Predicting functional relationship between DNA sequence and the epigenetic state
## Contents

1 Executive summary ........................................... 2  
  1.1 Background ............................................. 2  
  1.2 Challenge overview .................................... 3  
  1.3 Data .................................................... 4  
  1.4 Objectives ............................................. 4  
  1.5 Approach ................................................ 6  
  1.6 Main conclusions ...................................... 8  
  1.7 Limitations ............................................. 9  
  1.8 Recommendations and future work ..................... 10  

2 Data overview ................................................ 12  
  2.1 Dataset description .................................... 12  
  2.2 Data quality issues .................................... 17  

3 Data visualisation ............................................ 20  
  3.1 QTL connectivity ....................................... 20  

4 Experiments .................................................... 29  
  4.1 Model architecture ..................................... 29  
  4.2 Conditioning experiments .............................. 30  
  4.3 Receptive field analysis ............................... 35  

5 Future work and research avenues .......................... 41  

6 Team members ................................................. 46
1 Executive summary

1.1 Background

All major neurodegenerative diseases are characterized by substantial heritability and while recent large-scale genetic efforts have identified variants associated with disease, these often lie in non-coding, regulatory regions and cannot be linked to any functional outcomes [11]. Recent work has repeatedly highlighted that the genetic risk for Alzheimer’s disease acts primarily via microglia - the resident macrophages of the central nervous system [14]. However, most of these variants do not directly affect protein function, instead they are suspected of influencing gene expression by altering genomic regulatory elements.

One of the main challenges in genome biology is the understanding of highly cell-type specific gene regulatory mechanisms. In the context of diseases, research is typically targeted towards cell-types relevant to disease phenotype, or which are presumed to have a causal downstream effect. Genetic variants which are associated with the function or activity of regulatory elements (known as quantitative trait loci or QTLs) will exert their effects in a cell-type specific manner [5]. Mapping out molecular and regulatory QTLs comprehensively in disease-relevant cell-types would enable us to interpret functional outcomes of genetic variants on gene expression and regulation.

While we are focused on understanding its impact on dementias, the impact of improving regulatory genomic predictions will affect research on all biological traits for which good genetic traits are available (approximately 2000 diseases/phenotypes). Specifically in the context of Alzheimers, it would enable us to test for genome-wide associations with particular regulatory factors (e.g. transcription factors) which are directionally involved in disease. If we become able to detect such an enrichment, then we would be well positioned to start looking into drug target development.
1.2 Challenge overview

Genome-wide association studies (GWAS) are conducted to discover associations between diseases and the genetic sequences of individuals. They enable to uncover sequence mutations (single-nucleotide polymorphisms, SNPs) which are statistically enriched between groups of people. Interpreting the effect of GWAS SNP associations is arguably the primary challenge in human disease biology at this moment. The underlying datasets have been the outcome of the most substantial investment made in biological research in history. As a result, improvements in their analysis can be understood as providing one of the highest rates of return achievable in science: each computational advance can improve the analysis of all genetic datasets. In this challenge we proposed to use existing regulatory genomic data from peripheral immune cells to train machine learning algorithms to predict QTL-relationships, based on DNA sequence and cell-type specific epigenetic profiles.

Current state-of-the-art models predict local (cis-acting) effects, usually taking a continuous fragment of DNA sequence as input, and predicting the chance that the DNA is locally bound by TFs, or that it contains epigenetic modifications. This is usually modelled by fitting experimentally obtained DNA profiles as targets for multitask sequence classification (binary binding of each target), or regression (number of reads for each profile). As the input size to the model increases, the task is typically implemented as multichannel sequence-to-sequence prediction, with various levels of coarsening applied to the output (the experimental readings are rarely detailed to a level of single nucleotides - ChIP-seq TF peaks are typically a few hundred nts wide, histone modification peaks can be a few thousand nts wide).

The sensitivity of a model is limited by its input span, and the receptive field (RF), which for newest models can reach up to 200 kilobases (Enformer model [1]), but typically is a few kilobases from the predicted location. However, experimental QTL data shows that interactions can span much further (up to 1 megabase in the dataset of immune cell-types used for this challenge), because of the way the genome is folded and compacted in the cell nucleus. Currently, no methods exist in the literature to predict in silico the effect of such far-reaching mutations.
Although genomic sequence is important for predicting binding of many TFs, the sequence is generally the same across cell-types. Therefore, a more complete picture of tissue variability can only be obtained by analysing genetic and epigenetic data in concert. Studies show that a number of phenomena, such as chemical modifications to the DNA and histones which alter its shape and electric charge, and downstream properties such as chromatin accessibility, are crucial for accurate prediction \[10\]. These profiles can be measured by various types of experiments to describe the state of chromatin at given genomic coordinates, and used as inputs (conditioning) or outputs (targets) for the model. If the right cell-type specific signal (e.g. epigenetic profile) could be used to condition model prediction alongside the sequence, it might allow for adaptation to new cell types, generalisation of knowledge to avoid costly experimentation and training, and better understanding of causal phenomena underpinning gene expression.

1.3 Data

To tackle this challenge, the following data were provided:

1. DNA sequence data (A, C, G, T) - human genome build HG38;
2. Chromatin accessibility data - one dimensional tracks signifying how accessible the DNA is to protein binding along the genomic sequence:
   - DNase-seq data for 54 cell types;
   - ATAC-seq data for 4 cell types;
3. Histone modification data (H3K4me1, H3K27ac) for 58 cell types;
4. QTL data for 3 cell types (neutrophils, monocytes, T-cells).

1.4 Objectives

The main goal of the challenge was to predict the effects of sequence variation on epigenetic signals in microglia. A model capable of such prediction would allow the evaluation of functional relationships between
a large number of SNPs (out of which only some are causal QTLs), epigenetic profiles, and consequently gene expression. Due to limited data availability for this tissue we aimed to explore knowledge transfer approaches.

Current models are only able to learn from a single contiguous sequence, linear expansion of which becomes prohibitively expensive, as it requires deeper convolutions, and larger attention matrices. Increasing input length also provides diminishing returns for accuracy when the training is performed on a static reference genome. In such setup, the model is expected to predict the effect of unseen mutations, but can only learn to generalise them from different spatial relationships across the same genome. It is unlikely that such a model could accurately predict rare and specific high-dimensional effects.

While the experimental profiles for multiple genomes are unavailable at present, one approach could be to use QTL interaction data for training. A mutation could be introduced to the reference genome, and the model trained to predict the corresponding effect at a distance. While sacrificing the ability to use QTL data for validation, it might allow better understanding of trans-acting interactions by investigating the trained model with a number of feature attribution methods such as mutagenesis (perturbing input around the region of interest to see the change in output), integrated gradient analysis (averaging the non-linear response of a function), or attention matrices (weighing elements of input that are predicted to interact). QTL data indicates that these interactions can span much further than the RF of all current models, therefore investigating the ability to increase the predictive range spatially is crucial for solving this task.

An idealized model of the genome should in principle be able to infer both the graph edges (which regions interact), and how nodes in each neighbourhood influence the local prediction (by aggregating the connected nodes), while allowing generalisation across cell-types. Ultimately, we are interested in obtaining models which would be able to predict the effects of input variation in novel cell-types on targets located as far away as possible to cover all genome-wide interactions.
1.5 Approach

During the challenge, we firstly set out to explore the possibility of model conditioning with epigenetic input. Since we are interested in predicting the effects of sequence mutation on histone marks, we decided to use chromatin accessibility as a general conditioning signal. We used the Leopard model [12], which allows training a U-Net neural network sequence-to-sequence, and modified the inputs and outputs to determine how providing the accessibility tracks (DNase and average DNase difference across cell types - same as Leopard [12]) affects the prediction of histone marks (as opposed to TF binding in the original [12]). We trained this architecture on 4 mixtures of inputs and outputs, where the outputs are always trained to predict one of two epigenetic signals (H3K27ac and H3K4me1 - chromatin markers associated with genomic enhancer activity) in 3 cell types (neutrophils, monocytes and T-cells):

1. Model A - uses only sequence as input, predicts separate track for each histone modification and cell type;
2. Model B - uses sequence and local DNase signal as input, predicts one track per histone mark, for the cell type specified through accessibility input;
3. Model C - uses sequence and DNase signal as input (same as B), predicts separate tracks (same as A);
4. Model D - uses only sequence as input (same as A), predicts one track per histone mark (same as B).

The hypothesis of this experiment was that supplying accessibility information should improve the predictive performance (C should outperform A; B should outperform D). The difference in performance between models A and B should indicate how well the local accessibility signal can define the cell-type for the prediction. Model D should serve as a worst-case control, since it predicts the same output for every cell type.

Secondly, we investigated the RF of the convolutional Leopard model (U-Net), and attentional Enformer model. Any model with a RF which cannot
encapsulate the distance between the QTL and the targeted output will by
definition be insensitive at this distance, and incapable of predicting the
mutation effect on epigenetic signals. Two approaches to calculating the
empirical RF were used:

1. By evaluating the effect of forward pass through the model. First, a
prediction is made for a reference sequence. Then, a single
nucleotide substitution is introduced, and the prediction is
re-evaluated. The distance in the model output where any change
can be observed is the empirical RF, which is averaged across
multiple evaluations.

2. By calculating the backward pass (gradient) of the model input with
respect to a selected location of the output. Similarly to approach 1,
the non-zero gradient can only be observed within the RF on the
input.

Finally, to investigate the feasibility of constructing a graph model capable
of integrating non-local interactions, we analysed QTL data for 3 immune
cell-types. In our datasets, the statistically correlated interactions span up
to 1 million base-pairs (1 Megabase), which is a hard cut-off applied in
the study from which the data originates [5]. For the purpose of making a
model linking these interactions, we are interested in a number of inverse
links pointing from the effect site to the QTLs.

Specifically, if we wanted to create a model which for each prediction
region aggregates the nodes that might be functionally connected, how
many such regions do we typically observe? This information determines
the sparsity of the adjacency matrix connecting the nodes and has
implications for model design.

Furthermore, regions which are not predicted to interact with distal sites
would likely not benefit from such graph modelling approach. We
therefore created histograms representing the number of QTLs targeting
each location on the same chromosome (binned at 10k base-pairs), and
repeated the same process while also binning the QTL source location
(to avoid counting the same source region multiple times if many QTLs
are found in close proximity).
1.6 Main conclusions

Conditioning experiments indicate that the introduction of DNase tracks alongside the sequence allows reaching a lower training error, but also lowers test performance compared to using only the sequence. This is surprising, since we would expect that including additional information should only improve performance. It appears that the benefit of extra information in this architecture does not generalise to held-out regions when measured using a metric appropriate for imbalanced classes. Furthermore, the concatenation of extra channels to the input is likely to cause significantly different learning dynamics of the neural network. It is likely that the amount of information provided by DNase tracks is, per base-pair, much lower than the underlying sequence. Since the Leopard paper does not provide a comparison with and without DNase tracks, we can only conclude that this architectural setup is sub-optimal for cell-type conditioning on our histone mark data.

Receptive field analysis indicates that the RF of Leopard (with same model depth and 10KB input size) is approximately 1.1KB around the centre of mutation, or approximately 600 bases from the mutation to the target. In contrast, the RF of Enformer spans its entire input, which for the published model is 200KB (or approximately 100KB if the effect site is placed in the centre, however this is not required). It is therefore clear that the attention-based model is better suited for prediction of long-range interactions.

The analysis of QTL connectivity indicates a sparse graph of interactions, with hubs of increased connectivity (a scale-free network). In all three of our QTL datasets, the vast majority (approx. 90%) of genomic bins are not targeted by any QTLs. However, the remaining regions can be targeted by a number of QTLs simultaneously, ranging from one to over 1000 QTLs affecting a distal site. This result may vary, however, depending on the false discovery rate threshold used, and is likely to vary across datasets. In the context of models that would incorporate this information, we conclude that complex graph connectivity might only exist in a relatively small subset of regions, while most regions might not benefit significantly over currently used local modelling techniques. Which regions do benefit from extra proximal information is likely to be dynamic and cell-type specific.
1.7 Limitations

Conditioning experiments were performed using a baseline which is likely limited in terms of architecture and data. Specifically, for the convolutional model used for training:

1. Classification of a binarised ChIP-seq track, although easier to evaluate, is likely introducing a loss of information, particularly around the somewhat arbitrary position of cut-offs at the edges of each positive region;

2. Random sampling the training data genome-wide results in large class imbalance, with only a small proportion of regions containing positive training signal;

3. Concatenation of DNase tracks to the one-hot encoded sequence immediately at the level of input is likely suboptimal due to the different amount of information carried by each modality. While the sequence is informative at a base-pair resolution, the DNase signal peaks are broader and their variability across base-pairs is comparably lower. Many regions are likely to contain no variable DNase signal across training cell-types, making the signal uninformative for conditioning;

4. Differences between DNase and ATAC data, or other protocols of measuring chromatin accessibility, may make them unsuitable for being used interchangeably without identifying and correcting for such differences.

A key issue in learning functions which can predict interactions at a large distance is the exponential growth of parameters needed to provide connectivity between regions. In a convolutional model this manifests in a necessary increase of depth and size of filters to increase the RF. For self-attention models using a single contiguous input the growth of evaluated interactions is quadratic (as every latent seqlet embedding is tried against every other seqlet in the sequence). This makes the increase of input span costly, while providing diminishing returns when training on a static reference genome.

A single genome does not provide the model direct information required to learn its mutations, which can therefore only happen through
generalising across different loci, assuming spatial equivarience of interactions. This might make it impossible to predict the effects of rare disease-causing mutations in unique regions.

### 1.8 Recommendations and future work

Recommendations can be considered separately for the tasks of state conditioning and distal prediction. However, one can also express these problems jointly, by noticing that capturing distal effects requires conditioning of each local region prediction on (distal) signal containing sequence and state. Specifically, all current approaches can be generalised as graph node regression, where each node is represented by a latent vector of region embedding (of DNA sequence and possibly chromatin state over a certain genomic range), and the regression is performed to predict local state, by also aggregating influence of neighbourhood of each node (represented as a set of latent embeddings of nodes). In principle, the goal is to provide more conditioning signals (to generalise to different cellular states) and determine appropriate neighbourhoods between nodes (to allow conditioning on selected distal signals).

Ultimately, a generative model may be better suited for causal prediction. For example, convolutional variational autoencoders (VAEs) might be used to create a joint embedding of all available signals (sequence and state) for a given region. A key issue is making the local embedding shift-equivariant (so that small offsets in input do not result in different latent embeddings of state). This also precludes the use of maximum likelihood regression loss for edge inference with attention, as a model of sufficient capacity has all the information for reconstruction of each region, and does not benefit from extra distal information to minimise the local error.

In such cases regularisation with prior knowledge is required to make the model capture causal information. Adversarial contrastive methods can be used to classify valid and invalid neighbourhoods [17], and more generally test for known conditions. Regularisation often assumes that a small set of generalisable models is a better reflection of the underlying ground truth than a large number of independent parameters. For example, if TF binding or accessibility can be predicted from local
sequence by a regression sub-network, then it is preferable to learn this predictor over conditioning each region with the raw input data.

Further generalisation might also be achieved by using new methods of predicting molecular embeddings, such as AlphaFold [9] to predict 3D shape of macromolecules from sequence, and graph embedding methods to create latent representations from such shape [7]. Capturing the variability of binding domains in latent representations might enable the creation of models that generalise across different TFs (for input of DNA sequence and TF shape embedding predict the conditional probability of binding), including novel or mutated domains.
2 Data overview

2.1 Dataset description

Several different datasets were provided, containing information about the DNA sequence, and the associated chromatin state. DNA sequencing is the operation of determining the sequence of nucleotides in a DNA strand, which can contain one of four bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Sequencing can be used to determine the sequence of individual genes, full chromosomes or entire genomes of an organism. It has also become the most efficient way to sequence RNA, or infer the sequences of peptides in a protein. For model training, a single human reference genome was provided. This widely used, publicly available reference is a construct assembled from many individuals to achieve maximum sequence coverage.

Aligned chromatin accessibility data from DNase-seq and ATAC-seq experiments were also used in the training process. This cell-type specific data is provided as a general footprint of transcriptional activity and regulation. DNase I hypersensitive sites sequencing, or DNase-seq, is a molecular technique able to identify genome-wide open chromatin regions that are sensitive to cleavage by the enzyme DNase I. DNase I hypersensitive sites correlate with active gene regulatory elements, promoters, enhancers, silencers, insulators and locus control regions. Identifying such regions are key in understanding gene regulation and thus the mechanisms governing biological processes during development and disease. Similarly, the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) assesses genome-wide chromatin accessibility but is considered a faster and more sensitive analysis of the epigenome than DNase-seq. Accessible chromatin regions are identified using the hyperactive mutant Tn5 Transposase which inserts sequencing adapters into open chromatin regions.

Quantitative, or complex, traits are measurable phenotypes that are influenced by genetic and environmental factors. These traits can vary among individuals, over a range, to produce a continuous distribution of phenotypes. A gene or chromosomal region that contribute to a quantitative trait is called Quantitative Trait Locus (QTL). In our data set, QTLs signify the locations of Single Nucleotide Polymorphisms.
(sequence variants, SNPs), that are statistically inferred to contribute to a varying phenotype, such as epigenetic state, between populations. QTLs associated with the function or activity of regulatory elements often exert their influence in a cell-type specific manner. This type of interactions are known to occur in Alzheimer’s disease, but the functional effect of sequence mutations is not well understood.

We restricted our models to three cell types in order to speed up the training, and because QTL data was available for these cells. Monocytes are a type of white blood cell which are part of the vertebrate innate immune system and can differentiate into macrophages and dendritic cells. This cell population can be divided into subpopulations depending on the presence or absence of the surface markers CD4 and CD16. In our case the monocytes harboured CD4 but were depleted of CD16. Neutrophils represent the most common type of white blood cells and are the first line of defence after an infection, where they exert phagocytic activity. CD4+ T cells or T helper cells are part of the adaptive immune system and are capable of recruiting other immune cells through the release of cytokines.

The reason why solely immune cells were chosen to be part of the model training is that the cell type of interest, microglia, are a resident immune cell population of the brain and play a key role in the onset and progression of Alzheimer’s disease (AD).

In the following the data types are described in more detail and limitations are addressed.

2.1.1 DNA sequence data

For this challenge, we have used HG38 (Human Genome 38)\footnote{https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39} data, the reference human genome assembly released on December of 2013.

We use character-level one-hot encoding to represent the categorical data. For one hot encoding, each base is encoded as a vector of all zeros except one in a specific position, A is encoded as (1,0,0,0), T as (0,0,0,1), C as (0,1,0,0), and G as (0,0,1,0). We were provided four bigWig files for A, T,
C and G containing information for 23 chromosomes namely chromosome 1 to 22 and chromosome X.

2.1.2 Chromatin accessibility data

DNase-seq and ATAC-seq data for 55 cell types was originally obtained from EpiMap\(^2\) (Epigenome Integration across Multiple Annotation Projects), a repository of curated epigenome data from different projects such as ENCODE, Roadmap and GGR. Most of the experimental data came from DNase-seq experiments, only three cell types had ATAC-seq data instead. These were sources denoted with the following accession numbers: BSS00355, BSS00361, BSS00367. Although those data was available, the chromatin accessibility data from these 55 cell types were not used in this project. Instead, we focused on training models using immune cell lines from *mature neutrophils, CD4+ CD16- monocytes, and CD4+ T-cells*. While monocytes and T-cells had DNase-seq data available, neutrophils had ATAC-seq data. All data was downloaded from the BLUEPRINT consortium\(^3\).

The bigWig chromatin accessibility data from ENCODE EpiMap is pooled in 25 base pair intervals averaging the -log10 p-value track that is obtained from MACS2 (peak caller software used to align raw sequencing reads to the reference genome and identify significantly enriched regions called peaks). We do not perform an exact conversion to peaks, but instead take intervals with signal above 2 as positive examples for binary classification.

2.1.3 Histone mark data

Our data contains the epigenetic state signal of the two histone marks, H3K27ac and H3K4me1 for each cell type, except of microglia. H3K27ac mark refers to the acetylation of histone H3 at lysine residue 27, whereas H3K4me refers to methylation of histone H3 at lysine 4. Both marks represent active enhancer marks and are characteristic for an actively transcribed region. The data tracks for both marks were obtained from ChIP-seq experiments and can be expected to have a resolution of about

100bp. The data is available as a real-valued signal aligned to the reference DNA sequence. NaN values in the sequence were replaced by 0.

Similarly to accessibility data we used a threshold of 2 to binarise the signal into negative regions (value lower than 2) and positive regions (value larger than 2) for model training.

2.1.4 Quantitative Trait Loci data

In our dataset, QTLs are associated with regulatory elements marked with H3K27ac and H3K4me1 in specific cell types (CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells). Thus, we have six combinations of regulatory elements and cell types. Only significant QTLs are selected based on False Discovery Rate (FDR) threshold of 0.05. Only QTLs with effect target on the same chromosome are considered, within the maximum range of 1 megabase (see Figure 1). Further exploration is performed in order to understand the relationship between SNP position (single base-pair mutation), and the target region which the QTL affects (a span of distal sequence where the epigenetic signal is affected by the SNP).
Figure 1: Significant QTLs in BLUEPRINT datasets with effect site located at given distances within a chromosome.
2.2 Data quality issues

2.2.1 Gaps in the DNA sequence

While analysing the DNA sequence data, we found a small number of unassigned nucleotides across the chromosomes. These gaps can have different origin. The HG38 assembly contains the following principal types of gaps:

- short arm - short arm gaps (count: 5; size range: 5,000,000 - 16,990,000 bases)
- heterochromatin - heterochromatin gaps (count: 11; size range: 20,000 - 30,000,000 bases)
- telomere - telomere gaps (count: 48; all of size 10,000 bases)
- contig - gaps between contigs in scaffolds (count: 285; size range: 100 - 400,000 bases)
- scaffold - gaps between scaffolds in chromosome assemblies (count: 470; size range: 10 - 624,000 bases)

We calculated the fractions of unassigned nucleotides for each chromosome (Figure 2). Such gaps in the DNA sequence might confuse the model at training time and might need to be excluded from analysis. Due to the limited time frame of this project, we did not treat unassigned nucleotides any further in our processing and included them as-is in our model training (with all channels encoded as 0).

2.2.2 Gaps in target profiles

All profile data (DNase, ATAC, histone marks) included occasional not a number (NaN) values. The ATAC-seq data for the chromosome 5 of our neutrophil sample, for example, contained 45% of NaN values and no 0 values. By contrast, the DNase-seq profile for the T-cell sample only contained only 5% NaN values on chromosome 5 and a few 0 (4%) values.

These samples were not cherry-picked but obtained via random sampling of a chromosome.
While NaN values can indicate that a certain data point was not measured, they may equally well correspond to no observable signal, i.e. values close to 0, invalid measurement reads, or over-saturated signals. Depending on their meaning, it might be best to replace NaN values, impute them from adjacent data values, or exclude sequence snippets containing NaNs entirely. Unfortunately, our data sources did not include information on the meaning of NaN in the data, which limited our ability to treat them adequately.

To train the models in this project, we resorted to simply replacing NaN values by 0. For the DNase and ATAC signals this means that we treated NaN valued chromatin accessibility data as inaccessible chromatin. For the histone marks, replacing a NaN by 0 means that we assume the absence of histone marks in all NaN regions. For model scoring, we excluded histone mark predictions in regions with NaN values in the true histone mark labels to avoid affecting test scores.
2.2.3 Mixing of DNase-seq and ATAC-seq data for chromatin accessibility

Most of the samples were provided with associated DNase-seq data (monocytes, T-cells). Some cell lines (e.g. neutrophils) were provided with ATAC-seq data instead. Although both assays are broadly used to assess genome-wide chromatin accessibility and are based on cleavage enzymes, their footprints might differ, as illustrated in Figure 3. Since we only had either one of them available for each cell type, we resorted to using the tracks interchangeably after performing quantile normalisation.

Figure 3: Schematic illustration of differences between DNase-seq and ATAC-seq profiles. These differences might lead to problems when using them interchangeably, as done in the current analysis. Figure taken from [16].

While DNase-seq and ATAC-seq experimental data are highly correlated (~ 0.8 correlation⁶), using them interchangeably might cause improper generalisation of features by the model.

⁶We received a correlation analysis from Alan Murphy, but cannot disclose it here as it is part of a yet unpublished paper.
3 Data visualisation

3.1 QTL connectivity

We used QTL data as a reference of possible functional connectivity across the genome in order to investigate the feasibility of training a model conditioned on long-range interactions. This data consists of links which connect individual SNP mutations to distal regions in the genome where they presumably exert an effect (see Figure 4). We wanted to know how many regions are presumed to have distal links, and how many links are expected to occur per region. This information would influence model design in several ways. Regions with no inferred distal connectivity would be treated identically to currently used local prediction. For interconnected regions, the construction of a neural network and the implementation of neighbourhood aggregation needs to scale appropriately given a number of expected links that each region might have.

We performed two types of calculation to understand the density of QTL connectivity between regions, which we term connected QTLs and binned QTLs. Firstly, we bin each chromosome at 10,000 base-pair intervals. For example, chromosome 1 has length of approximately 240 million base pairs so it can be divided into 240 thousand bins. To obtain the distribution of connected QTLs we calculate the number of QTLs that target each bin. Therefore, if a model was created to predict the epigenetic signal at the
Table 1: Number of individual QTLs connected to binned target regions.

<table>
<thead>
<tr>
<th>Connected QTLs</th>
<th>mono K27ac</th>
<th>neut K27ac</th>
<th>tcel K27ac</th>
<th>mono K4me1</th>
<th>neut K4me1</th>
<th>tcel K4me1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected QTLs</td>
<td>4642287</td>
<td>4124684</td>
<td>392350</td>
<td>12771939</td>
<td>10379199</td>
<td>3936845</td>
</tr>
<tr>
<td>Connected regions</td>
<td>24253</td>
<td>21104</td>
<td>19486</td>
<td>53278</td>
<td>45371</td>
<td>23496</td>
</tr>
<tr>
<td>All regions</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
</tr>
<tr>
<td>Avg QTLs per connected region</td>
<td>191</td>
<td>195</td>
<td>184</td>
<td>239</td>
<td>228</td>
<td>167</td>
</tr>
<tr>
<td>Avg QTLs per region</td>
<td>15</td>
<td>13</td>
<td>11</td>
<td>42</td>
<td>34</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 2: Number of binned QTL regions connected to binned target regions.

<table>
<thead>
<tr>
<th>Binned QTLs</th>
<th>mono K27ac</th>
<th>neut K27ac</th>
<th>tcel K27ac</th>
<th>mono K4me1</th>
<th>neut K4me1</th>
<th>tcel K4me1</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL bins</td>
<td>650762</td>
<td>572460</td>
<td>500804</td>
<td>1735453</td>
<td>1418704</td>
<td>571662</td>
</tr>
<tr>
<td>Connected regions</td>
<td>24253</td>
<td>21104</td>
<td>19486</td>
<td>53278</td>
<td>45371</td>
<td>23496</td>
</tr>
<tr>
<td>All regions</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
<td>303114</td>
</tr>
<tr>
<td>Avg QTL bins per connected region</td>
<td>26</td>
<td>27</td>
<td>25</td>
<td>32</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>Avg QTL bins per region</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

As expected, there are more connected QTLs than binned QTL regions, shown in Table 1 and 2. The average number of QTLs targeting each region is significantly lower if we consider the QTLs originating in close proximity as a single link. Furthermore, it can be seen that the vast majority of regions is not targeted by any significant QTL.

For each QTL dataset we then created histograms of the number of connected QTLs (single SNPs and binned regions) per each target region. Our data consists of six combinations of regulatory elements and cell types: monocytes H3K27ac (Figure 5), neutrophils H3K27ac (Figure 6), T-cell H3K27ac (Figure 7), monocytes H3K4me1 (Figure 8), neutrophils H3K4me1 (Figure 9), and T-cell H3K4me1 (Figure 10).
Regions without any associated QTL account for the vast majority of the total (about 92%) in both categories. Majority of regions associated with QTLs tend to have less than 100 links, however, a significant number of regions also exists with over 100 links (up to around 1000 links for the most targeted regions). For connected QTLs, 3.8% of bins have more than 100 QTLs. In case of binned QTLs, only 0.3% of bins have more than 100 QTL regions targeting them. This shows that in many cases the QTLs targeting a given region are located close to one another.

Overall, it can be seen that there are few target loci with large number of QTL effects, while there are many loci with small or no QTL targeting them at all. This information could be useful for determining the correct graph modelling approach, as a small number of regions with large number of QTL effects could be treated as hubs (nodes with a number of links that greatly exceeds the average). Networks containing such hubs are known as scale-free networks [3].
Figure 5: Top: Connected QTLs per region. Bottom: Binned QTLs per region for monocytes H3K27ac.
Figure 6: Top: Connected QTLs per region. Bottom: Binned QTLs per region for neutrophils H3K27ac.
Figure 7: Top: Connected QTLs per region. Bottom: Binned QTLs per region for T-cells H3K27ac.
Figure 8: Top: Connected QTLs per region. Bottom: Binned QTLs per region for monocytes H3K4me1.
Figure 9: Top: Connected QTLs per region. Bottom: Binned QTLs per region for neutrophils H3K4me1.
Figure 10: Top: Connected QTLs per region. Bottom: Binned QTLs per region for T-cells H3K4me1.
4 Experiments

4.1 Model architecture

Most deep learning models can be expressed as a series of calculations aggregating information on a graph \([4]\). In case of convolutional models for DNA sequence, initial layers of convolution are used to create a dense embedding of underlying fragments of sequence (motifs or seqlets), while the following layers are trained to aggregate their interactions and make a prediction. In a purely convolutional model, the interaction between these nodes has a fixed structure, where neighbours proximal in sequence are always presumed to interact (edge weights are 1 between close neighbours and 0 otherwise). In an attentional model, the interactions between nodes are predicted dynamically for a given input, implementing a form of parameterised sum-pooling. While the interactions can be evaluated between freely chosen query and key inputs, even at a large distance, the consequence is invariance to mutual positioning (permutation) of nodes. This information has to be supplied back to the model (in the form of basis functions) if global spatial relationships are to be considered.

For this challenge we decided to evaluate and adapt an existing sequence-to-sequence convolutional model called Leopard, based on the U-Net architecture \([12]\). Leopard aims to identify transcription factor binding sites at single-nucleotide resolution in a cell type-specific manner. The model takes DNA sequence data and DNase-seq accessibility data as inputs, and outputs TF binding probability for each input location. The model performs binary classification of each nucleotide position (binding/no binding) for each predicted track. The performance of Leopard was assessed by its authors using the area under receiver operating characteristic curve (AU-ROC), and area under precision-recall curve (AU-PRC) due to large class imbalance between positive (binding) and negative (non-binding) regions.

The architecture of Leopard consists of two main components: the encoder and the decoder. Encoder comprises repeating convolution-convolution-pooling (ccp) blocks, the decoder consists of repeating upscaling-convolution-convolution (ucc) blocks (see Figure \([11]\). We follow this exact architecture in our Model B, which receives a total of
6 input channels: the first four channels contain the one-hot encoded DNA sequence, the remaining two contain a quantile normalised DNase signal and an reference DNase signal (obtained from averaging DNase signals across 55 cell types from the ENCODE challenge). The full network consists of 5 convolution blocks and 5 deconvolution blocks with shortcut connections between them. Each convolutional block contains two convolutional layers with a kernel size of 7 and stride 1, followed by a max-pooling step with pool size 2. Layers are batch normalised between convolutions. We transform the two output channels with a sigmoid transformation to lie between 0 and 1. The two output channels correspond to the probabilities of histone marks $H3K27ac$ and $H3K4me1$ in the given cell-type.

4.2 Conditioning experiments

Based on the Leopard model architecture we set out to assess the different conditioning approaches to predict the cell-type specific epigenetic state determined by the histone marks $H3K27ac$ and $H3K4me1$. We trained similar models using different combinations of input and output data.

Model A was trained on DNA sequence data only, and predicted two epigenetic histone mark tracks ($H3K27ac$ and $H3K4me1$) for three different immune cell types (neutrophils, monocytes and T-cells). Thus, model A receives the input via four channels and outputs six channels. It is of note that model A is similar in its input/output setup to many existing multitask models that predict from primary DNA sequence, including Enformer [1].

Model B was conditioned on accessibility data from DNase-seq experiments in addition to the sequence information. In total model B has six input channels, four for the DNA sequence, one for cell-type specific DNase-seq, and one for the averaged DNase-seq signal across cell-types (same as in Leopard). The output for model B are predictions for the two epigenetic tracks $H3K27ac$ and $H3K4me1$ for the cell-type specified by DNase conditioning. Model B is architecturally the same as Leopard, but predicting histone marks instead of TF binding.

Two further model architectures were used in the conditioning experiments to provide a reference. Model C uses sequence and
accessibility as input, and predicts separate tracks per histone mark per cell type. The accessibility information consists of two tracks (same as Model B), where one is the averaged profile from all cell-types, and the other a (randomly selected) profile from one of the target cell-types. Model C is therefore similar to multitask model A, but with extra accessibility information provided as input.

Model D uses only sequence as input, and predicts one output per histone mark, which is shared by all cell types (receives no cell-type information, and outputs the same result for each cell-type). It is therefore similar to model B without any conditioning.

Table 3 lists the tested architectures with their respective input and output channels.

<table>
<thead>
<tr>
<th>Model</th>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Sequence (4)</td>
<td>Histones mark per cell type (2x3=6)</td>
</tr>
<tr>
<td>B</td>
<td>Sequence (4), accessibility (2)</td>
<td>Histone marks (2)</td>
</tr>
<tr>
<td>C</td>
<td>Sequence (4), accessibility (2)</td>
<td>Histone marks per cell type (2x3=6)</td>
</tr>
<tr>
<td>D</td>
<td>Sequence (4)</td>
<td>Histone marks (2)</td>
</tr>
</tbody>
</table>

Table 3: Number of input and output channels for models A-D.

All models were trained on the three immune cell lines (mature neutrophils, CD4+ CD16- monocytes, and CD4+ T-cells) with chromosomes 1, 8 and 21 held out for testing. The choice of chromosomes used for evaluation follows Leopard implementation [12]. For training, 100,000 random genomic regions of 10,240bp window-size were sampled from the training chromosomes; hyperparameters such as learning rate (1e-3) and learning rate decay (1e-5) were used, same as in Leopard. We binarised the histone mark signal with a threshold of 2 to train the models as a binary classification task, using binary cross-entropy loss for optimisation (using Adam optimiser). All models were trained using the default settings of 5 epochs with a batch size of 128. No hyper-parameter tuning was performed at this stage to establish baseline models. The training curves can be seen in Figure 12.

Evaluation of models was performed using the area under the precision recall curve (AU-PRC) which was calculated on test chromosomes 1, 8 and 21. This metric is appropriate for measuring classification
Figure 11: Schematic of Model B (adapted from [12]). U-Net architecture is used to predict two histone mark profiles from input DNA sequence (4 channels) and concatenated accessibility information (2 channels).
performance with class imbalance in the test set. Figure 13 shows test performance of the four models on predicting the two histone labels H3K27ac and H3K4me1.

The results indicate that the lowest training loss is reached for models B and C, which both receive accessibility information in the input. Model B can reach the lowest training loss, as it predicts the marks specific to the input conditioning. As expected models A and D, lacking accessibility information, reach worse (higher) training loss, in particular model D which outputs the same values for each cell-type.

However, test results indicate that model A performs best, followed by model D. This was surprising, since both models take only the DNA sequence data as input, and therefore cannot use extra accessibility information. Concatenating the chromatin accessibility data as additional input channels appears to result in a significant drop in test performance for both models in this test.

The test metrics do not follow the same ranking as the training losses, indicating that while the accessibility information might provide benefit to the training of this architecture, this benefit does not appear to generalise.
to held-out test data, and when using a metric appropriate for imbalanced classes (PR-AUC). We therefore conclude that the apparent benefit in training is likely due to overfitting to the extra input information, rather than improved prediction of transferable causal effects.

Figure 13: Test performance (Precision-Recall Area Under Curve).
4.3 Receptive field analysis

The receptive field (RF) of the model is the span of the model input which can contribute (in any magnitude) to the calculation of a given output. Inputs outside of RF by definition can not influence a given prediction. The calculation of RF enables us to define a maximum distance at which a convolutional model with (non-global) pooling can theoretically remain sensitive to input variation.

The definition of a RF mirrors the concept in biological neurons of the eye retina. According to Wikipedia [18], the receptive field (of a biological neuron) is “the portion of the sensory space that can elicit neuronal responses, when stimulated”. The sensory space can be defined in any dimension (e.g. a 2D perceived image for an eye). In the context of deep learning, the definition could be “the region of the input space that affects a particular unit of the network”. It is therefore crucial to understand the receptive field because it provides a description of the location at which a stimulus must be presented in order to elicit a response from an output, which in our case could tell us the impact of sequence mutation on predicted distal epigenetic signal.

The RF for Leopard was not given in the paper [12], we determine it in order to understand the input sensitivity of U-Net-based models such as models A-D.

Further, we analyse the RF of Enformer, which uses attention-pooling to aggregate distal information.

To estimate the empirical receptive field of a model, we use two methods.

1. **Window perturbation method**: A random one-hot encoded DNA sequence is generated. This sequence is forward passed through the model to produce a baseline output. A copy is mutated at a single position (preferably at the center to avoid edge effects) and then passed through the network to obtain a perturbed output. Subtracting baseline output from the perturbed output results in a difference vector, where non-zero values show which elements were affected by the given input feature that was mutated. The procedure is illustrated in Figure [14]. This process is repeated multiple times
with random inputs to determine the maximum affected output span. This method does not require being able to compute the gradient of a model and was particularly useful for calculating the RF of Enformer (calculating the gradient requires more memory than a forward pass).

2. Gradient-based method: This method directly yields the empirical receptive field, but requires the ability to compute the gradient of a model. As a first step, a random DNA sequence is generated. The sequence is subsequently passed through the model to predict an output sequence. We then calculate the gradient of an output feature of interest (e.g. the centre feature) with respect to the input. The input values with non-zero gradient are within the receptive field. An illustration of the method is shown in Figure 15. The process is repeated multiple times and the results are averaged.

![Diagram of perturbation-based effective receptive field computation method](image)

**Figure 14:** Perturbation-based effective receptive field computation method.
We observe that the maximum receptive field of our Leopard-like (U-Net) model A is around 1100bp (Figure 16). This indicates that the maximum range of non-local interactions that this model will remain sensitive to is only around 550bp from the point of mutation.

We performed the analysis using a trained and an untrained version of the model. The untrained model was evaluated after random weight initialisation, without any gradient descent training. In principle, the receptive field should remain the same in both cases, as its range is determined only by the depth of the network, filter span, stride, and pooling operations. This is indeed the case (Figure 17 and 18), however, we observe than an untrained network is much less sensitive to input outside a very narrow middle range compared to a trained model. This indicates that while network structure determines the theoretical limits of RF, training greatly increases the level of sensitivity to relevant features throughout the RF.
Figure 16: Perturbation method: RF of trained model A (U-Net). Effect shown in log-scale on the right. Position is relative to the mutation location in the input.

Figure 17: Perturbation method: RF of an untrained model A (U-Net). Effect shown in log-scale on the right. Position is relative to the mutation location in the input.

Figure 18: Gradient method: RF of trained (left) and untrained (right, log scale) model A. Position is relative to the prediction location in the output.
Figure 19: Perturbation method: Enformer RF when perturbed at the center of input.

The RF analysis of Enformer (Figure 19 and 20) indicates that the model remains sensitive to input perturbation throughout its entire span (200,000bp). Attention mechanism can select features from any input position to attend to any output position, therefore the RF is equal to the maximum distance at which a mutation can influence the outcome (unlike U-Net where this distance is half the RF). We observe, however, that when mutations are placed at the centre of input (Figure 19), the effect of mutation is largely concentrated around this centre. This might indicate lower sensitivity at larger distance, but might also be an expected outcome of random mutations primarily affecting their local chromatin environment.
Figure 20: Perturbation method: Enformer RF when perturbed at the edge of input. Top bars indicate the location of perturbation.
5 Future work and research avenues

Capturing the available long-range QTL interactions requires a model which remains sensitive to mutations up to 1 million base-pairs away from their target. At present, the model with the largest receptive field is Enformer, reaching up to 200KB, with typical convolutional models having RF less than a few KB. Therefore, no current deep learning approach is sufficient to model known interactions. Given that DNA is coiled in the nucleus, rather than a straight line, it could be beneficial to capture 3D genomic distances in the model, instead of a linear sequence.

Graph modelling can be considered as a way to deal with the challenge. The basic idea of graph modelling is that based on graph structure, machine learning models can learn the relationships or links between nodes in the graph, and the information learnt from relationships of neighboring nodes is useful in predicting state of the target node. One way to learn such relationships is through message passing process (Figure 21) which updates the latent state of each node by aggregating its neighbour information, sequentially updating neighbourhoods in the graph.

![Figure 21: Illustration of a graph and a GNN. Adapted from 15.](image)

Our analysis indicates that the graph created using the available QTL data would contain a sparse structure, with many nodes (regions) not having distal functional links, and only interacting with their sequence neighbours. A significant number of regions, however, form hubs of hundreds or thousands of potential connections. This type of network
holds a scale-free property, and is widely observed in nature, for example in a yeast interactome (Figure 22).

Figure 22: Example of a scale-free property: yeast protein interaction network. While most proteins participate in only a few interactions, a few participate in many dozens. Adapted from [2].

The modelling of sparse interactions therefore appears to be an important avenue of research, with applications to protein-DNA, protein-protein, and other types of macro-molecular binding phenomena. Attention-based mechanisms seem particularly well suited to modelling such interaction programs. One can observe similarities between programs learned by transformer-type neural networks, Neural Turing Machines (which dynamically select appropriate instructions [8]), and sequential interactions of macro-molecules, such as proteins. Such biological objects contain multiple interaction domains, which only interact under the right conditions. An attention mechanism can evaluate the existence of interactions between molecular domains (check if they implement the right class protocol), and produce the result of the interaction (return the
output of the function). Deep learning therefore appears particularly well suited for end-to-end learning of programs on graphs of molecular interactions.

Detailed recommendations are given below to address specific aspects of the modelling approaches discussed in this report.

1. In order to (locally) condition the prediction on signal additional to DNA sequence consider:

   (a) Differences in the amount of information each modality provides per base-pair. Supplying each type of input through modality-specific initial layers, and concatenating the latent variables deeper in the network will allow better hyper-parameter adjustment, and should result in a more balanced information flow;

   (b) Downsampling the local epigenetic/TF profiles (binning the input). After the input of certain tracks is downsampled, it could be provided for a region wider than the input sequence itself (in combination with 1a.). Increasing the window of conditioning signal results in more informative global state being available to the model;

   (c) Using global cellular state as a conditioning vector. Experiments in relevant tissues, in particular at a single-cell level, can be used to obtain dense embeddings of state, e.g. scRNA-seq provides information about protein expression (typically 10k+ genes), scATAC-seq provides information about open chromatin regions (typically 100k+ binarised regions). Such information can be embedded into a dense lower-dimensional space, providing a single vector descriptive of expression and accessibility differences between cell-types. In contrast to using a categorical conditioning variable (e.g. one-hot encoding signifying one-of-many cell-types), a dense embedding should provide a meaningful distance measure, which would better allow the model to generalise by learning that cells similar in expression or accessibility space may have similar ChIP profiles at certain sequences.
2. To enable distal prediction of QTL effects consider:

(a) Defining a sampling distribution of possible trans-acting functional elements. Our QTL data indicates that only around 10% of regions would benefit from additional distal information, however this depends on thresholding of significant interactions. Apart from GWAS data, other experiments can be used as a source of presumed interactions: chromatin conformation capture (e.g. single cell Hi-C) provides information about chromatin looping and is also available in microglia [19]. Single-cell ATAC-seq co-accessibility can be used to infer correlated functionality indirectly;

(b) Training a proof-of-concept convolutional model where the embedding of distal regions is supplied to the network as conditioning for the local prediction. Increase in predictive performance, or gradient flow through the conditioning channel, would suggest that the chosen extra region is functionally connected. A small number of distal regions could be sampled (from links assumed in 2a.) to prevent overfitting to known edges. Without informative distal conditioning such a model should reduce to standard local prediction (see Figure [23]). Further, QTLs could be used for training by introducing the mutation and training to predict the effect on the target site;

(c) Attentional models can be used to predict the edges of distal interactions automatically. The problem of quadratic complexity in evaluating attention matrices is often addressed by performing sparse attention [13], where not all keys are matched to all queries. In our case, a local prediction (query) could be evaluated in training only on distal regions (keys) which are presumed to interact. Basis functions should be selected to provide meaningful information about mutual positioning of regions at a large, possibly discontinuous genomic distance;

(d) After training, such models could be used to evaluate connectivity in arbitrary locations (not used in training). One might also attempt to use a model to predict on a wider window than it was trained on. The validity of such predictions will
largely depend on the type of pooling functions and basis functions used in the model.

3. General remarks:

(a) Deep learning models are sensitive to the distribution of the training data. Consider performing the training only on a subset of significant regions (e.g. peaks with variable chromatin accessibility, with known distal connections). Consider real-valued regression to fully utilise the training signal and adapt metrics accordingly;

(b) Applying the same loss to multiple tracks with different distributions is likely sub-optimal and may violate modelling assumptions (homoscedasticity for least-squares regression, variance equal to mean for Poisson regression). Visualising the distribution of residuals (errors) against ranges of data values may help to determine the correct type of loss function. Negative binomial regression can be used for multiple targets with a dynamically determined dispersion for each [6];

(c) The discriminative type of models discussed here is likely to lose sensitivity to the causal signal of interest when the conditioning signal (input) becomes redundantly informative of the target prediction (output). For example, accessibility signals might be correlated with binding of chromatin remodeller proteins, which in turn bind to specific sequence motifs. However, providing such signals as input may make the model insensitive to the underlying sequence, and therefore not responsive to tested mutations. Generative modelling and additional regularisation might offer a further solution to this problem.
Figure 23: Schematic of a regression model incorporating distal as well as local conditioning.

6 Team members

Principal Investigator

Mike Phuycharoen - postdoctoral researcher at the University of Manchester.

Facilitators

Alan Murphy - first-year PhD student at Imperial College, London.

Christina Morgenstern - university lecturer and trained molecular biologist.

Umran Yaman - first year PhD student in Hardy Group in Institute of Neurology, University College London.

Participants

Simon V Mathis - first year PhD student in Computer Science with Prof. Liò at the University of Cambridge.

Sethuraman T V - Research Associate at Games 24x7, India.
Thu Trang Dinh - PhD student in Business and Management at the University of Manchester. Her research focuses on integrating multiple sources of information for exchange rate forecasting via the adoption of Machine learning combining with Network analysis.

Melania Abrahamian - Data scientist and Project manager as PostDoc at UCLA. Her research is on the application of bigdata and ML in cancer and neurodegenerative diseases (personalised medicine).

Prachi Sharma - working in PriceWaterhouseCoopers India in data and analytics competency. Her major focus area is Business Intelligence and Visualisation Domain.

Hanz Tantiangco - PhD student in the Information School at the University of Sheffield. His research interests are in chemoinformatics and deep learning.

Sukanya Mandal - Data Scientist by profession having experience in various domains. Her interest lies in using Data Science techniques for addressing problem statements for the greater good of the society and humans.

Srijit Seal - PhD student at the University of Cambridge. He works at the Center for Molecular Informatics, Department Chemistry, using machine learning for toxicity prediction.
References


