# Epigenetic Data Boosts the Accuracy of Recombination Hotspot Prediction by Machine Learning Models

Joy Linyue Fan
Biological Engineering
Massachusetts Institution of Technology
jlfan@mit.edu

Lawrence Wong
Computer Science and Molecular Biology
Massachusetts Institution of Technology
lcwong@mit.edu

***Abstract*— Genetic recombination plays an integral role in generating genetic diversity in a population, but the mechanisms of the processes governing double-strand break (DSB) formation and subsequent ligation remain poorly understood. Recent advances in machine learning as applied to genetic data have demonstrated an ability to predict the location of recombination hotspots in the genome based on raw DNA sequences. However, these models neglect potential contributing factors from epigenetic marks and chromatin structure. Specifically, H3K4me3 and H3K36me3 are known to be correlated with the activity of PRDM9, a zinc finger protein that plays a role in determining sites of recombination in humans and mice, and open chromatin structure is required for the activity of the DSB-forming protein, Spo11. Furthermore, some correlation may exist between hotspot regions and SNP density. We demonstrate using simple classification models that the accuracy of hotspot prediction is significantly improved with the inclusion of ChIP-Seq epigenomic data, DNase hypersensitivity data, and Single Nucleotide Polymorphism (SNP) density data. A similar trend was observed in our deep learning model consisting of a hybrid deep convolutional and recurrent neural network trained on the new datasets as added features. This allowed us to produce a comprehensive predictive model for locations of hotspots in the human genome. Concurrently, we utilized the Gibbs sampling motif discovery technique in an attempt to discover binding motifs for Spo11 and PRDM9. These results combined will help shed light on the mechanisms of recombination and set the stage for better informed GWAS and linkage analysis studies.**

## I. INTRODUCTION

As one of the key driving factors behind evolution, recombination has played an increasingly significant role in genetic analyses such as GWAS and linkage analyses, both of which have been foundational in furthering our understanding of human inheritance patterns. On a small scale, there is a clear correlation between recombination hotspots, areas of the genome where recombination occurs at a high frequency, and correlations/lack of correlations in genetic markers such as SNPs. However, longer-range patterns are more complex due to the uncertain behaviors of weaker recombination hotspots [11]. Furthermore, errors due to recombination, such as translocation events and other mispairings, are heavily implicated in genetic diseases. Non-homologous recombination events and unequal homologous recombination often predicate cancer and inherited disease [12]. Other errors in recombination, such as gene duplications or deletions, creation of hybrid genes, and formation of mismatched coding and regulatory sequences have been shown to be a cause of disorders in steroid biosynthesis and a contributor towards hypertension [13].

Thus, it becomes increasingly important to understand the processes governing genetic recombination. Meiotic recombination occurs during prophase I of the cell cycle when homologous chromosomes are held in close proximity to one another via the synaptonemal complex. Segments of the chromosomes then exchange genetic material with one another in a process known as crossing over. This process is mediated by Spo11, a type II topoisomerase-like protein, which produces double-strand breaks (DSBs) in the DNA that are asymmetrically nicked to produce 3' overhangs. These overhangs then allow for strand exchange by homologous repair of some subset of DSBs, leading to the formation of a recombinant product [1].

While the specific targeting of sequences by Spo11 remains an elusive process, it has been demonstrated that Spo11 possesses some sequence bias and may play a role in determining where DSBs are formed [2]. Recent advances in technology such as high-density microarray analysis and high-throughput linkage disequilibrium (LD) mapping in humans have allowed for the generation of maps of recombination hotspots. In humans, it has been observed that genetic crossovers occur at high frequencies within 2 kbp regions that are spaced 50-100 kbp apart, with approximately 72% of

these crossovers overlapping a recombination hotspot [1], [3]–[5]. Thus, the ability to predict the distribution of recombination hotspots constitutes a preliminary step towards generating a predictive model for recombination itself.

However, most current models fail to take into consideration the influence of higher-order chromatin information and other genetic regulators on the formation of DSBs. For example, open chromatin structure may be required for Spo11 to access the DNA [6], [7]. Thus, we propose to include DNase Hypersensitivity data as a method of determining the locations of open chromatin. Additionally, a correlation has been observed between the presence of DSBs and the enrichment of tri-methylated H3K4 (H3K4me3). Known to be a marker of active transcription, H3K4me3 has also been demonstrated to be a marker for regions where meiotic recombination is initiated [8]. Lastly, the density of SNPs in a region may also be indicative of active hotspots, due to an increased amount of gene conversion [25].

Notably, it has been shown that the zinc finger protein PRDM9, which has a histone methyltransferase domain, plays a significant role in the determination of hotspot localization in humans and mice. It is hypothesized that when PRDM9 binds, it promotes the enrichment of H3K4me3 on nucleosomes in its vicinity, which has downstream effects on the activation of recombination machinery and the ultimate recruitment of Spo11 [9]. In hotspots governed by PRDM9, it has been shown that the presence of both H3K4me3 and H3K36me3 is a stronger predictor than each mark alone [17]. PRDM9 is known to have an affinity for a previously identified degenerate 13-mer hotspot motif [9]. This degenerate motif was identified through LD studies in humans and was reported to be critical in recruiting crossover events to at least 40% of all human recombination hotspots [10]. The existence of a recognizable motif in hotspots suggests a further degree of predictability in the location of DSB formation. This raises the possibility of applying motif discovery techniques to further elucidate the process of PRDM9 binding.

We developed a more robust predictive model for genetic recombination using the above data as features to train a deep learning neural network. We combined epigenetic marks including H3K4me3 and H3K36me3 with DNase hypersensitivity data, SNP density data, and raw DNA sequence to train a classifier to predict the precise location of hotspots of recombination. We also used Gibbs sampling as a motif discovery technique to search for the presence of hotspot motifs that may aid in our understanding of this fundamental biological process.

## II. MATERIALS & METHODS

The HapMap Phase I, II datasets include a finescale genetic map of locations with high recombination rates across a variety of human populations [19]. The NIH Roadmap Epigenomics Project provides H3K4me3 and H3K36me3 chromatin mark annotations and DNAse I Hypersensitivity data for the same genomic regions as the recombination hotspots [22]. SNP data were obtained from the NCBI SNP database [32]. For each chromosome, we constructed a length interval $l$ by number of features $f$ matrix such that the first four columns were one-hot encodings of the raw DNA sequence and the last four columns were binary indicators of the exogenous variables e.g. chromatin marks, DNase Hypersensitivity, and presence of an SNP at the corresponding genomic coordinate. Therefore, each data point was represented as a matrix $D \in \mathbb{R}^{lxf}$ (Fig. 1). For consistency, all data were mapped to human reference genome GRCh37, and all epigenetic marks data and DNase hypersensitivity data came from B-lymphocytes in the peripheral blood. We chose the cell type to match the dataset from HapMap in order to minimize the effect of epigenetic variance due to differential gene expression in different cell types.

In order to create a sizable dataset, we defined cutoff thresholds for recombination rates to differentiate between genomic hotspots and coldspots. Nucleotide sequences from hotspot regions were assigned positive binary labels while sequences from coldspot regions were assigned with negative binary labels. Principal component analysis (PCA) was performed to see if any significant patterns existed within the datasets for raw sequences alone and raw sequences combined with exogenous features. For the following classification models, the training data were flattened into a one-dimensional vector $D \in \mathbb{R}^{lf}$. These models included logistic regression, naive Bayes classifier, random forest classifier, and a simple neural network. Each model was first trained on the raw sequences alone and then on raw sequences combine with exogenous variables. We chose to use logistic regression for the baseline model, which has a relatively simple model complexity and associated assumptions. We then used a hybrid convolutional and recurrent network model (hybrid-CRNN) on the unflattened dataset to improve the predictive accuracy.

Each model was then 10-fold cross-validated using the training dataset to assess their predictive performance and ability to generalize to other data set. All models were implemented using packages from Scikit-learn and the Keras API [33], [38].

*Logistic regression* is a generalized linear model designed for the classification of binary variables. The algorithm aims to find the optimal weights $\theta, \theta_0$ that minimizes the objective function that contains a negative loss likelihood $L_{nll}$ between predicted label and actual labels $y^{(i)}$.

$$J(\theta,\theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{nll}(\sigma(\theta^T x^{(i)} + \theta_0), y^{(i)}) + \frac{\lambda}{2}||\theta||^2 \quad (1)$$

We included a ridge regression (l2) term that reduced model complexity with regularization strength $\lambda$ to prevent overfitting. We adjusted weights inversely proportional to the frequencies of hotspots and coldspots in the dataset. The Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm was used for the solver. This algorithm does not require a linear relationship between the dependent and independent variables, and homoscedasticity of the error residuals. Additionally, regularization reduces overfitting issues with multicollinearity among the predictors. However, this algorithm tends not to perform well on datasets with large feature space.

*Naive Bayes Classifier* applies Bayes' theorem with the assumption of independence between every pair of features. We used Bernoulli naive Bayes since the dataset consisted of binary features. While this algorithm is computationally fast and works well with high dimensional data, its performance relies on the assumption of independence between the features. This did not hold in our case since nucleotide sequences and chromatin marks are correlated to a certain degree. However, we still opted to include this model since it often performs better in practice than in theory.

*Random Forest Classifier* creates a set of decision trees by, for each tree, drawing a bootstrap sample from the the dataset and then growing a tree on that sample using a set of randomly selected variables. Each decision tree creates rules for the given subset and predicts from observations by traversing from the branches to the leaves. Construction starts from the top and goes to the bottom by choosing a variable at each internal node that best splits the dataset by some criteria, e.g. information gain by Gini index. It then aggregates votes from the the ensemble of trees to make a prediction on a new data point. While this method reduces the overfitting issue seen in simple decision tree models, it is not as easy to visually interpret.

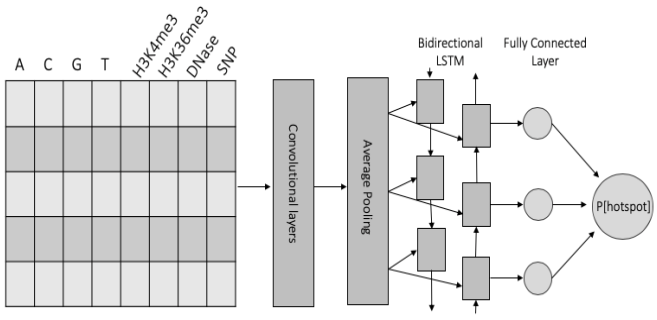


Fig. 1. Structure of our dataset and architecture of hybrid convolutional-recurrent neural network.

*Simple Neural Network* utilizes a single fully connected layer with sigmoidal activation, a nonlinear activation function that generates outputs bounded in the interval between 0-1. This model is trained by minimizing the objective function of negative loss likelihood $L_{nll}$ between the predicted label $p^{(i)}$ and the actual labels $y^{(i)}$ to find the optimal weights $\theta, \theta_0$ for the neural network (1). The input matrix was flattened following the procedure described above, and outputs represented the probability of a given sequence is a hotspot.

*Hybrid Convolutional and Recurrent Neural Network* follows the framework used by models such as FactorNet and DanQ [26], [27]. This type of hybrid model has the ability to capture motifs via convolution, while finding longer-range relationships between motifs via a long short term memory (LSTM) layer [27]. Our model takes as an input an $l$ by 8 matrix, where the first four columns represent the one-hot encoding of the raw DNA sequence, while the next columns represent data for H3K4me3, H3K36me3, SNP density, and DNase I Hypersensitivity (Fig. 1). For the sake of training, we set the length of the matrix $l$ to be 1000, since the length of hotspot regions is known to range between 1-2kb [28]. The matrix was then fed into two 1D convolutional layers, with 50 filters each using ReLU activation functions. Max pooling was utilized after each convolutional layer to extract the most significant contributors for the sequence [29]. The weights were then passed to a bidirectional LSTM layer. Global average pooling was used, before passing to a fully connected layer with a sigmoidal activation function. A dropout rate of 0.1 was used to prevent overfitting. The outputs of the model represented the probability

that a given sequence is a hotspot. The model was validated on a separate testing set of data to obtain accuracy metrics.

For motif discovery, we used the Gibbs sampling algorithm, which searches for frequently occurring motifs in a set of sequences. This algorithm outputs a position weight matrix (PWM) that represents the likelihood of each of the four nucleotides occurring at a given position in the motif. For the sake of runtime in our current iteration, we limit our search to 50 hotspot sequences.

The algorithm first generates an initial probability weight matrix (PWM) $P$ based on the length of motif $L$, the given list of sequences $S$, and a list of random guesses $G$ for the starting position the motif in each sequence. A uniform background distribution $B$ is assumed across all nucleotides since we have no prior knowledge about the nature of the sequences given. Next, we create a profile matrix $Z$ such that each entry $Z_{ij}$ corresponds to the probability that a motif instance starts at position $j$ in the sequence $i$. We loop through the list of sequences and exclude one sequence $S_i$ from the set of sequences $S$ at each iteration. The PWM is updated using the remaining set of sequences $S - S_i$ by calculating the frequency of observing nucleotide $n$ at each position while adding a pseudo-count $d$ to prevent calculation issues due to zero probabilities. One iteration is completed after each sequence is excluded once and each starting position is updated once.

Using the updated PWM, we update each entry $Z_{ij}$ in the profile matrix, which is the probability of the motif instance starting at position $j$ in the removed sequence $S_i$. The update is calculated by dividing the probability that the sequence was generated based on updated PWM by the probability that the sequence was generated based on the background noise (2).

$$Z_{ij} = \frac{\Pi_{k=j}^{j+L-1} P(c,k)}{\Pi_{k=j}^{j+L-1} B(c,k)} \tag{2}$$

The profile matrix $Z_i$ is then normalized into a proper probability distribution. The new starting position for a sequence $S_i$ is decided by choosing positions based weighted based on $Z_i$. Finally, the previous two steps are repeated until convergence. The convergence condition is fulfilled when the change in each entry of the PWM is less than some value $\varepsilon$ between iterations, which ensures convergence to any maximum. In our current model, we use an $\varepsilon$ of 0.001. We run the Gibbs Sampler 10,000 times to find 10,000 motifs, which are then compared to the known degenerate hotspot motif CCNCCNTNNCCNC, which is enriched in approximately 40% of hotspots [10]. An average percent identity is obtained and used to evaluate the success of our sampling, and the most frequently occurring motifs are reported.

Finally, we ran the sequence into our model and created a probability distribution of hotspot likelihood. Using a user-defined threshold, we selected new regions that are likely to be hotspots (and even coldspots). Using these sequences and the motif discovery method, we attempted to identify new motifs and validate existing ones. These sequences were also validated by exploring their sequence properties and association with GWAS SNPs.

## III. RESULTS

### *A. Test for significant correlations within the dataset using simple classification models and neural network with different architectures.*

We trained our models on chromosome 1, 3, 5, 7, and 9 and tested on chromosome 11. This is so that we can make comprehensive predictions on chromosome 11, which is associated with over 150 human diseases and has a unusually high number of genes [39]. The chromosomes used to train the model were chosen such that there was a 85:15 training-testing dataset split. PCA was performed on the training and testing dataset consisting of hotspots and coldspots regions with and without the exogenous variables (Fig. 2). The two-dimensional plots were generated based on the principal components that explained the most variance in their respective directions. The majority of the hotspot and coldspot regions overlapped with one another in the PCA of the dataset consisting of only the sequence, hinting at the possibility of a lack of separable patterns. However, the PCA of the dataset including the exogenous variables created a visual pattern that separated the hotspot and coldspot regions. This suggested that the data can be classified with high degrees of accuracy given the right choice of models.

10-fold cross-validation using the training dataset was performed on each model to assess its predictive performance (Fig. 2). The cross-validation accuracy confirmed the results of the PCA analysis since each model performed, on average, 3% better with the inclusion of the exogenous variables with the exception of Naive Bayes. This makes sense since the performance of naive Bayes depends on independence between its features. The inclusion of exogenous variables broke the independence assumption and, therefore, lead to

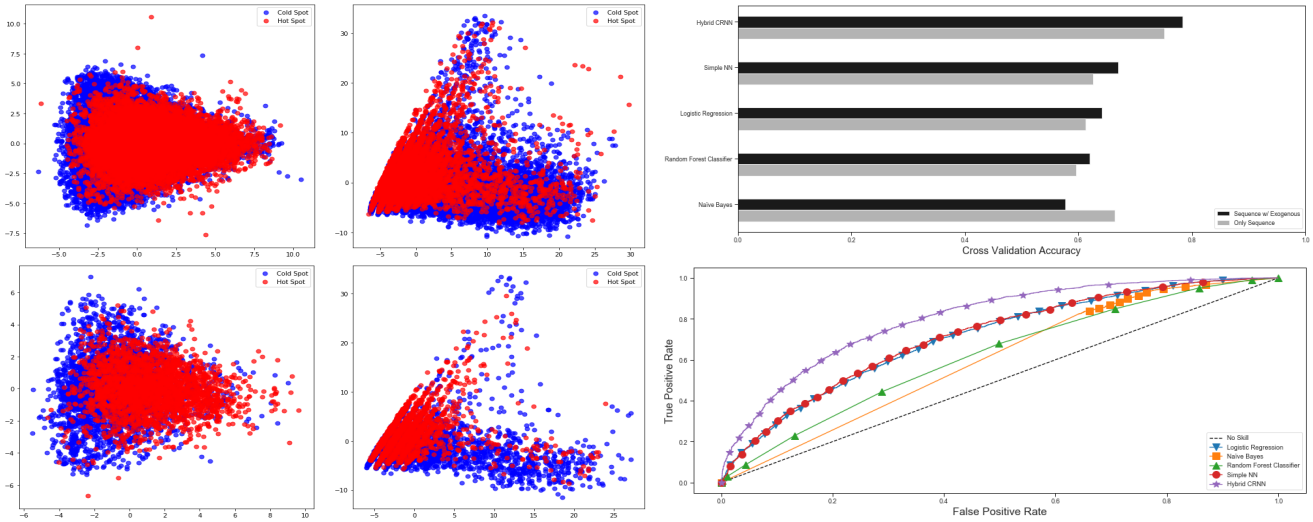


Fig. 2. PCA on the training set (top) and PCA on the testing set (bottom) without exogenous variables (left) and with exogenous variables (right). Cross validation accuracy (top right) for all models trained on dataset with only the sequence and sequence with exogenous variables. ROC (bottom right) of each tested model trained on sequence with exogenous variables.

a higher prediction accuracy on the dataset with only raw sequences. Logistic regression and random forest trees both performed reasonably well considering that they are relatively simple models. Likewise, simple neural network only performed slightly better despite the large increase in complexity of the model. Finally, the hybrid CRNN model performed the best amongst all of the models, achieving a cross validation accuracy of 78.35% on the dataset with raw sequence combined with exogenous variables.

The Receiver Operator Characteristic (ROC) for each model (Fig. 2) served as another method to measure the ability of these binary classification models. It demonstrated the trade-off between sensitivity and specificity, such that classifiers closer to the top-left corner of the graph are more ideal. This suggested that the hybrid CRNN had the highest predictive accuracy with an Area Under the Curve (AUC) of 0.802. Therefore, we proceeded with the hybrid CRNN for further analysis.

The improved level of accuracy can be further visualized in the training accuracy of the hybrid CRNN. A training/validation data split of 0.2 was used to obtain an estimate of training and validation accuracy at each epoch. When trained on raw sequence data alone, the model achieved 78.2% accuracy in training (Fig. 3, top). However, a significant tendency towards overfitting was observed as the number of training epochs increased.

Conversely, when trained on all 8 parameters (4 nucleotides, 2 chromatin marks, DNase and SNP data) the

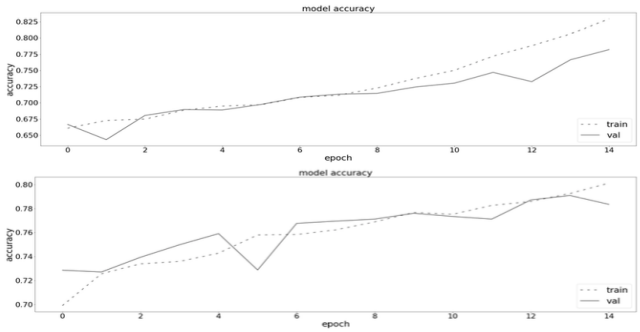


Fig. 3. Training and validation accuracy of the hybrid CNN-RNN.

model achieved 78.8% in training (Fig. 3, bottom). Notably, the validation accuracy increases proportionally with the training accuracy. This showed that contrary to the 4 parameter model, the model with 8 parameters does not overfit. These results confirmed our hypothesis that adding data related to chromatin marks, chromatin structure, and SNP density can boost the accuracy of machine learning models and allow us to make more powerful hotspot predictions.

### *B. Utilize motif discovery techniques to search for motifs enriched within hotspot regions.*

We ran the Gibbs Sampler to obtain 10,000 potential motifs. Table 1 documents the top ten most frequently occurring motif and their percentage identity to the known hotspot motif CCNCCNTNNCCNC. Percentage identity is calculated by dividing the number of matched nucleotides in a pair of sequences by the total

length of the sequence. The average percentage identity $\mu_2$ is 0.655 with a standard deviation $\sigma_2$ of 0.119. The expected baseline percent identity for a given 13-mer assuming a random distribution of nucleotide is calculated as follows, where $n$ represents the number of bases in the motif that are degenerate (denoted by $N$):

$$P = \frac{n}{13} + \frac{.25(13-n)}{13} \tag{3}$$

That is, the first half of the equation describes the free percentage identity due to the $n$ degenerate nucleotides and the second half assumes each non-degenerate nucleotide to match a quarter of the time. This formula evaluated to a baseline percent identity of 0.538. To confirm this calculation, we generated 10,000 13-mer motifs using a uniform background frequency for all nucleotides. The average percentage identity $\mu_1$ is 0.537 with a standard deviation $\sigma_1$ of 0.094. We used a two-sample t-test with unequal variances to further quantify the statistical significance of these results. The null hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis is $H_A : \mu_1 \neq mu_2$. We selected the significance level of 0.05. Let $n_1 = n_2 = 10000$ represent the sample size of each population. The degree of freedom is calculated as $min(n_1 - 1, n_2 - 1) = 9999$. The $t$ statistic is defined by the following equation:

$$t = \frac{(\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}} \tag{4}$$

This gives a $t$ value of 77.1, which at 9999 degrees of freedom in the student t-distribution yields a p-value of 0.00. This is less than the threshold of significance ($p < 0.05$), which shows that the percent identity of the motifs we found through Gibbs sampling are statistically significant.

Importantly, the 10 most frequently found motifs contain almost exclusively C's and G's. This aligns well with the current hypotheses that GC content either drives recombination or is a direct result of biased gene conversion due to recombination [35]. These results support the validity of using raw DNA sequence as a training feature, since they show clear bias depending on nucleotide frequency. Furthermore, they suggest that other hotspot motifs may exist besides the known degenerate motif. However, it is also worth noting that PRDM9 is a rapidly evolving protein, and there is evidence that the known hotspot motif may be undergoing self-driven removal in the human genome [30]. Thus, the biological validity of using such motif discovery techniques needs to be further evaluated.

TABLE I
10 MOST FREQUENTLY FOUND MOTIFS DISCOVERED BY GIBBS SAMPLING

| Motifs | Percent Identity | Frequency |
|---|---|---|
| CCCCCCCCCCCCC | 0.920 | 8 |
| GCCCCCCCCCCCC | 0.846 | 4 |
| CCCTCCCCCCCCC | 0.846 | 4 |
| CCCCCTCCCCCCC | 0.920 | 4 |
| CCCCCCCCCCGCC | 0.846 | 4 |
| CCCCTCCCCCCCC | 0.846 | 4 |
| CCCCGGCCCCCCC | 0.846 | 3 |
| CCCCCGGCCCCCC | 0.920 | 3 |
| CCCCCCCCCTCCC | 0.846 | 3 |
| CCCGCCCCCCCCC | 0.846 | 3 |

### *C. De novo hotspot and motif search in the human genome*

To fully visualize the success of our model, we used the CNN-RNN hybrid model to predict the location of hotspots on human chromosome 11 (Fig. 4). The locations of known hotspots are displayed as red spots, and the locations of known coldspots are displayed as blue spots. Our predicted probabilities are shown in black.

Several patterns can be observed from this graph. First, there exists a gap in the data points around coordination 0.55e8 that correlates to the centromere of chromosome 11. Due to the mechanisms of recombination machinery, we do not expect to see a high frequency of recombination near the centromere, so we would expect to see a high density of coldspots [34]. Our results align with this prediction, as we do see a decrease in hotspot density and an increase in coldspot density near the centromere. We also see that our model generally does a good job of identifying locations of hotspots and coldspots. However, there exist several instances where a hotspot was predicted where none exists. Moreover, coordinates 0-0.2e8 indicate many incorrect coldspots predictions. Further literature research is required to learn about the special structural properties of these genomic coordinates that may be causing our model to misclassify these data points.

We then filtered the predictions to only display the predictions with probabilities greater than 95% for and less than 5% (Fig. 4, bottom). New hotspots can be found in regions with prediction probabilities greater than 95%. For now, we labeled regions with a probability greater than 95% that are close to hotspot hubs in the chromosome as our newly discovered hotspots

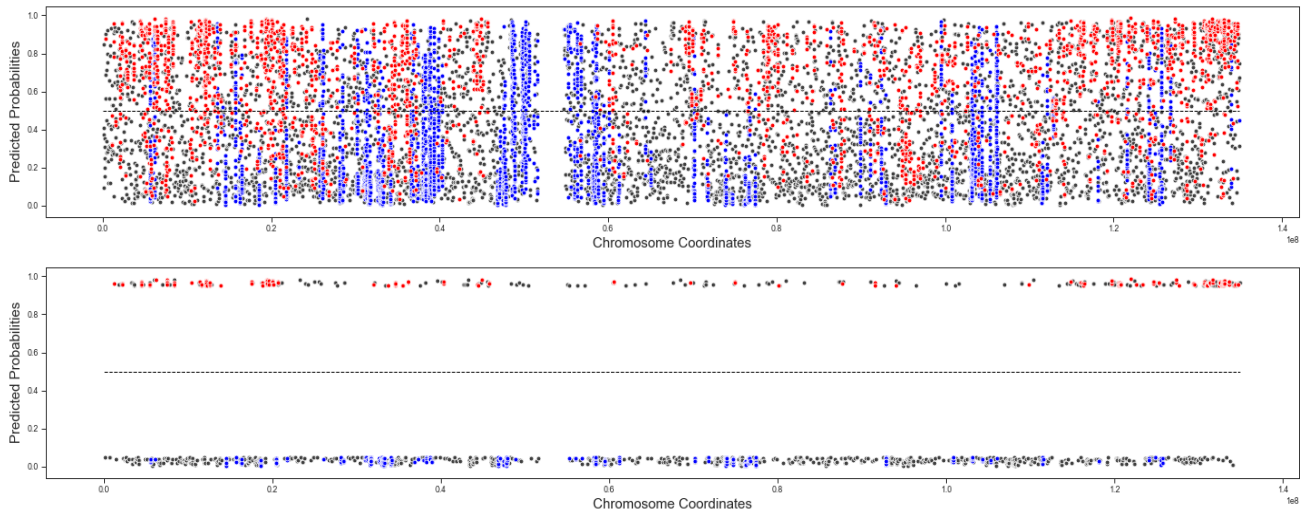


Fig. 4. CNN-RNN hybrid predictions along chromosome 11 with intervals of 1kb (top). Results are filtered for prediction probabilities of greater than 95% for hotspots and lower than 5% for coldspots (bottom).

(approximately 1519). To assess the significance of these results, we searched the GWAS database [36] to see how many of the discovered hotspots contain SNPs that are associated with diseases / traits. We found that 62 of these motifs (4%) contain GWAS SNPs associated with schizophrenia, cognitive abilities, and Alzheimer's disease. However, we later realized that this approach is faulty, since disease-associated SNPs tend not to be found in recombination hotspots [40]. We also performed a two-sample t-test between the hotspot sequence percentage identities and the background percentage identity. The mean $\mu_3 = 0.4825$ with a standard deviation of $\sigma_3 = 0.088$, leading to the conclusion that our results are significant. However, the motifs had an overall lower percentage identity and more $A$ and $T$ nucleotide compare to the motifs found using the actual sequences.

## IV. DISCUSSION

Recent attempts to predict hotspot distribution have included models using SVM, auto-cross covariance, and principal component analysis [14] and random forest classification [15]. These models trained on features such as codon composition [16] and gapped dinucleotide composition [15], and together constitute significant advances in the field of recombination prediction. However, to our knowledge, no models have been built for the mapping of recombination hotspots in the human genome. Furthermore, current models fail to take into consideration the contributions from higher-order epigenetic information.

We demonstrated here that there is a significant amount of information in these epigenetic data that can be used to identify more cohesive patterns in training features. For almost every model, we showed that model accuracy increases by a significant measure when we include the four exogenous features, comprising of H3K4me3 and H3K36me3, DNase hypersensitivity, and SNP density. We also found that, for our hybrid CRNN, the inclusion of these four features greatly reduced overfitting. This suggests that, at least in the context of human genomics, raw sequence information may not be enough for machine learning models to draw meaningful patterns from without becoming overly reliant on training data.

This finding raises the question of whether existing models for other species can be improved upon by the incorporation of more features. One potential concern with including epigenetic marks such as H3K4 is that the results may be conflated, due to the fact that H3K4 is enriched at many other significant sites in the genome, such as promoter regions. These varied enrichment patterns may be a source of noise in our model, and could potentially be causing an increase in the number of false-positive classifications. To mitigate this in future iterations, we would like to construct a negative dataset using coldspot regions that happen to be enriched with these histone marks. This could help our model better differentiate between histone marks that are correlated with hotspots and histone marks that are correlated with promoters and other genomic regions.

We also explored the efficacy of the hybrid CRNN model, which has been previously applied with notable success to other genomic contexts, in hotspot prediction. We found that, out of all the models tested, the hybrid CRNN performed with the greatest accuracy. This may be due to the fact that this model possesses the unique ability to find patterns both sequentially and spatially, via LSTM and convolutional layers, respectively. Thus, it proves to be particularly well fitted to genomic problems, in which a wide range of features can be assessed at once, while simultaneously considering longer-range patterns that may be present due to the sequential nature of DNA. Furthermore, our model identified hotspots that were previously unmapped. More work needs to be done to validate these results, from both computational and experimental standpoints, but this suggests that deep learning models may be able to find new hotspots in the human genome.

Another possible method of increasing our accuracy is reverse complement parameter sharing, a process in which convolutional filters are passed between forward and reverse reads of input sequences [26], [29] This method, originally proposed by Shrikumar et al. [31] is used as a way of assuring that both strands lead to the same predictions, and has been shown to improve accuracy and aid in faster learning. Finally, it is worth noting that our choice of making predictions on chromosome 11 may be leading to worse results, due to the unusually high density of genes on chromosome 11 [39], which may result in a higher than average incidence of hotspots. However, we have shown that the false positive rate decreases with the inclusion of these four features, and we postulate that the patterns drawn from the combination of the four features help to mitigate any bias that may be the result of any single feature. The possibility of a single feature tending towards bias may also explain why a decreased accuracy was observed for models trained on raw sequence data alone.

As a preliminary step towards even more robust hotspot prediction in the future, we sought to use Gibbs sampling to identify motifs that may be strongly associated with recombination hotspots. To measure the success of our sampling, we compared the motifs returned by the Gibbs sampler to a previously identified hotspot motif known to be enriched in 40% of hotspots. By calculating the percent identity of our motifs to the hotspot motif, we found that the ones returned by our Gibbs sampler corresponded more significantly to the hotspot motif than those generated by random chance. More work needs to be done to build a more refined sampling mechanism-however, our results suggest that motif discovery methods may be able to offer new insights and perhaps identify new hotspot-associated motifs.

To improve the performance of our motif discovery techniques, we may try to run the Gibbs sampler for more iterations over a greater array of hotspot sequences. We might also increase the stringency of the termination threshold of $\varepsilon$. Alternative approaches that may yield better results include general expectation maximization (EM), which is a deterministic method and considers the average of all entries of $Z_{ij}$ rather than taking a random sampling.

Additionally, we would like to attempt to extract the motifs from the convolutional layers in our deep learning model as implemented by Wang et al. [29]. To do this, the authors scanned the filters in the rectification layer following convolution and extracted the position that carries the most weight. This was repeated for each subsequence in the pool of sequences, and the results were aligned to obtain a PWM [29]. We would then compare the results obtained by this filter scanning method with the results generated by our Gibbs model.

Finally, we sought to place our results in the context of clinical significance by searching to see if disease-associated SNPs could be found within the hotspots we identified. We specifically chose to make predictions on human chromosome 11, due to the high incidence of disease-associated alleles present on this particular chromosome [39]. The low level of correlation (4%) reflects the fact that, contrary to our initial expectations, disease-associated SNPs actually congregate in areas of low recombination [40]. Thus, more extensive correlation studies still need to be done, and more revealing results may possibly be achieved with the LD database, since linkage disequilibrium has a direct correlation with recombination [37]. However, whether or not the identified hotspots can be used as a mechanism of finding disease-associated alleles, it remains that the ability to predict hotspots and understand in more depth the mechanisms behind recombination constitutes an important step forward in the fields of GWAS and genetic disease study.

## V. CONTRIBUTIONS

JF reviewed relevant literature regarding the optimal architecture of neural network models for this dataset, trained the hybrid CRNN, performed the motif discovery. LW consolidated the datasets, trained the simplistic

ML models, and performed the PCA, ROC, and t-test analyses. We contributed equally to the write up of the project. What worked in this collaboration was that JF had a biological background to interpret the results while LW had a computer science background to create results.

## VI. COMMENTARY

If we were to start over on the project, we would like to fully understand the dataset prior to creating models. Originally, our models were performing poorly. However, after reviewing the source of the dataset, we realized that there was a mismatch between the versions of human genome reference used by the HapMap dataset and the SNPs/Chromatin Marks. Also, we realized that each dataset came from different types of cells, making it much more incomparable. After correcting these mistakes, however, these models performed much better. We managed to balance the time between writing the paper and coding to obtain the results.

One of the most challenging aspects of the project was the complex nature of recombination in humans. There is still much that is not understood about the mechanisms of action of key proteins such as Spo11 and PRDM9, which raised questions early on into the process of whether or not the project would be successful. Extensive literature research was conducted to shed some light on the matter. However, the area of recombination study is still in its early stages. Finding and consolidating the necessary datasets also proved to be a challenge, and we were concerned that using a variety of histone marks and other higher-order genome information would lead to a high number of spurious hits since there does not exist any unique genomic markers for recombination. A lot of work was also done to make sure that all of the markers were mapped appropriately. Lastly, we also encountered difficulties in assessing the significance and validity of our results. Our model demonstrated the ability to find hotspots that have not been previously mapped, but we had no way of verifying that these were true hotspots because the experimental data does not exist. This was a major roadblock in the process of determining if our model was functioning correctly or if we were encountering spurious hits.

The peer-review process helped us narrow down our topic. Some of the feedback from other students were useful, but the majority of them were rehashing what was already stated in the proposal. They did identify potential pitfalls, such as pointing out that marks such as H3K4 are highly enriched at other sites within the genome-namely promoters. This was an extremely valid observation that we later brought up to our mentor, and it is something that needs to be addressed in more detail in the future. The concern of matching cell-types was also brought up, which, as previously mentioned, had a significant impact on the success of our model. Lastly, peer reviewers pointed out that motif discovery may not be a particularly applicable tool to this problem due to the highly mutable nature of PRDM9. We decided to proceed with the Gibbs Sampler anyway and achieved positive results. However, the question of whether or not it is possible to find a true hotspot motif that can be used for hotspot prediction is one that will need to be researched in greater depth.

Overall, this project was a greatly rewarding and enjoyable experience. We appreciated the opportunity to ask questions and receive feedback from experts in the field, and we also found the peer review process to be very helpful. The experience of pitching a project in the form of an NIH proposal to writing the final report was very valuable, and we hope to continue working on this project after the end of the semester.

## ACKNOWLEDGMENTS

We would like to thank Dr. Manolis Kellis, Dr. Wouter Meuleman, and Samuel Kim for their guidance on this project.

## REFERENCES

[1] Lam, Isabel, and Scott Keeney. "Mechanism and Regulation of Meiotic Recombination Initiation." *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. 1, 2014, doi:10.1101/cshperspect.a016634.

[2] Murakami, H., and A. Nicolas. "Locally, Meiotic Double-Strand Breaks Targeted by Gal4BD-Spo11 Occur at Discrete Sites with a Sequence Preference." *Molecular and Cellular Biology*, vol. 29, no. 13, 2009, pp. 3500–3516., doi:10.1128/mcb.00088-09.

[3] Myers, Simon, et al. "A Common Sequence Motif Associated with Recombination Hot Spots and Genome Instability in Humans." *Nature Genetics*, vol. 40, no. 9, 2008, pp. 1124–1129., doi:10.1038/ng.213.

[4] Myers, Simon R, and Steven A Mccarroll. "New Insights into the Biological Basis of Genomic Disorders." *Nature Genetics*, vol. 38, no. 12, 2006, pp. 1363–1364., doi:10.1038/ng1206-1363.

[5] Coop, G., et al. "High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans." *Science*, vol. 319, no. 5868, July 2008, pp. 1395–1398., doi:10.1126/science.1151851.

[6] Tock, Andrew J., and Ian R. Henderson. "Hotspots for Initiation of Meiotic Recombination." *Frontiers in Genetics*, vol. 9, May 2018, doi:10.3389/fgene.2018.00521.

[7] Noor, Mohamed, and Caitlin Smukowski. “Faculty of 1000 Evaluation for A Hierarchical Combination of Factors Shapes the Genome-Wide Topography of Yeast Meiotic Recombination Initiation.” *F1000 - Post-Publication Peer Review of the Biomedical Literature*, Jan. 2011, doi:10.3410/f.11685958.12773057.

[8] Borde, Valérie, et al. “Histone H3 Lysine 4 Trimethylation Marks Meiotic Recombination Initiation Sites.” *The EMBO Journal*, vol. 28, no. 2, Nov. 2008, pp. 99–111., doi:10.1038/emboj.2008.257.

[9] Verlhac, Marie-Hélène, and Marie-Emilie Terret. “Faculty of 1000 Evaluation for PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice.” *F1000 - Post-Publication Peer Review of the Biomedical Literature*, Feb. 2016, doi:10.3410/f.1867956.793515107.

[10] Myers, Simon, et al. “A Common Sequence Motif Associated with Recombination Hot Spots and Genome Instability in Humans.” *Nature Genetics*, vol. 40, no. 9, 2008, pp. 1124–1129., doi:10.1038/ng.213.

[11] Altshuler, D., et al. “Genetic Mapping in Human Disease.” *Science*, vol. 322, no. 5903, July 2008, pp. 881–888., doi:10.1126/science.1156409.

[12] Abeysinghe, Shaun S., et al. “Translocation and Gross Deletion Breakpoints in Human Inherited Disease and Cancer I: Nucleotide Composition and Recombination-Associated Motifs.” *Human Mutation*, vol. 22, no. 3, 2003, pp. 229–244., doi:10.1002/humu.10254.

[13] Pascoe, Leigh, and Kathleen M. Curnow. “Genetic Recombination as a Cause of Inherited Disorders of Aldosterone and Cortisol Biosynthesis and a Contributor to Genetic Variation in Blood Pressure.” Steroids, vol. 60, no. 1, 1995, pp. 22–27., doi:10.1016/0039-128x(94)00003-u.

[14] Liu, Bingquan, et al. “IRSpot-DACC: a Computational Predictor for Recombination Hot/Cold Spots Identification Based on Dinucleotide-Based Auto-Cross Covariance.” *Scientific Reports*, vol. 6, no. 1, 2016, doi:10.1038/srep33483.

[15] Jiang, P., et al. “RF-DYMHC: Detecting the Yeast Meiotic Recombination Hotspots and Coldspots by Random Forest Model Using Gapped Dinucleotide Composition Features.” *Nucleic Acids Research*, vol. 35, no. Web Server, Aug. 2007, doi:10.1093/nar/gkm217.

[16] Zhou, Tong, et al. “Support Vector Machine for Classification of Meiotic Recombination Hotspots and Coldspots in Saccharomyces Cerevisiae Based on Codon Composition.” *BMC Bioinformatics*, BioMed Central, 26 Apr. 2006

[17] Storey, Aaron J., et al. “Chromatin-Mediated Regulators of Meiotic Recombination Revealed by Proteomics of a Recombination Hotspot.” *Epigenetics Chromatin*, vol. 11, no. 1, 2018, doi:10.1186/s13072-018-0233-x.

[18] Yelina, Nataliya E., et al. “DNA Methylation Epigenetically Silences Crossover Hot Spots and Controls Chromosomal Domains of Meiotic Recombination InArabidopsis.” *Genes Development*, vol. 29, no. 20, 2015, pp. 2183–2202., doi:10.1101/gad.270876.115.

[19] “A Second Generation Human Haplotype Map of over 3.1 Million SNPs.” *Nature*, vol. 449, no. 7164, 2007, pp. 851–861., doi:10.1038/nature06258.

[20] Guo, Jing, et al. “LDSplitDB: a Database for Studies of Meiotic Recombination Hotspots in MHC Using Human Genomic Data.” *BMC Medical Genomics*, vol. 11, no. S2, 2018, doi:10.1186/s12920-018-0351-0.

[21] Liu, Bin, et al. “IRSpot-EL: Identify Recombination Spots with an Ensemble Learning Approach.” *Bioinformatics*, vol. 33, no. 1, 2016, pp. 35–41., doi:10.1093/bioinformatics/btw539.

[22] Chadwick, Lisa Helbling. “The NIH Roadmap Epigenomics Program Data Resource.” *Epigenomics*, vol. 4, no. 3, 2012, pp. 317–324., doi:10.2217/epi.12.18.

[23] “Encyclopedia of DNA Elements.” ENCODE, https://www.encodeproject.org/.

[24] Ernst, Jason, and Manolis Kellis. “Chromatin-State Discovery and Genome Annotation with ChromHMM.” *Nature Protocols*, vol. 12, no. 12, Sept. 2017, pp. 2478–2492., doi:10.1038/nprot.2017.124.

[25] Walker, Michael, et al. “Affnity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifcations on mammalian recombination hotspot usage.” *Epigenetics Chromatin* 2015, doi:10.1186/s13072-015-0024-6.

[26] Quang, Daniel, and Xie, Xiaohui. "FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data." *ScienceDirect* 2019, https://doi.org/10.1016/j.ymeth.2019.03.020

[27] Quang, Daniel, and Xie, Xiaohui. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic Acids Res* 2016, 44(11):e107. doi:10.1093/nar/gkw226

[28] Mackiewicz, Dorota et al. “Distribution of recombination hotspots in the human genome–a comparison of computer simulations with real data.” *PloS one* vol. 8,6 e65272. 11 Jun. 2013, doi:10.1371/journal.pone.0065272

[29] Wang, Meng et al. “DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional noncoding variants.” *Nucleic acids research* vol. 46,11 (2018): e69. doi:10.1093/nar/gky215

[30] Myers, Simon et al. “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination.” *Science* vol. 327,5967 (2010): 876-9. doi:10.1126/science.1182363

[31] Shrikumar, Avanti et al. "Reverse-complement parameter sharing improves deep learning models for genomics." *biorxiv* 2017, doi:https://doi.org/10.1101/103663

[32] Kitts A, Sherry S. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. 2002 Oct 9 [Updated 2011 Feb 2]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 5. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21088/

[33] Chollet, Franọis, et al. "chollet2015keras," 2015. Available at: https://keras.io/

[34] Gerton, J., et al. "Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomycescerevisiae*." *PNAS* 2000, 97 (21) 11383-11390; DOI: 10.1073/pnas.97.21.11383

[35] Marie-Claude Marsolier-Kergoat and Edouard Yeramian, "GC Content and Recombination: Reassessing the Causal Effects for the Saccharomyces cerevisiae Genome." *Genetics* September 1, 2009 vol. 183 no. 1 31-38, https://doi.org/10.1534/genetics.109.105049

[36] GWAS catalog. https://www.ebi.ac.uk/gwas/

[37] Bulik-Sullivan, et al. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nature Genetics*, 2015.

[38] Pedregosa, Fabian & Varoquaux, Gael & Gramfort, Alexandre

& Michel, Vincent & Thirion, Bertrand & Grisel, Olivier & Blondel, Mathieu & Prettenhofer, Peter & Weiss, Ron & Dubourg, Vincent & Vanderplas, Jake & Passos, Alexandre & Cournapeau, David & Brucher, Matthieu & Perrot, Matthieu & Duchesnay, Edouard & Louppe, Gilles. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12.

[39] Davis, Nicole. "Chromosome 11 rolls high number." *Broad Institute News*. (2006) https://www.broadinstitute.org/news/chromosome-11-rolls-high-number

[40] Elands, R., Simons, C., Riemenschneider, M. et al. A systematic SNP selection approach to identify mechanisms underlying disease aetiology: linking height to post-menopausal breast and colorectal cancer risk. *Sci Rep* 7, 41034 (2017) doi:10.1038/srep41034