A Deep Learning Approach to Prioritizing Genetic Variants in Coronary Artery Disease

Katherine Erdman
Department of Computer Science
Stanford University
kerdman@stanford.edu

Lucas Lin
Department of Computer Science
Stanford University
lucaslin@stanford.edu

Xiran Liu
Institute for Computational and Mathematical Engineering
Stanford University
xiranliu@stanford.edu

Sukolsak Sakshuwong
Department of Management Science and Engineering
Stanford University
sukolsak@stanford.edu

***Abstract*—Coronary artery disease (CAD) is a leading cause of mortality worldwide. Risk in coronary disease loci is largely determined by altered expression of its causal genes. A growing number of studies have been focusing on analyzing the genetic composition for disease progression, including the genome-wide association studies that report single-nucleotide polymorphisms (SNPs) associated to CAD. However, none of the existing methods attempt to infer the disease causal mechanism ab initio from the DNA sequence. In this project, we build a convolutional neural network model to predict chromatin profiling of DNA sequences, which enables the fine-mapping of disease-causal genetic variants for CAD. Our model was worse at predicting ATAC-seq profiles than the state-of-the-art Basenji and also less specific at predicting profile changes from SNPs identified by a genome-wide association study (GWAS). Combining SNP rankings from Basenji and our model may help isolate SNPs with high predicted differences due to hypothesized causation and not due to poor prediction on that region of the genome.**

I. Background and Motivation

Coronary artery disease (CAD) is the most common form of heart disease and a leading cause of mortality in many countries [1], [2]. CAD is caused by plaque buildup in the walls of the arteries, which blocks the blood flow over time and could lead to heart failure.

CAD has several environmental risk factors, such as high LDL-cholesterol, diabetes, and high blood pressure, but the underlying genetic composition may also substantially modify the disease risk, hence is critical for disease progression [2]. Risk in coronary disease loci is determined primarily by altered expression of the causal gene, due to variation in binding of transcription factors and chromatin-modifying proteins that directly regulate the transcriptional apparatus [3]. There are a growing number of studies in the genetic basis of coronary artery disease (CAD). Pjanic et al. have reviewed several genetic and genomics assays and approaches applied to coronary artery disease research [2].

Genome-wide association study (GWAS), with its growing databases, has been used to discover novel susceptibility loci for complex diseases through hypothesis-free case-control studies. Studies have provided large-scale association analysis for the coronary artery disease using data from multiple individual GWAS studies, identifying a number of loci that contain candidate causal genes [4], [5]. Several statistical methods have been applied to fine-mapping GWAS loci by calculating the posterior probabilities of causality for candidate variants [6]–[8]. Since the first genome-wide association studies of CAD reported in 2007, with increasing numbers of cases and controls, the power to discover loci associated with CAD continues to steadily rise [9]–[12]. However, when interpreting the results of GWAS, even true associations may not be causal due to linkage disequilibrium (LD). A GWAS association could represent either a causal variant or a non-causal variant that is in LD with the true causal variant, which leaves it a challenge to unravel the true causal mechanism behind the genetic composition of the disease [13].

Studies have used RNA-Sequencing to examine how genetic variation influences gene expression changes through expression quantitative trait loci (eQTL) and detect allele-specific expression (ASE) involving the differential expression between two alleles at heterozygous sites. Both eQTL and ASE have been used to prioritize functional variants among candidate GWAS variants and study underlying causal mechanism(s) of the disease/trait association [8], [14], [15]. The chromatin immunoprecipitation sequencing (ChIP-seq) and the assay of transposase accessible chromatin high-throughput sequencing (ATAC-seq) have been shown to be useful to detect local nucleosome occupancy and positioning, thus also valuable to investigate the genetic mechanisms of CAD loci [16]. For example, a recent study that performed genomic studies in human coronary artery smooth muscle cells, including chromatin immunoprecipitation sequencing, RNA sequencing, and protein-protein interaction studies, has shown that TCF21 and JUN regulate expression of two presumptive causal coronary disease genes, and the co-localization of AP-1 and TCF21 are enriched in coronary disease loci where they broadly modulate H3K27Ac and chromatin state changes linked to disease-related processes in vascular cells [3].

Meanwhile, there is a growing interest in building models for predicting phenotypic outcomes from genotypes using machine-learning methods. Prediction of cell type-specific epigenetic and transcriptional profiles from DNA sequence enables predictions for the influence of genomic variants on gene expression, which aligns well to causal variants underlying eQTLs in human populations and can be useful for generating mechanistic hypotheses to fine map disease loci. Previous studies have used both classical machine learning

models and deep learning models to predict a genetic variant's influence on gene expression and prioritize functional variants including expression quantitative trait loci (eQTLs) and disease-associated variants [17]–[19]. These methods have the advantages of performing predictions ab initio from the DNA sequence data and also avoiding some complicates brought by LD.

Therefore, we look into these deep learning ideas for prioritizing disease-causal variants for CAD. Given chromatin profiling datasets in coronary smooth muscle cells which have an important role in coronary artery disease, our objective for this project is to train neural networks to map DNA sequence to these profiles and use interpretation methods to score and fine map genetic variants associated with CAD.

II. Overview of existing work

A. Classic Machine Learning Approaches

Classic machine learning based methods have been developed for prediction of epigenetic and transcriptional profiles. Lee et al. have developed the gkm-SVM method based on gapped k-mer support vector machines for predicting the effect of regulatory variation. The model classifies putative regulatory sequences and matched negative-control sequences and gives each of the 10-mers an SVM weight that quantifies its contribution to the prediction of regulatory function, which is then used to score the predicted impact of any sequence variant on regulatory activity [20].

Following similar ideas, we build an SVM-based model to predict ATAC-seq profiles from DNA sequences and use it as the baseline for comparing the performances of our deep learning model.

B. Deep Learning Approaches

Zhou et al. developed a deep learning-based framework, ExPecto, which made ab initio prediction of variant effects on expression and disease risk using only the genome sequence [21]. It was based on the previous work, DeepSEA, by Zhou and Troyanskaya [17], with some notable improvements. The deep neural network was first trained to predict 2,002 different transcription factor, DNA accessibility, and histone mark profiles for 218 tissues and cell types. Then, the features were spatially transformed using exponential bases to generate a reduced set of features. Finally, tissue-specific regularized linear models used the transformed epigenetic information centered on the transcription start sites to predict expression of genes. The model achieved a 0.819 median Spearman correlation between predicted and observed gene expression levels across 218 tissues and cell types. It was also able to prioritize GWAS loci effectively; They found that loci containing SNPs with stronger predicted effects were significantly more likely to be replicated in another GWAS. This demonstrated that sequence-based deep learning frameworks could be an effective tool in fine-mapping GWAS loci.

C. State of the Art

Kelly et al. developed Basenji, a model similar to the deep learning-based approach of ExPecto with a few notable differences [19]. They trained the deep neural network to predict 4,229 chromatin profiles from over 1,000 tissues and cell types. The model first performed convolutional and pooling layers like ExPecto, but instead of using exponential bases to capture the effects of distal regulatory elements, it then used dilated convolutional layers, which increased the receptive field width exponentially [22]. To test the effectiveness of the receptive field width, they trained models with one to seven dilated convolutional layers and found that the test accuracy increased with each additional layer. (They did not use an 8th layer as it would reach outside the bounds of the sequence too often.) In addition, their data pre-processing pipeline also made use of multimapping read alignments, which have been shown to be important for gene regulation analysis. [23]

To evaluate Basenji's performance in eQTL prediction, Kelly et al. used the eQTL data on 19 tissues from the Gene-Tissue Expression project. For each SNP-gene pair, they calculated the "SNP expression difference" (SED) score as the absolute difference between the predicted CAGE coverage at that gene's transcription start sites for the two alleles. They also took LD in the eQTL data into account by adjusting SED according to variant correlations. They found that the adjusted SED significantly correlated with the observed eQTL ($p < 1 \times 10^{-54}$).

III. Methods

Our aim was to create a model for predicting Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) profiles in smooth cardiac cells based solely on DNA sequence.

A. Data

In total, there were 105,642 peaks within the ATAC-seq profile across chromosomes one to twenty-three, X and Y. Of these 105,642 peaks the minimum peak length in base pairs was 73 and the maximum was 2,499. Ultimately, input sequences were composed of 13k base pairs with the sequence corresponding to the peak in the ATAC-seq profile at the center of the 13k base pair sequence. The output was a prediction for the ATAC-seq profile corresponding to the center 3k base pairs. As noted in Figure 1, as all peaks were less than 2,532 base pairs, the center 2,532 base pair block in the 105,642 input sequences included both peak and non-peak data points. Like in past work, the inclusion of sequence outside of the 2,532 center region whose ATAC-seq profile was being predicted allowed for distal regulatory elements embedded within the surrounding sequence to be accounted for.

Notably, the model was trained twice. Once on just the reference sequence and another on the reference sequence, along with the corresponding complementary sequence. Details are shown in next section.

We have a list of SNPs with information about their associated to CAD from a GWAS study. There are 2,420,360 SNPs

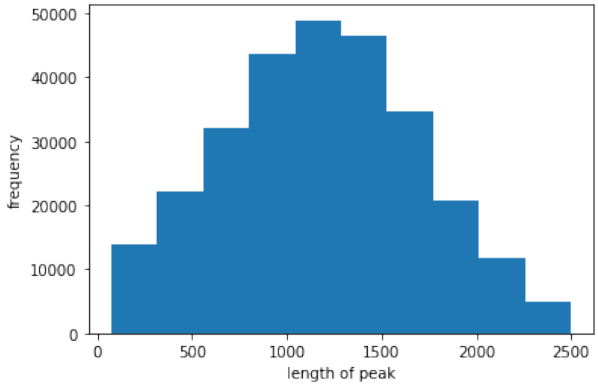

Fig. 1. Distribution of peak lengths

in the list. The information provided include the reference and alternative allele, the frequency of reference allele, number of cases and controls used in the study, the p-value for heterozygosity and log odds. The distribution of p-values are shown in Figure 2. This list of SNPs provides candidates of causal variants that we will use our prediction model to prioritize. By simple analysis, we found that 41,277 SNPs have p-value less than 0.01, 7,376 SNPs have p-value less than 10^{-3}, 509 SNPs have p-value less than 10^{-5} and 230 SNPs have p-value less than 5×10^{-7}.

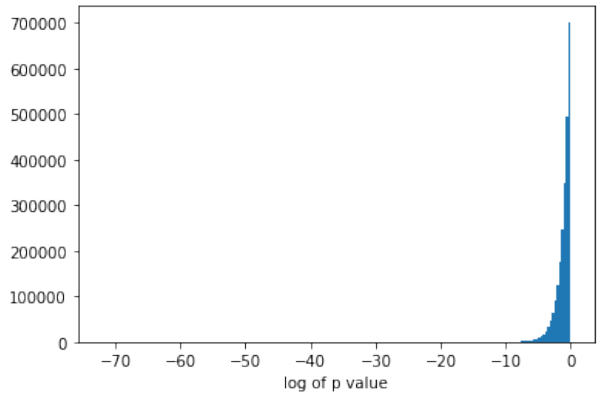

Fig. 2. Histogram of log(p value) in GWAS study of CAD-associated SNPs.

We also have a list of peaks in the ChIP-seq profile of TCF21 in coronary artery smooth muscle cells. TCF21 is a transcription factor encoded by the TCF21 gene that has been shown to be important for the development of vital organs, including heart [24]. A recent study has shown that TCF21 regulates expression of potentially causal coronary disease genes and is enriched in coronary disease loci [3].

B. Model

Our model made use of dilated convolution to reduce the 13k input to a 2,532 length output. Dilated convolutions allow for access to an exponentially increasing receptive field without also exponentially increasing the number of parameters. This is advantageous with only 105,642 examples. The dilated convolution between signal f and kernel k and dilation factor l is defined as [22]:

$$(k *_l f)_t = \Sigma_{\tau=-\infty}^{\infty} k_\tau \cdot f_{t-l\tau}$$

Our model (Figure 3) utilizes sequential convolutional layers to expand the input field in order to consider distal elements that affect chromatin accessibility. As the dilation rate increased, the number of filters decreased, while the kernel size increased. The number of filters started at 32 in the initial convectional layer and decreased by a factor of two every other time the dilation rate increased. The kernel size started at 24 when the dilation rate was 1 and dropped to 4 when the dilation rate was 5. From there, the kernel size increased by a factor of 2 every 6 layers. The model was trained using the adam optimizer to minimize mean squared error.

Our model differs from that of Basenji in a few key ways. One of which is the initial input window. Basenji determines outputs based on a 131k base pairs, while our model only looks at a 13k region. The first section of the Basenji model is a series of connected layers to generate a single representation for each 128 base pair bucket, as such predictions are made on the scale of 128 base pair regions. Our model predicts on the scale of base pairs.

C. SNP Analysis

Our model will be utilized to hypothesize the causality of single nucleotide metamorphism (SNPs) in regards to CAD. From the 2,420,360 SNPs with association data with CAD from a GWAS study, the top 10,000 with the lowest p-value, and thus the highest association with CAD, will be analysed to predict how they impact chromatin accessibility by modifying the reference genome with the alternative allele and centering the modified base pair at the center of the input sequence. 10,000 was chosen as it was computationally tractable while also containing all 7,376 SNPs that have p-value less than 10^{-3}. The max difference between the ATAC-seq profile predicted for the sequence containing the reference allele and that containing the alternative allele will serve as a measure to rank SNPs.

IV. ATAC-seq Prediction Results

Model	R^2 score	Pearson correlation	Spearman correlation
SVM	0.12	0.39	0.40%
Our Model	0.38,	0.70	0.64
Basenji	0.60	0.78	0.70

TABLE I
SUMMARY OF ATAC-SEQ PREDICTION RESULTS

A. Baseline from Classical Machine Learning Approach

Due to constraints on computational power and time, we implemented our baseline SVM model on only 20% of the samples available. Additionally, to simplify the prediction task, the top 15% of the data with the largest ATAC-seq profile

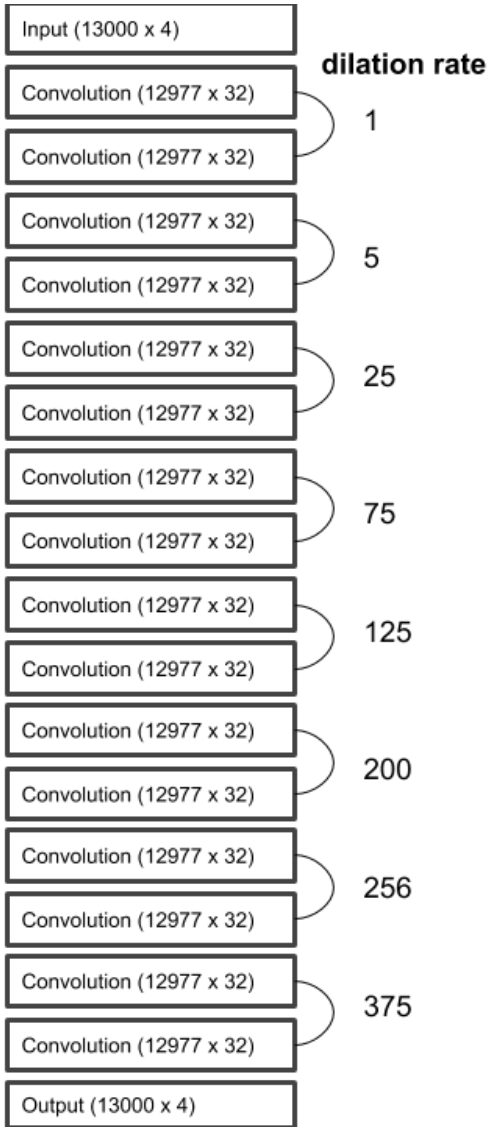

Fig. 3. Model Overview. Output sizes are shown in parentheses and associated dilation rates are shown to the right. Each convolution layer also includes batch normalization prior to ReLU activation.

values was removed to decrease the range of values. Each sample has features obtained from DNA sequence in a window size of 100-bp centered at the interested base pair location. The SVM model that uses this subset of samples yields a mean squared error (MSE) of 1.233, an R^2 score of 0.117, a Pearson correlation of 0.386 and a Spearman correlation of 0.394. A scatter plot of the prediction results is shown in Figure 4.

B. Basenji

To evaluate the performance of our model against state of the art, we trained the Basenji model on the ATAC-seq data. Instead of predicting 4,229 chromatin profiles like the original study, we modified the last layer of the model to predict only the ATAC-seq profile. Basenji was trained on all regions of the genome, not only peaks of chromatin accessibility. The model took as input a 131,072-bp region and output read count of

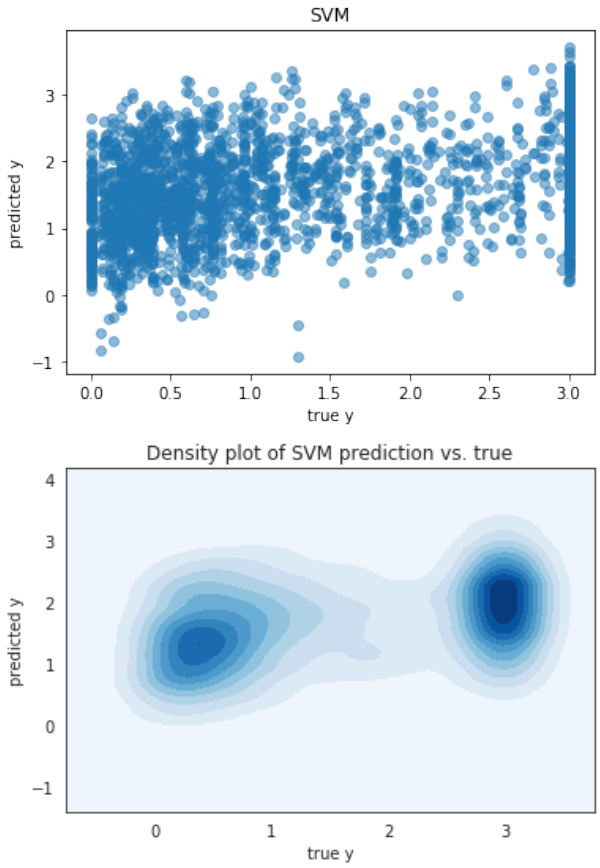

Fig. 4. Performance of SVM-based prediction model: a baseline. x axis is the ATAC-seq profile value; y axis is the predicted profiles.

960 bins, each bin representing a 128-bp region. Chromosomes 9 and 17 were withheld during training. Chromosome 9 was used for validation, and chromosome 17 was used for testing.

We first trained the Basenji model with random weight initialization but found the model to overfit quickly and achieved worse performance than the original study. We hypothesized that reducing the dimensions of the output (from 4,229 to 1 profile) while keeping almost the same number of parameters might contribute to the overfitting. We then initialized the model using the weights from the pre-trained model from the original study, except for the last fully connected layer, which had a different shape. The Basenji model achieved an R^2 score of 0.60, a Pearson correlation of 0.78 and a Spearman correlation of 0.70.

C. Dilated Convolutional Model

While training, all peaks originating from chromosomes 9 and 17 have been withheld. Chromosome 9 was used for validation, and reported results are from chromosome 17.

Our model was initially trained on just the reference sequence. This gives a mean squared error (MSE) of 3.52, an R^2 score of 0.38, a Pearson correlation of 0.70 and a Spearman correlation of 0.64. The predictions against true profiling are shown in Figure 5.

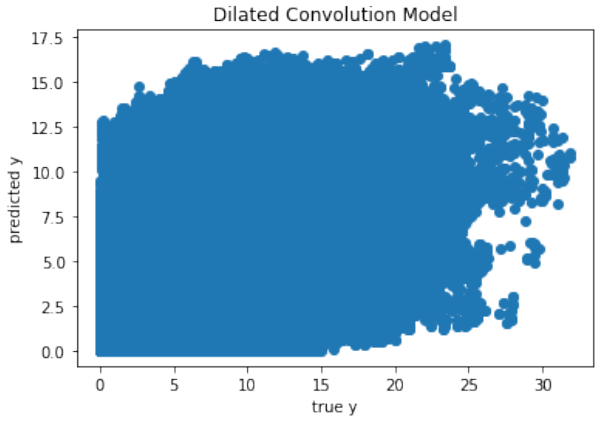

Fig. 5. Performance of our prediction model using only the reference genome. x axis is the true profiles; y axis is the predicted profiles.

Our model was separately trained on the reference sequence, along with the complementary sequence. This doubled the size of our training data and resulted in a mean squared error (MSE) of 6.46, an R^2 score of -7.13×10^{12}, an undefined Pearson correlation and an undefined Spearman correlation. The undefined correlation is caused by the fact that there is zero variance in the predictions. As shown in Figure 6, the model only learns the single predicted value that minimizes the overall MSE.

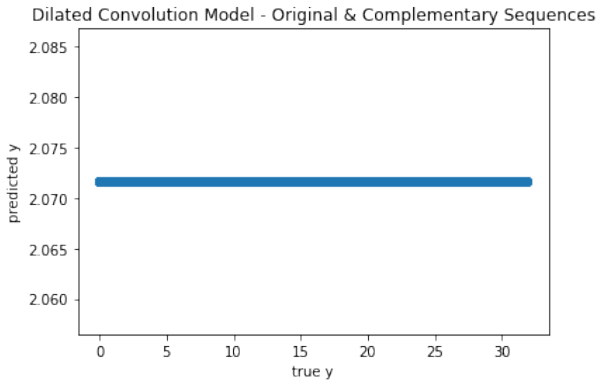

Fig. 6. Performance of our prediction model using both the reference genome and its complement. x axis is the true profiles; y axis is the predicted profiles.

V. SNP Analysis

SNPs were ranked by our model and Basenji based on the max difference in ATAC-seq profile predictions between the reference and the alternative allele. Example predictions for the five highest-rated SNPs are included in Figure 7.

An ideal model would be highly specific, with most SNPs causing very few changes in the predicted ATAC-seq profile, as most SNPs identified by GWAS are not causative of, but rather just associated with, CAD. As shown in Figure 8, Basenji was highly specific while the baseline SVM was not. This is expected as Basenji is able to more accurately predict ATAC-seq profiles and thus should be able to better predict

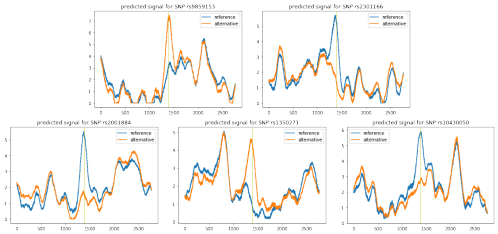

Fig. 7. Predictions for reference allele and alternative allele on Top 5 SNPs of our model. SNP rs9859153 is the top 1 predicted by both our model and Basenji.

the difference, or lack thereof, caused by a single base pair change. Our model behaves slightly worse than Basenji in terms of specificity, but much better than baseline model and shows the pattern of concentration towards zero difference as desired.

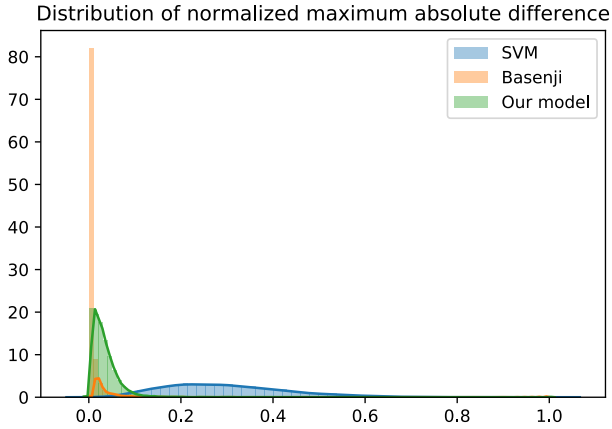

Fig. 8. Distribution of maximum absolute differences between predictions for reference and alternative alleles for top 10,000 SNPs associated with GWAS.

When comparing the ranking of SNPs between Basenji and our model, there were a significant number of SNPs that were ranked highly by both models (Table II). Specifically, when considering the set of SNPs in common between Basenji's and our model's top 20, 50 and 100 ranked SNPs, around 20% of those top SNPs were shared between both models. While it's encouraging to note that results were not completely desperate, it's difficult to determine which ranking is more accurate without additional, *in vivo* validation.

SNPs in common		Our Model				
		Top 20	Top 50	Top 100	Top 500	Top 1000
Basenji	Top 20	4	7	7	10	12
	Top 50	7	11	13	18	22
	Top 100	9	13	17	26	35
	Top 500	14	29	37	84	122
	Top 1000	15	33	49	125	190

TABLE II
NUMBER OF COMMON SNPS PREDICTED BY BASENJI AND OUR MODEL.

A. Validation Against Literature

Across literature, 32 SNPs have been found to be likely causal with CAD. rs2075650 (APOE) is a SNP found to cause

higher levels of high-sensitivity C-reactive protein [25]. High-sensitivity C-reactive protein contribute to chronic inflammation that contributes to coronary events like myocardial infarction [25]. rs7412 (APOE), rs1746048 (CXCL12), rs10757274 (9p21), rs17465637 (MIA3) and rs646776 (SORT1) are all SNP alleles found to be associated with CAD by the CARDIoGRAMplusC4D consortium [26]. These SNPs were found to decrease cytokine levels in serum, which caused a cytokine imbalance that is linked to an immune inflammatory pathogenic pathway of CAD [26]. In a biological assay with HeLa cells, disruptive missense variants in the low-density lipoprotein receptor gene caused higher plasma LDL-cholesterol, which is linked to myocardial infarction [27].

Of these 32 SNPs, only two of them had low enough p-values to be considered in our study of 10,000 SNPs. Of these two SNPs, rs1746048 (CXCL12) was ranked as the 19th top SNP in Basenji and as the 22nd ranked SNP in our model. rs646776 (SORT1) was ranked as the top 1544th SNP in Basenji and as the top 3107th SNP in our model. Overall, our model and Basenji aligned well with the high-ranking SNP and did not align as well for the other. This could be because only had a few SNPs had a large predicted change, so agreeing on ranking those large changes is an easier task. On the other hand, when considering a SNP that did not have a large predicted change, as was the case for the majority, small margins separated max differences for many SNPs, so getting similar rankings when the predicted change was smaller was more difficult.

B. Validation Against TCF21 Binding

Of the 10,000 GWAS SNPs with the lowest p-value, 345 SNPs are located within peaks of TCF21 binding in coronary artery smooth muscle cells. TCF21 is a transcription factor that has been implicated in CAD by multiple studies [3], [16]. To test whether the models were able to prioritize TCF21 SNPs, we plotted the cumulative distribution of these SNPs against the baseline where the SNPs were distributed uniformly (Figure 9). We found that Basenji was able to prioritize SNPs that are located within TCF21 peaks even without explicitly using TCF21 ChIP-seq profile as an input, while our model did not do as well as Basenji. We hypothesized that this might be because, by using a pre-trained model trained with many transcription factors, Basenji was able to identify certain motifs.

VI. Discussion

While many of the SNPs ranked highly by our model and Basenji cannot be validated by biological assays in other studies, certain SNPs occur on genes that could be causal to CAD. One such SNP is rs9859153, the top ranked SNP by both our model and Basenji. This SNP in inositol hexakisphosphate kinase 1, a gene that regulates the accumulation of fat stores in by the body by affecting AMPK-mediated adipocyte energy metabolism [28]. It was determined that inositol hexakisphosphate kinase 1 regulates energy metabolism in a way that

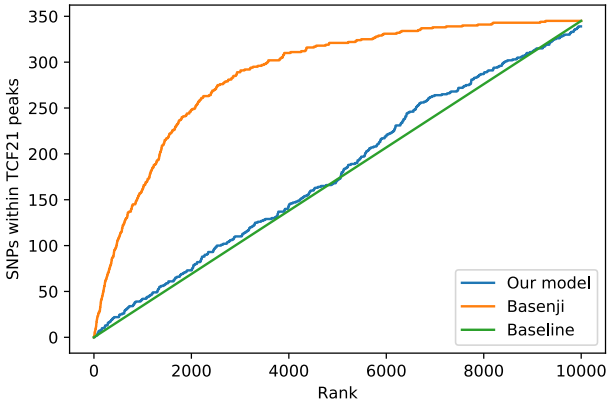

Fig. 9. Cumulative distribution of SNPs that are within TCF21 peaks. The baseline shows a uniform distribution of the SNPs.

implies it could affect obesity, of which CAD is a comorbid condition [28].

Another such SNP is rs6905958, which is only ranked in the top 25 by our model. rs6905958 is a variant within the SLC22A3, which encodes organic cation transporter 3 (OCT3) that helps to inactivate biogenic amines and remove toxic substances [29]. SLC22A3 has been shown to be involved in coronary vascular development and could explain how this SNP may be causal [29].

Though our model did not beat the state-of-the-art in terms of ATAC-seq profile prediction, even Basenji, the state-of-the-art model, did not always perfectly predict. As shown in Figure 10, certain SNPs had a higher MSE between the actual and predicted ATAC-seq profile and were ranked highly by Basenji. We hypothesized that the high ranking may not be because of a causal link between the SNP and CAD, but rather there is a large max difference in the prediction on the reference versus alternative allele sequence because Basenji predictions were not accurate in that region of the genome. As such, any changes may be as a result of poor modeling. To test this theory, the subset of SNPs ranked in the top 20, 50, 100, 500 and 1000 by both our model and Basenji were compared to just those ranked by Basenji. As show in Table III, for the top 20, 50, and 500 SNPs, the average MSE dropped by more than 12%. This seems to imply that cross referencing models identifies SNPs that are in regions of the genome that are better able to be predicted on. As such, any change due to swapping the alternative allele for the reference allele is more likely to occur *in vivo* and less likely to be due to poor predictive capability.

VII. Future Work

Due to computational limits, this study only considered the top 10,000 SNPs with the lowest p-values as determined by GWAS for our analysis. However, many of the causal SNPs identified in literature were not within this set. In the future, we would like to predict the potential effect of more than 2 million SNPs provided in the data set in order to determine if some SNPs that have low association with CAD may have

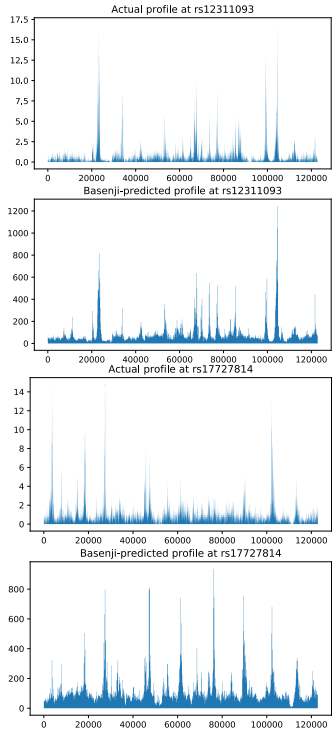

Fig. 10. Actual ATAC-seq profiles and predicted ATAC-seq profiles for two SNPs ranked within the top 20 by Basenji. SNP rs17727814 has a higher MSE between the actual and predicted values than rs12311093.

Number of SNPs	All SNPs	SNPs in common	Percent Δ
Top 20	16843.76	14713.52	-12.65%
Top 50	15589.74	12544.83	-19.53%
Top 100	14199.12	12440.82	-12.38%
Top 500	14080.50	14322.98	1.72%
Top 1000	14223.22	14595.79	2.61%

TABLE III
AVERAGE MEAN SQUARED ERROR BETWEEN THE ACTUAL AND BASENJI'S PREDICTED ATAC-SEQ PROFILE. PREDICTIONS WERE ACROSS A 71,680 BP SEQUENCE AND ON THE REFERENCE GENOME. ALL SNPS REFERS TO THE AVERAGE MSE ACROSS ALL TOP SNPS. SNPS IN COMMON REFERS TO THE AVERAGE MSE ACROSS ALL SNPS THAT WERE RANKED IN THE TOP N SNPS BY BOTH BASENJI AND OUR MODEL.

high causality. As CAD is such a widespread disease, it would make sense for there to be many genetic causes and perhaps there is a SNP that is causal for some small subsection of the population. This would explain why the SNP would not be ranked highly in GWAS, but could potentially still be causal.

Additionally, we would like to consider more complex models. Our initial model has 12 dilated convolutional layers and the updated model has 16. However, even the updated model still collapses to a single prediction when run on the original and complementary sequences. We hypothesize that this is due to underfitting and would like to consider a more sophisticated model with a slower rate of increase in dilation, allowing for more convolutional layers. As Basenji considers both the original and complementary sequences, it may be beneficial to generate a model that targets predicting on both sequences. That would not only double the size of the training data, but also provide some means of normalization as the weights would have to consider both the original and the complementary sequence that generate the same ATAC-seq profile.

ACKNOWLEDGMENT

We would like to thank Laksshman Sundaram and Anshul Kundaje for providing us the dataset and guiding us on this project.

CODE AND DATA

Our code and data are available on GitHub at https://github.com/xr-cc/DilatedCNNforCAD

REFERENCES

[1] E. J. Benjamin, P. Muntner, and M. S. Bittencourt, "Heart disease and stroke statistics-2019 update: a report from the american heart association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[2] M. Pjanic, C. L. Miller, R. Wirka, J. B. Kim, D. M. DiRenzo, and T. Quertermous, "Genetics and genomics of coronary artery disease," *Current cardiology reports*, vol. 18, no. 10, p. 102, 2016.

[3] Q. Zhao, R. Wirka, T. Nguyen, M. Nagao, P. Cheng, C. L. Miller, J. B. Kim, M. Pjanic, and T. Quertermous, "TCF21 and AP-1 interact through epigenetic modifications to regulate coronary artery disease gene expression," *Genome medicine*, vol. 11, no. 1, p. 23, 2019.

[4] P. Deloukas, S. Kanoni, C. Willenborg, M. Farrall, T. L. Assimes, J. R. Thompson, E. Ingelsson, D. Saleheen, J. Erdmann, B. A. Goldstein, *et al.*, "Large-scale association analysis identifies new risk loci for coronary artery disease," *Nature genetics*, vol. 45, no. 1, p. 25, 2013.

[5] M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, *et al.*, "A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease," *Nature genetics*, vol. 47, no. 10, p. 1121, 2015.

[6] S. A. Gagliano, M. R. Barnes, M. E. Weale, and J. Knight, "A bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization," *PLoS One*, vol. 9, no. 5, p. e98122, 2014.

[7] K. K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, *et al.*, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature*, vol. 518, no. 7539, p. 337, 2015.

[8] G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc, "Integrating functional data to prioritize causal variants in statistical fine-mapping studies," *PLoS genetics*, vol. 10, no. 10, p. e1004722, 2014.

[9] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Baker, A. Palsson, *et al.*, "A common variant on chromosome 9p21 affects the risk of myocardial infarction," *Science*, vol. 316, no. 5830, pp. 1491–1493, 2007.

[10] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, D. R. Cox, D. A. Hinds, L. A. Pennacchio, A. Tybjaerg-Hansen, A. R. Folsom, *et al.*, "A common allele on chromosome 9 associated with coronary heart disease," *Science*, vol. 316, no. 5830, pp. 1488–1491, 2007.

[11] P. Burton, D. Clayton, L. Cardon, *et al.*, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, p. 661, 2007.

[12] S. L. Clarke and T. L. Assimes, "Genome-wide association studies of coronary artery disease: recent progress and challenges ahead," *Current atherosclerosis reports*, vol. 20, no. 9, p. 47, 2018.

[13] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits," *Nature Reviews Genetics*, vol. 10, no. 4, p. 241, 2009.

[14] O. Mayba, H. N. Gilbert, J. Liu, P. M. Haverty, S. Jhunjhunwala, Z. Jiang, C. Watanabe, and Z. Zhang, "Mbased: allele-specific expression detection in cancer tissues and cell lines," *Genome biology*, vol. 15, no. 8, p. 405, 2014.

[15] B. Van De Geijn, G. McVicker, Y. Gilad, and J. K. Pritchard, "Wasp: allele-specific software for robust molecular quantitative trait locus discovery," *Nature methods*, vol. 12, no. 11, p. 1061, 2015.

[16] O. Sazonova, Y. Zhao, S. Nürnberg, C. Miller, M. Pjanic, V. G. Castano, J. B. Kim, E. L. Salfati, A. B. Kundaje, G. Bejerano, *et al.*, "Characterization of TCF21 downstream target regions identifies a transcriptional network linking multiple independent coronary artery disease loci," *PLoS genetics*, vol. 11, no. 5, p. e1005202, 2015.

[17] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, p. 931, 2015.

[18] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome research*, vol. 26, no. 7, pp. 990–999, 2016.

[19] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks," *Genome research*, vol. 28, no. 5, pp. 739–750, 2018.

[20] D. Lee, D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer, "A method to predict the impact of regulatory variants from dna sequence," *Nature genetics*, vol. 47, no. 8, p. 955, 2015.

[21] J. Zhou, C. Theesfeld, K. Yao, K. Chen, A. Wong, and O. Troyanskaya, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nature Genetics*, vol. 50, 08 2018.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv*, vol. 1511, no. 07122, 2015.

[23] C. Feschotte, "Transposable elements and the evolution of regulatory networks," *Nature reviews. Genetics*, vol. 9, no. 5, pp. 397–405, 2008.

[24] A. Acharya, S. T. Baek, G. Huang, B. Eskiocak, S. Goetsch, C. Y. Sung, S. Banfi, M. F. Sauer, G. S. Olsen, J. S. Duffield, E. N. Olson, and M. D. Tallquist, "The bHLH transcription factor Tcf21 is required for lineage-specific EMT of cardiac fibroblast progenitors," *Development*, vol. 139, no. 12, pp. 2139–2149, 2012.

[25] M. K. Christiansen, S. B. Larsen, M. Nyegaard, S. Neergaard-Petersen, R. Ajjan, M. Würtz, E. L. Grove, A.-M. Hvas, H. K. Jensen, and S. D. Kristensen, "Coronary artery disease-associated genetic variants and biomarkers of inflammation," *PloS one*, vol. 12, no. 7, 2017.

[26] W. Ansari, S. Humphries, A. Naveed, O. Khan, D. Khan, and E. Khattak, "Effect of Coronary Artery Disease risk SNPs on serum cytokine levels and cytokine imbalance in Premature Coronary Artery Disease," *Cytokine*, 2017.

[27] A. S. Thormaehlen, C. Schuberth, H.-H. Won, P. Blattmann, B. Joggerst-Thomalla, S. Theiss, R. Asselta, S. Duga, P. A. Merlini, D. Ardissino, E. S. Lander, S. Gabriel, D. J. Rader, G. M. Peloso, R. Pepperkok, S. Kathiresan, and H. Runz, "Systematic cell-based phenotyping of missense alleles empowers rare variant association studies: a case for LDLR and myocardial infarction," *PLoS genetics*, vol. 11, no. 2, 2015.

[28] Q. Zhu, S. Ghoshal, A. Rodrigues, S. Gao, A. Asterian, T. M. Kamenecka, J. C. Barrow, and A. Chakraborty, "Adipocyte-specific deletion of Ip6k1 reduces diet-induced obesity by enhancing AMPK-mediated thermogenesis," *The Journal of clinical investigation*, vol. 126, no. 11, 2016.

[29] Q. Zhao, H. Wei, D. Liu, B. Shi, L. Li, M. Yan, X. Zhang, F. Wang, and Y. Ouyang, "PHACTR1 and SLC22A3 gene polymorphisms are associated with reduced coronary artery disease risk in the male chinese han population," *Oncotarget*, vol. 8, no. 1, 2017.