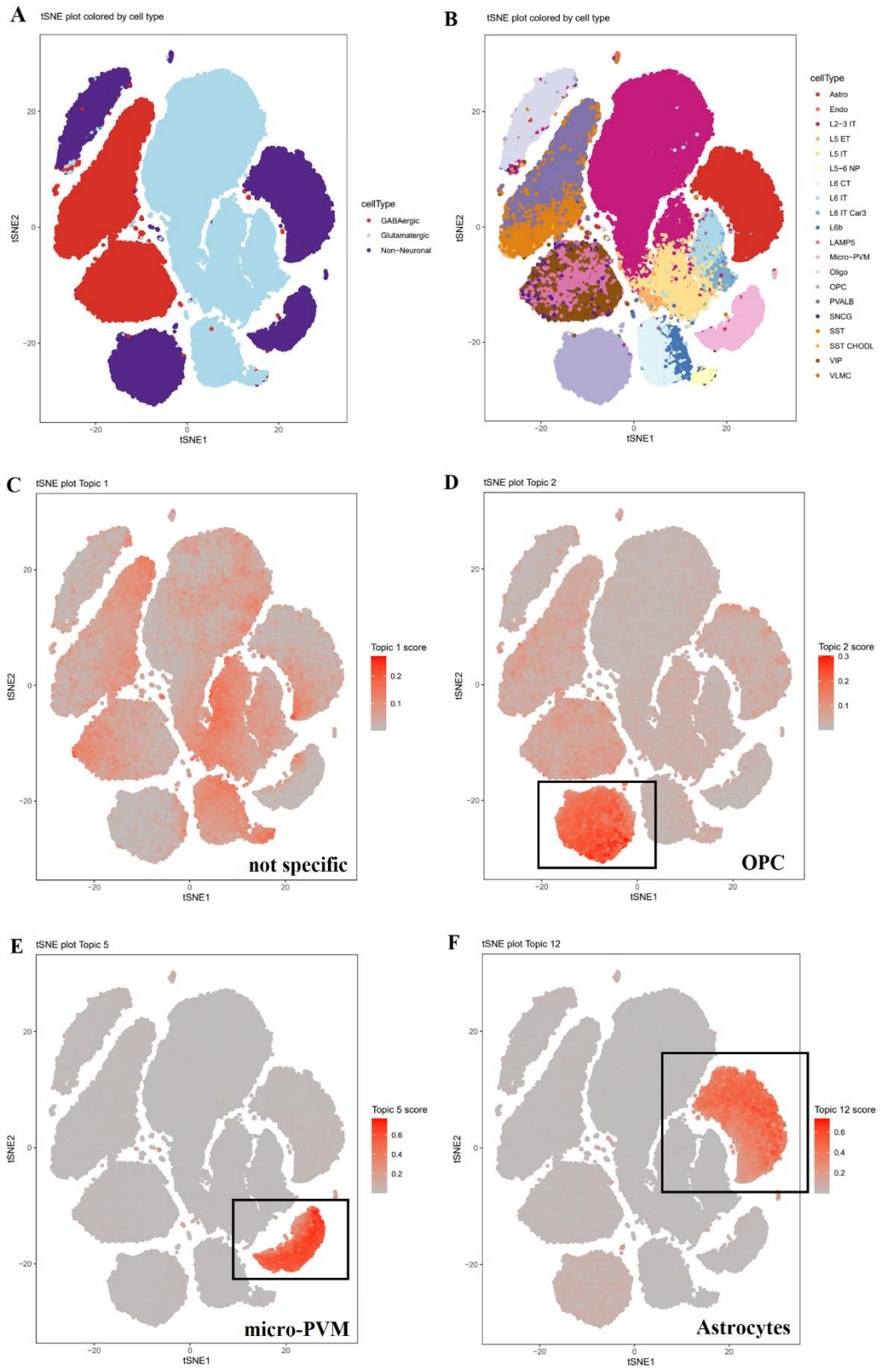

Figure 2: tSNE showcasing the clustering of the cells in the dataset based on the cell-topic distributions. The cells are coloured based on cell type, following the (A) highest and (B) lowest level classification known. The topic scores are superimposed on the clustering plot, showing which topics are active in which cells: Topic 1- no cell specificity (C), Topic 2 – OPC cells (D), Topic 5 – micro-PVM cells (E) and Topic 12 – astrocytes (F).

1.2. Pathway analysis

We wished to see if this topic-cell type specificity could be explained through an investigation into the cellular pathways that are active in each topic. Based on the RNA expression levels, we were able to draw links between the ATAC-seq peaks and relevant genes. Using the connection between topics and DNA regions in a GSEA analysis, we obtained a list of cellular pathways that were significantly enriched in each topic. The results are available in the **Supplementary Materials** folder ('GSEAresults.xlsx').

When trying to link these pathways to the cellular functions, we were able to discover several relevant connections. Topic 5, for instance, overlays the micro-PVM cell type, which are a population of brain cells whose main function is maintaining the integrity of the blood-brain barrier (BBB) through regulation of immune responses (Faraco et al., 2017). These cells can function as antigen-presenting cells (APCs), thus being part of the immunological synapse between a T cell and an APC, which relates to one of the pathways correlated with topic 5. Topic 7 shows a high contribution to PVALB cells, which are a subtype of GABAergic neurons that modulate the activity of voltage-gated calcium ion channels (Baimbridge et al., 1992). This reflects in the corresponding pathway related to neuronal ion channel clustering. Finally, topic 12 overlaps with the cluster of astrocyte cells, which, like micro-PVM cells, also contribute to the maintenance and regulation of the BBB. According to our analysis, the genes linked to the astrocyte-dominated topic are active in T-cell as well as chondrocyte differentiation pathways, both connections which have been previously suggested in scientific literature (Onore Beurel et al., 2014; Kepes et al., 1984). Nevertheless, most of the cell-pathway connections were not as obvious and proved difficult to interpret directly.

1.3. Cell type prediction from topic score

Because of this difficulty in deciphering the biological significance of the topics, we decided to assess whether they were indeed sufficiently predictive to distinguish between cell types. Thus, we used the topic scores in a multi-class classification model, trained to predict the high-level type of each cell (GABAergic, glutamatergic or non-neuronal). The evolution of the loss function and accuracy during training is stored in the **Supplementary Figure S3**. **Figure 3** illustrates the results of the class prediction on the test set, with 4501/4512 samples correctly predicted as non-neuronal, 4711/4713 as glutamatergic and, finally, 4743/4777 accurately assigned the label GABAergic, corresponding to an accuracy of 99.67%. Thus, at a high level, the topic scores appear to hold enough information to make this distinction.

In order to gather further insights into the logic behind the topic clustering, this experiment can be extended to a model that distinguishes between more specialized cell types. Furthermore, in order to grasp exactly which topics and, consequently, which DNA regions, are active within specific cell types, we need to look into the predictive mechanism of the model.

Figure 3: Confusion matrix showcasing the results of the neural network model used to predict the cell type using the topic scores as input. The accuracy of prediction is 99%, suggesting that predicting the cell type by using the topic scores alone is a simple task for such a model.

2. Transformer models for cell type prediction

Since it did not prove intuitive to determine the performance of an unsupervised model on understanding the intricate regulatory networks of genes, we also experimented with using a supervised model. To this end, we fine-tuned the GENA-LM BigBird encoder model, which had been pre-trained on the entire genome, to differentiate between the three high-level cell types (non-neuronal, GABAergic and glutamatergic) when given the respective ATAC-seq peaks as input.

2.1. Model fine-tuning

To begin with, we trained *model 1* on a dataset which was pre-processed to capture the most relevant peaks through a differential accessibility analysis. The evolution of the loss function during training is recorded in **Supplementary Figure S4.** It was first tested on the *test 1* dataset that had gone through the same pre-processing steps as the training. The results of this evaluation are portrayed in the confusion matrix from **Figure 4A**, as well as in **Table 1,** showing that 2114 non-neuronal, 2111 glutamatergic and 2112 GABAergic cells were classified correctly out of the 2114 representing each class, resulting in 99.92% accuracy. However, when testing the same model on the *test set 2*, which was not subset using DARs, the performance was significantly poorer, as almost all samples were predicted as Glutamatergic, resulting in a weighted average F1 score of 0.29 (**Figure 4B, Table 2**). These results show that the model is not able to generalize on a less pre-processed dataset, and it might have not learnt to associate regulatory motifs with cell type, as was expected.

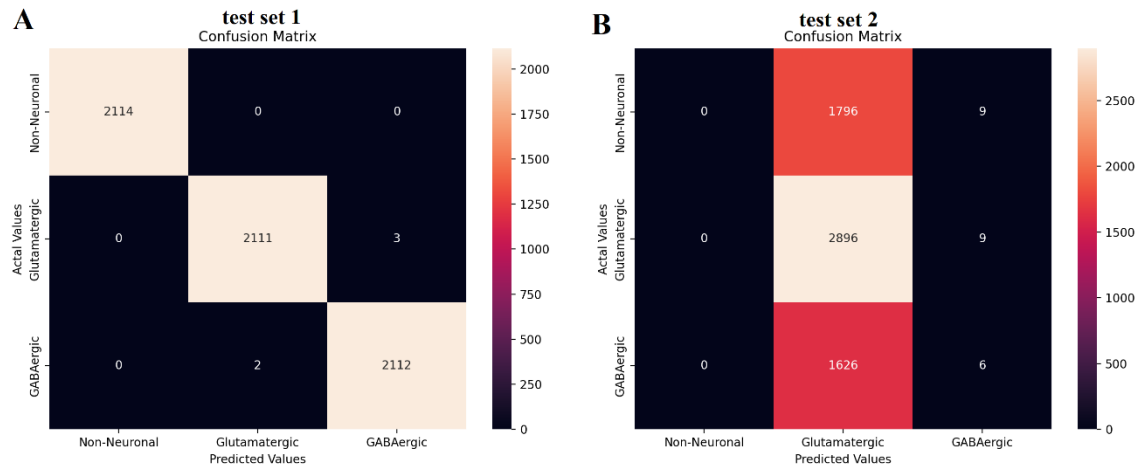

Figure 4: Confusion matrices from the results of the first fine-tuned GENA-LM-BigBird model on test set 1 (A) and test set 2 (B). The prediction accuracy of the model on test set 1 is 99%, however the performance significantly decreases when the model is tested on a less pre-processed set of peaks (test 2). Most cells are now predicted as Glutamatergic, which can be explained by the initial high variability in this class's differentially accesible peaks, which are now re-introduced as input sequences for the other classes and are likely recognised as glutamatergic by the trained model.

Since it is not clear from these results whether the model was able to learn the patterns of regulatory motifs, we decided to test that by fine-tuning the model on a different set of input data, in which case the DARs were split between the training and test sets. The rationale behind thi is that we expect the motifs we are looking for to actually be shared between different peaks that are cell-specific. Thus, if the model were to actually learn the sequence features we are interested in, it should be able to recognize them in the test set despite not having been trained on the corresponding entire peak sequences.

The evolution of the loss function and accuracy during training is stored in the **Supplementary Figure S5. Figure 5** shows the results of the model on the corresponding test set, while **Table 3** indicates the F1 score per class, as well as on average. Firstly, the prediction accuracy for the non-neuronal cells is 0%, suggesting that none of the patterns that the model had learnt for this class during training were recognised in the test set. This could be caused by the low variety of DARs that were left for this testing subset – eight peaks. We can see that, while the other two classes have better performances, with F1 scores of 0.61 (GABAergic) and 0.81 (Glutamatergic), the prediction power appears to decrease with the number of DARs that can be used as input. Therefore, while the results suggest that the model is indeed capable of learning and recognising cell-specific motifs in these open chromatin regions, testing is limited by the amount of available data. Alternatively, we could try to remove the non-neuronal class and re-train the model on more specific cell types to further assess its learning abilities.

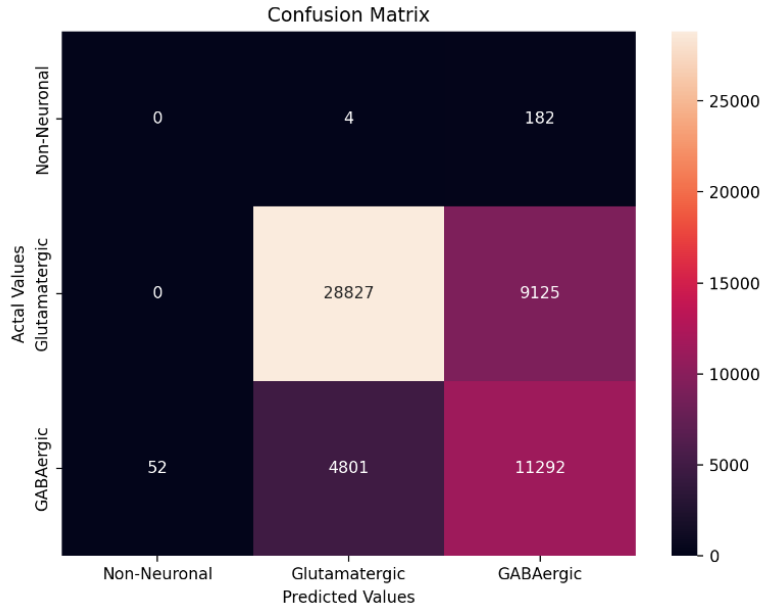

Figure 5: Confusion matrix showcasing the results of the second fine-tuned GENA-LM-BigBird model. The prediction accuracy per class decreases with the number of available DARs that were used for preprocessing, such that Glutamatergic and GABAergic classes have good predictions, while Non-Neuronal, which had 44 differentially accesible peaks, has an accuracy of 0%. However, for the former two classes, the model is able to recognise regulatory elements in the test set even if it has not encountered the peak sequences in the training set.

Table 1. Table showing the performance metrics of the first fine-tuned model on test set 1

	Precision	*Recall*	*F1*	*Support*	*Accuracy*
Glutamatergic	1.00	1.00	1.00	2114	
GABAergic	1.00	1.00	1.00	2114	0.99
Non-Neuronal	1.00	1.00	1.00	2114	
Weighted average	1.00	1.00	1.00	2114	

Table 2. Table showing the performance metrics of the first fine-tuned model on test set 2

	Precision	*Recall*	*F1*	*Support*	*Weighted Accuracy*
Non-Neuronal	0.00	0.00	0.00	1805	
Glutamatergic	0.46	1.00	0.63	2905	0.46
GABAergic	0.25	0.00	0.01	1632	
Weighted average	0.27	0.46	0.29	6342	

Table 3. Table showing the performance metrics of the second fine-tuned model

	Precision	*Recall*	*F1*	*Support*	*Weighted Accuracy*
Non-Neuronal	0.00	0.00	0.00	186	
Glutamatergic	0.86	0.76	0.81	37952	0.74
GABAergic	0.55	0.70	0.61	16145	
Weighted average	0.76	0.74	0.75	54283	

2.2. Interpretation of attention scores

The next step toward the goal of identifying cell type-specific features within the open chromatin regions involves extracting and interpreting information from the transformer model. Therefore, we decided to use the token attention scores to determine the specific motifs that the model was considering when making each prediction. For an initial look, we used the first model we fine-tuned and randomly selected the peak located at chr5:76952593-76954930 to visually inspect the results. We passed the peak through the model, mapped the cumulative attention scores of the existent tokens onto the DNA sequence and plotted heatmaps for visualization of three sub-regions of the chosen peak (**Figure 6**). The same figure also illustrates the comparison with previously annotated transcription factor binding sites form the Swiss Regulion database (Pachkov et al., 2013). A visual examination supports a certain degree of overlap between the sites most important for the prediction and the known regulatory element locations, however this overlap also appears skewed in certain places, such as for RCOR1 (**Figure 6A**) and RUNX1 (**Figure 6B**). Furthermore, there appear to be other locations which, while highlighted by the attention mechanism, have no previous annotation (**Figure 6C**).

To determine whether this overlap was indeed informative, we randomly shuffled the attention scores within these three regions 100 times (**Figure S6**) and compared the number of overlapping base pairs of each resulting score vector with the known locations of the regulatory elements. The results are recorded in **Table 4.** For the three sub-regions of the chosen peak, the difference between the overlap counts is positive in all cases, showing a better recognition of annotated elements. To get a better idea of the significance of our findings, we repeated the analysis on the entire DNA sequences of 18 peaks chosen for high topic score assignments (**Table 4**). The differences are also positive in the regions where known TF binding sites were actually present, and a Wilcoxon sum-ranked test proves them to be significant (*p-value* 0.00865). Therefore, it is possible to identify sites of regulatory elements by leveraging the attention scores, however, more precise results are likely to be obtained when intepreting a model with higher accuracy.

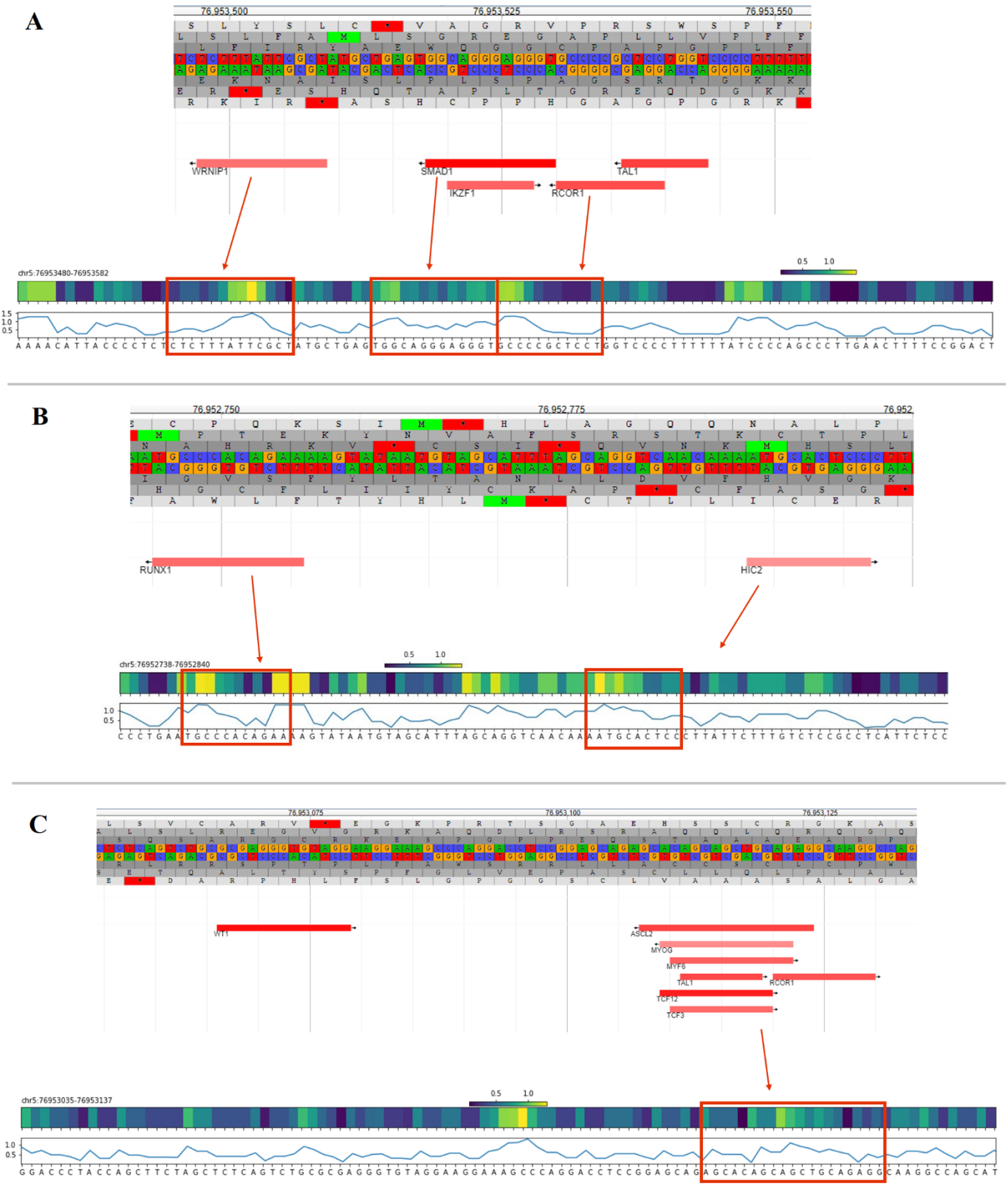

Figure 6: Token attention scores mapped on a the region chr5:76953480-76953582 compared to annotated TF binding sites from the SwissRegulion database. The figure indicates a rough overlap between some of the DNA regions with high cummulative mapped attention scores (higher than 1) and the annotated elements. Shown are (A) chr5:76953480- 76953582; (B) chr5:76952738-76952840; (C) chr5:76953035-76953137.

Table 4. Table showing the difference in the overlap count between the mapped shuffled scores

Location	*Overlap mapped attention*	*Overlap shuffled*	*Difference*
chr5:76953480-76953582*	76	61.72	**14.28**
chr5:76952738-76952840*	30	20.94	**9.06**
chr5:76953035-76953137*	15	8.98	**6.02**
chr1:633543-634316	0	0	**0**
chr3:93470467-93471024	0	0	**0**
chr5:76952593-76954930	205	157.19	**47.81**
chr13:110306246-110308993	405	311.20	**93.70**
chr5:100900736-100903868	172	173,34	**-1.34**
chr2:17877787-17879898	309	255.45	**53.55**
chr11:16605771-1660816	0	0	**0**
chr7:8433246-8434867	611	481.97	**129.3**
chr13:37868657-37871376	777	319.44	**457.56**
chr2:170815757-170818430	478	373.37	**104.63**
chr18:12253588-12256062	374	260.43	**113.57**
chr1:18071366-18074416	0	0	**0**
chr9:135166103-135168844	0	0	**0**
chr3:194621781-194624359	0	0	**0**
chr7:45407218-45409817	0	0	**0**
chr9:127679563-127681320	0	0	**0**
chr18:51857753-51859599	0	0	**0**
chr11:119416959-119418818	0	0	**0**

***** *the three hand-picked regions for the initial analysis*

Wilcoxon Test (>)	
statistic	*p-values*
35.0	**0.00865**

Discussion and Conclusion

The aim of this research was to develop a strategy involving natural language tools to discover DNA features that can aid us in determining which regulatory elements determine which cell type within a dataset of human CNS cells. In order to address this, we used a dataset of open chromatin regions and their respective read count for each cell. We began our investigation by using topic modelling in order to cluster the ATAC-seq peaks into 14 regulatory 'topics' that represent groups of regions that are active together. This is done by optimizing two distributions: the peak regions are assigned a probability for participating in each topic, while the contribution of all of these topics in each cell is also scored. Following this analysis, we are able to create a link between which DNA regions are likely to be open together in specific cell types. However, when we tried to connect this information to known gene pathways for a more in-depth look into the biological meaning behind the topics, we were not able to find many easily interpretable connections. Thus, in order to further assess the reliability of the topic assignment itself, we tested whether the 14 scores were informative enough to predict the cell type. To tackle this, we created a neural network model which was able to predict with high

accuracy whether an input cell was glutamatergic, GABAergic or non-neuronal, based on its scores in each of the topics.

A suitable follow-up would involve further training of a similar model to classify more specific cell types. In order to extract relevant DNA motifs from such a model, however, a new protocol still needs to be developed that can help us understand the model's inner workings. A simple analysis strategy can be exploited, in which different input scores are removed before being passed onto the model, in order to determine their importance in making the final prediction. Thus, a selection of topics could be obtained which are relevant for each predicted class. On the other hand, for a more direct approach, we could try to translate this into a linear model, for which one can easily extract the weights of each topic in relation to the predicted class. Ultimately, a different method would still be necessary to determine the most relevant open chromatin regions that are active together as a topic, before being able to test for any motif enrichment. To address that, the tSNE reduction could be repeated on the region-topic distribution to try to cluster which regions function together within which topic combinations.

On the other hand, we chose to also assess if a supervised NLP technique would be able to learn the regulatory elements that determine the different cellular functionalities. To achieve this, we utilized a pre-trained BigBird model and fine-tuned it on the task of classifying the cells into the three high-level types (glutamatergic, GABAergic and non-neuronal) based on the input peak sequences. While the model had a great performance on the initial test set (99.92% accuracy), where the input peaks were selected based on a differential accessibility analysis, when we removed this pre-processing step from the test set, the accuracy dropped significantly, and almost all the cells were predicted as glutamatergic. To explain this, we have to consider how the transformer model and the tokenization step work. We used a pre-trained BPE tokenizer, which extracted high-frequency motifs from our peak sequences, which were then given as input to the model. The differential accessibility analysis allowed us to pre-determine the regions that were open in a significantly larger percentage for each cell type, which can then lead to more specific motifs being picked up during the tokenization process. After all the pre-processing, the glutamatergic cellular class had a far larger number of possible peaks than the other two types. Therefore, when so many regions are introduced back into the test dataset, it is likely that they were recognized as glutamatergic purely due to the higher initial variety in the training sequences. However, these results raise a different question, namely whether it was indeed relevant regulatory motifs that the model was able to learn, or it simply exploited some other pattern in the data.

In order to address this concern, we split the DARs between the training and test set and fine-tuned a new model to see whether it would be able to recognize relevant motifs within peak sequences it had not previously seen during training. The results of this second model suggest that is indeed the case, however what it also made obvious was the loss of predictive power given a lower number of available cell specific DARs. For a successful model, a different pre-processing strategy would need to be developed in order to address this lack of balance between the peaks in each class. One could try to simply provide the model with all peaks that have a significant read count, without using the differential accessibility analysis results, while still randomly splitting them between training and test per class. However, it is possible that this

introduces a large amount of noise within the data, and whether the model is capable of finding the important features within this noise remains to be seen. On the other hand, the transformer could be fine-tuned on a larger number of classes, representing more specific cell-types, which could help lower the variability between the available DARs.

Nonetheless, once a model with suitable accuracy is obtained, it further needs to be interpreted in order to reach our initial research goals. Going a step beyond training, we attempted to implement a protocol that could help us visualize which tokens were most relevant for the prediction by diving into the attention score matrix. By adding the attention contribution of each token to the [CLS] token, which represents an input-wide encoding and is used for the classification, we can obtain an 'importance' score of each input token for the final prediction. When we mapped these scores back to the ATAC-seq peak regions, we were able to roughly overlap the regions the model found important with the locations of known TF binding sites. A Wilcoxon sum-ranked test showed that the overlap obtained through this process was significantly better (*p-value* 0.00865) than achieved through random attention score assignments. It is likely that this significance would only improve with a more accurate fine-tuned model and would allow us to more reliably extract cell-specific DNA motifs.

Nevertheless, we wish to propose other techniques of explainable AI which could provide us with more insights into the model's predictive behavior. A deep dive into the attention matrix would prove difficult due to its complexity and high dimensionality. However, a simple yet effective strategy could involve token nullification, where random combinations of tokens are removed from the input in order to determine the effect on the prediction performance. This could allow us to piece together which token combinations are relevant to which classes by assessing the change in the class probabilities given the removal.

In conclusion, these results show that there is sufficient predictive power in ATAC-seq data to be able to differentiate between CNS cell types at a high level, and natural language tools are indeed able to pick up those distinguishing characteristics. However, understanding which DNA features control the specificity of these cells and how they work together to regulate gene transcription still requires exploration of these models. Further work is needed to address cell differentiation at a lower level, as well as to perfect the protocols that would allow us to extract reliably relevant features.

References

1. Auer, P. L., & Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, *7*(1). https://doi.org/10.1186/S13073-015-0138-2
2. Baimbridge, K. G., Celio, M. R., & Rogers, J. H. (1992). Calcium-binding proteins in the nervous system. *Trends in Neurosciences*, *15*(8), 303–308. https://doi.org/10.1016/0166-2236(92)90081-I
3. Bakken, T. E., Jorstad, N. L., Hu, Q., Lake, B. B., Tian, W., Kalmbach, B. E., Crow, M., Hodge, R. D., Krienen, F. M., Sorensen, S. A., Eggermont, J., Yao, Z., Aevermann, B. D., Aldridge, A. I., Bartlett, A., Bertagnolli, D., Casper, T., Castanon, R. G., Crichton, K., … Lein, E. S. (2021). Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature 2021 598:7879*, *598*(7879), 111–119. https://doi.org/10.1038/s41586-021-03465-8
4. Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research*, *3*, 993–1022.
5. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., & Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods 2019 16:5*, *16*(5), 397–400. https://doi.org/10.1038/s41592-019-0367-1
6. Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol*, *109*. https://doi.org/10.1002/0471142727.mb2129s109
7. Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, *37*, 1452–1457. https://doi.org/10.1038/s41587-019-0290-0
8. Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved June 18, 2023, from https://github.com/tensorflow/tensor2tensor
9. Faraco, G., Park, L., Anrather, J., & Iadecola, C. (2017). Brain perivascular macrophages: characterization and functional roles in health and disease. *Journal of Molecular Medicine (Berlin, Germany)*, *95*(11), 1143. https://doi.org/10.1007/S00109-017-1573-X
10. Fishman, V., Kuratov, Y., Petrov, M., Shmelev, A., Shepelin, D., Chekanov, N., Kardymon, O., & Burtsev, M. (n.d.). *GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences*. https://doi.org/10.1101/2023.06.12.544594
11. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573. https://doi.org/10.1016/J.CELL.2021.04.048
12. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, *37*(15), 2112–2120. https://doi.org/10.1093/BIOINFORMATICS/BTAB083

13. Kepes, J. J., Rubinstein, L. J., & Chiang, H. (n.d.). *The Role of Astrocytes in the Formation of Cartilage in Gliomas An Immunohistochemical Study of Four Cases*.
14. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. https://doi.org/10.1093/BIOINFORMATICS/BTR260
15. Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y. C., Regev, A., & Buenrostro, J. D. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, *183*(4), 1103. https://doi.org/10.1016/J.CELL.2020.09.056
16. Minnoye, L., Taskiran, I. I., Mauduit, D., Fazio, M., van Aerschot, L., Hulselmans, G., Christiaens, V., Makhzami, S., Seltenhammer, M., Karras, P., Primot, A., Cadieu, E., van Rooijen, E., Marine, J. C., Egidy, G., Ghanem, G. E., Zon, L., Wouters, J., & Aerts, S. (2020). Cross-species analysis of enhancer logic using deep learning. *Genome Research*, *31*(12), 1815–1834. https://doi.org/10.1101/GR.260844.120/-/DC1
17. Onore Beurel, E., Harrington, L. E., Buchser, W., Lemmon, V., Jope, R. S., & Wilson, E. H. (n.d.). *Astrocytes Modulate the Polarization of CD4 + T Cells to Th1 Cells*. https://doi.org/10.1371/journal.pone.0086257
18. Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E., & Van Nimwegen, E. (n.d.). *SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates*. https://doi.org/10.1093/nar/gks1145
19. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., & Satija, R. (2021). Single-cell chromatin state analysis with Signac. *Nature Methods 2021 18:11*, *18*(11), 1333–1341. https://doi.org/10.1038/s41592-021-01282-5
20. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG
21. Uffelmann, E. ;, Huang, Q. Q., Munung, N. S., De Vries, J. ;, Okada, Y. ;, Martin, A. R., Martin, H. C., Lappalainen, T. ;, & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762
23. Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, *21*(1). https://doi.org/10.1186/S13059-020-1929-3
24. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, *2020-December*. https://arxiv.org/abs/2007.14062v2
25. Zhang, S., Fan, R., Liu, Y., Chen, S., Liu, Q., & Zeng, W. (2023). Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances*, *3*(1). https://doi.org/10.1093/BIOADV/VBAD001