Bioinformatics, 2019, 1–8
doi: 10.1093/bioinformatics/btz767
Advance Access Publication Date: 6 November 2019
Original Paper

Sequence analysis

Quantifying functional impact of non-coding variants with multi-task Bayesian neural network

Chencheng Xu [1,2], Qiao Liu[1,3], Jianyu Zhou[1,2], Minzhu Xie[4], Jianxing Feng[5,*] and Tao Jiang[1,2,6,*]

[1]Bioinformatics Division, BNRIST, [2]Department of Computer Science and Technology, [3]Department of Automation, Tsinghua University, Beijing 100084, China, [4]College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China, [5]Haohua Technology Co., Ltd, Shanghai 200041, China and [6]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 27, 2019; revised on September 29, 2019; editorial decision on October 2, 2019; accepted on November 4, 2019

Abstract

Motivation: Advances in high-throughput genotyping and sequencing technologies during recent years have revealed essential roles of non-coding regions in gene regulation. Genome-wide association studies (GWAS) suggested that a large proportion of risk variants are located in non-coding regions and remain unexplained by current expression quantitative trait loci catalogs. Interpreting the causal effects of these genetic modifications is crucial but difficult owing to our limited knowledge of how regulatory elements function. Although several computational methods have been designed to prioritize regulatory variants that substantially impact human phenotypes, few of them achieve consistently high performance even when large-scale multi-omic data are integrated.

Results: We propose a novel multi-task framework based on Bayesian deep neural networks, MtBNN, to quantify the deleterious impact of single nucleotide polymorphisms in non-coding genomic regions. With the high-efficiency provided by the multi-task Bayesian framework to integrate information from different sources, MtBNN is capable of extracting features from genomic sequences of large-scale chromatin-profiling data, such as chromatin accessibility and transcript factor binding affinities, and calculating the distribution of the probability that a non-coding variant disrupts regulatory activities. A series of comprehensive experiments show that MtBNN quantifies the functional impact of cis-regulatory variations with high accuracy, including expression quantitative trait locus, DNase I sensitivity quantitative trait locus and functional genetic variants located within ATAC-peaks that affect the accessibility of the corresponding peak and achieves significantly better performance than the existing methods. Moreover, MtBNN has applications in the discovery of potentially causal disease-associated single-nucleotide polymorphisms (SNPs), thus helping fine-map the GWAS SNPs.

Availability and implementation: Code can be downloaded from https://github.com/Zoesgithub/MtBNN.

Contact: fengjianxing@harmon.health or jiang@cs.ucr.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic mutations are closely related to human genetic diseases. In the past decades, genome-wide association studies (GWAS) have identified tens of thousands of disease- or trait-associated single-nucleotide polymorphisms (SNPs), of which more than 90% reside in non-coding regions (Hindorff *et al.*, 2009; Khurana *et al.*, 2016), revealing the vital role of non-coding variants in the occurrences of disease and the changes of phenotype (Zheng *et al.*, 2010). Understanding the functional impact of non-coding genomic variants is still challenging, partly due to our limited knowledge of how non-coding regions impact the regulation and expression of genes in different tissues and cell types. Meanwhile, the existence of linkage disequilibrium (LD) also makes it more complicated to identify causal SNPs from other benign variations within disease-associated intervals. Numerous efforts have been made to analyze the link between SNPs in non-coding regions and phenotype variations. Degner *et al.* (2012) used DNase I sequencing (DNase-seq) to detect chromatin accessibility disturbance caused by SNPs and found about 50% of expression quantitative trait loci (eQTL) SNPs are also DNase I sensitivity quantitative trait locus (dsQTLs), and 16% of dsQTLs are associated with variations in the expression of nearby genes. Moreover, a large proportion of disease-

associated variants in non-coding regions are proven to influence gene expression by disrupting the binding of important transcript factors (TFs) (Barrera *et al.*, 2016; Del Rosario *et al.*, 2015; Tehranchi *et al.*, 2016) such as CCCTC-binding factor (CTCF) (Bonder *et al.*, 2017). These results suggest that non-coding variants often affect the activities of regulatory elements by altering chromatin accessibilities or disturbing bindings of TFs.

Due to the importance of functional impact of non-coding genomic variants, several computational methods have been proposed to prioritize these variants by measuring their perturbation on the activities of regulatory elements (Lee *et al.*, 2015; Li *et al.*, 2016b; Quang and Xie, 2016; Ritchie *et al.*, 2014; Zhou and Troyanskaya, 2015). Among these work, deltaSVM (Lee *et al.*, 2015) trained a gapped k-mer supporting vector machine (gkm-SVM) (Ghandi *et al.*, 2014) to predict chromatin accessibilities and derived a score directly by summing the weight changes between alleles. In spite of its impressive performance in identifying dsQTLs, deltaSVM was found to be inadequate in the analysis of complex disease-associated SNPs (Li *et al.*, 2016b; Liu *et al.*, 2018). GWAVA (Ritchie *et al.*, 2014) utilized multi-omic data, including TF binding sites (TFBSs), phylogenetic conservation scores, genic context and histone modifications, to train a random forest classifier on the Human Gene Mutation Database (HGMD) to distinguish disease-associated variants from benign ones. Since the variants in HGMD are germ-line mutations, GWAVA is unable to make cell-type-specific predictions, which is essential when considering mutations in non-coding regions. DeepSEA (Zhou and Troyanskaya, 2015) and DanQ (Quang and Xie, 2016) are deep-learning-based frameworks, which integrated large-scale chromatin-profiling data of different cell types and outperformed the majority of the existing methods in predicting chromatin effects of genetic variants. However, similar to GWAVA, tissue-specific prioritization of SNPs cannot be achieved with the above two methods. Another SVM-based method, CAPE (Li *et al.*, 2016b) learned regulatory sequence signatures from a large amount of cell-type-specific regulatory signal tracks associated with enhancers and constructed an SVM classifier to identify cell-type-specific functional mutations.

Here, we introduce a multi-task Bayesian neural network framework, MtBNN, to quantify the tissue-specific functional impact of SNPs residing in non-coding genomic regions. Bayesian neural networks (BNNs) (Hernández-Lobato and Adams, 2015; Orre *et al.*, 2000) combine the strengths of stochastic models and deep neural networks (DNNs) that have achieved the state-of-the-art performance in a variety of classification tasks (Collobert and Weston, 2008; Glorot *et al.*, 2011b; Silver *et al.*, 2016). BNNs often overcome the over-confident estimation and over-fitting problems of DNNs (Gal and Ghahramani, 2015) by learning the posterior distributions of model parameters and estimating the probability distribution of predicted values. The probability distribution predicted by a BNN is more comprehensive than the point estimation given by a traditional DNN since it provides additional information such as uncertainty (Lacoste *et al.*, 2018; Lakshminarayanan *et al.*, 2017). More importantly, when informative priors are provided, BNNs can easily infer parameter posterior distributions even with a limited-size dataset, which is exactly the situation in prioritizing non-coding genetic variations since the amount of experimentally validated functional variants within non-coding regions of a specific cell type is usually limited (Li *et al.*, 2016b). In MtBNN, in order to integrate different sources of chromatin-profiling data, the identifications of different regulatory sequence signatures are defined as different tasks and a prior across multiple tasks are learned using a Bayesian neural network framework (Lacoste *et al.*, 2018). After training MtBNN on large-scale chromatin-profiling datasets including TFBSs, chromatin accessibility and histone modifications, the learned prior is transferred to quantify the impact of functional non-coding SNPs by fine-tuning the pre-trained model on labeled non-coding genomic variations. Our comprehensive experiments demonstrate that in the prediction of TF binding affinity, MtBNN is comparable to the state-of-the-art method (Liu *et al.*, 2018), and in the tasks of prioritizing functional regulatory SNPs including identifying eQTLs, dsQTLs and ATAC-QTLs (i.e. functional genetic variants located within ATAC-peaks that affect the accessibility of the peaks), MtBNN significantly outperforms the existing methods. Our case studies also show that MtBNN can help fine-map GWAS SNPs.

2 Materials and methods

2.1 Design of MtBNN

MtBNN is based on a multi-task Bayesian neural network framework (Lacoste *et al.*, 2018). Here, the prediction of each chromatin-profiling dataset is designated as a task. In other words, the identification of binding sites of different TFs, the prediction of chromatin accessibility and the identification of peak locations of different histone marks are regarded as different tasks. By using a hierarchical Bayesian model across N tasks and assuming each task has its own parameters W_i, the posterior of parameters can be calculated as:

$$p(\boldsymbol{W}, \alpha|\boldsymbol{D}) = \frac{p(\alpha)\prod_i p(D_i|W_i)p(W_i|\alpha)}{p(\boldsymbol{D})}, \tag{1}$$

where $p(\alpha)$ is the hyper-prior distribution, $\boldsymbol{D}$ is the entire data in all N tasks, $\boldsymbol{W}$ represents the set of W_i and D_i denotes the data used in the ith task. Since W_i is high-dimensional and simple distributions do not fit well the intricate correlations in $p(W_i|\alpha)$, a random variable z_i is introduced as $W_i = h_\alpha(z_i)$ so that the parameter W_i can be separated into a common part α, which is shared across different tasks, and a task-specific part z_i. Hence,

$$p(W_i|\alpha) = \int_{z_i} p(W_i, z_i|\alpha)dz_i = \int_{z_i} p(z_i|\alpha)p(W_i|z_i, \alpha)dz_i. \tag{2}$$

To solve the model with variational inference, distribution families $q(W_i|z_i, \alpha)$ and $q(z_i|\alpha)$ are used to fit the posterior distributions. The Kullback–Leibler (KL) divergence of N tasks can be expressed as

$$\mathrm{KL}[q(\boldsymbol{W}, \boldsymbol{z}, \alpha)||p(\boldsymbol{W}, \boldsymbol{z}, \alpha|\boldsymbol{D})] = \mathbb{E}_q \log \frac{q(\boldsymbol{W}, \boldsymbol{z}, \alpha)}{p(\boldsymbol{W}, \boldsymbol{z}, \alpha|\boldsymbol{D})} \tag{3}$$

Note that $\boldsymbol{z}$ represents the set of all z_i. Here, we make two assumptions: (i) the data D_i is conditionally independent from z_i and α given W_i; (ii) the variational distribution for approximating $p(W_i|z_i, \alpha)$ is fixed as $W_i = h_\alpha(z_i)$, i.e. only $q(z_i|\alpha)$ and $q(\alpha)$ are variational. With these two assumptions, the KL divergence can be further simplified as

$$\begin{aligned}
&\mathrm{KL}[q(\boldsymbol{W}, \boldsymbol{z}, \alpha)||p(\boldsymbol{W}, \boldsymbol{z}, \alpha|\boldsymbol{D})] \\
&= \mathbb{E}_q \sum_i \log \frac{q(W_i, z_i|\alpha)}{p(W_i, z_i|D_i, \alpha)} + \mathbb{E}_q \log \frac{q(\alpha)}{p(\alpha, \boldsymbol{D})} \\
&= \sum_i \mathbb{E}_{q_i} \log \frac{q(W_i, z_i|\alpha)}{p(W_i, z_i|D_i, \alpha)} + \mathbb{E}_{q(\alpha)} \log \frac{q(\alpha)}{p(\alpha|\boldsymbol{D})} \\
&= \sum_i \mathbb{E}_{q_i} \log \frac{q(W_i|z_i, \alpha)q(z_i|\alpha)}{p(D_i|W_i)p(W_i|z_i, \alpha)p(z_i|\alpha)} + \mathbb{E}_q \log p(\boldsymbol{D}|\alpha) \\
&\quad + \mathbb{E}_{q(\alpha)} \log \frac{q(\alpha)}{p(\alpha|\boldsymbol{D})} \\
&= \sum_i \mathbb{E}_{q_i} \log \frac{q(z_i|\alpha)}{p(D_i|W_i)p(z_i|\alpha)} + \mathbb{E}_{q(\alpha)} \log \frac{q(\alpha)}{p(\alpha|\boldsymbol{D})} + \mathbb{E}_q \log p(\boldsymbol{D}) \\
&= \sum_i \mathrm{ELBO}_i + \mathbb{E}_{q(\alpha)} \log \frac{q(\alpha)}{p(\alpha|\boldsymbol{D})} + \log p(\boldsymbol{D}),
\end{aligned} \tag{4}$$

where $q = q(\boldsymbol{W}, \boldsymbol{z}, \alpha)$ and $q_i = q(W_i, z_i|\alpha)$. This allows the KL divergence over all tasks to be factorized into the task-dependent part ELBO_i and a task-common part $\mathbb{E}_{q(\alpha)} \log \frac{q(\alpha)}{p(\alpha|\boldsymbol{D})}$. The evidence lower bound (ELBO) can be derived as

$$\mathrm{ELBO} = -\sum_i \mathrm{ELBO}_i - \mathbb{E}_{q(\alpha)} \log \frac{q(\alpha)}{p(\alpha|\boldsymbol{D})}. \tag{5}$$

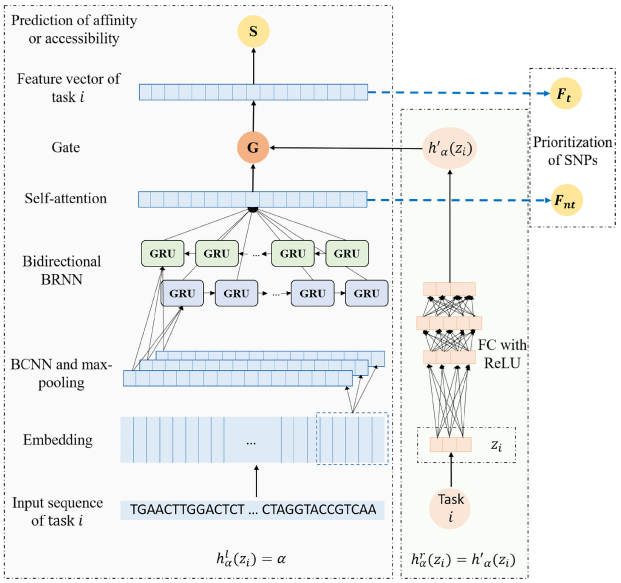

Fig. 1. The network architecture of MtBNN. The left network takes genomic sequences as the input and extracts features using the BCNN layers and BRNN layers. The output of the BRNN layer is integrated at a self-attention layer. In this part, the approximation function $h^l_\alpha(z_i) = \alpha$ is shared among all tasks. In the right network, the task-specific hidden variable z_i is transformed into a vector $h'_\alpha(z_i)$ by rectified linear units (Glorot *et al.*, 2011a) (or ReLU) activated FC layers (Supplementary Fig. S1), and the vector $h'_\alpha(z_i)$ is fed into a gate to control how much of each component should be passed on, thus filtering information output from the self-attention layer. The approximation function in the right network is set as $h^r_\alpha(z_i) = h'_\alpha(z_i)$. When the input sequence concerns SNPs, the logit, which measures how likely this SNP is deleterious, is estimated from F_t or F_{nt}, depending on what features are selected as the input of the multi-layer ReLU-activated FC network for fine-tuning

Therefore, our problem can be interpreted as a classical variational Bayes problem and there is no need to figure out the posterior distribution of $\boldsymbol{W}$. Many existing algorithms can be applied to optimize the posterior distribution of the remaining parameters including $\boldsymbol{z}$ and α, such as adaptive moment estimation (Adam) (Kingma and Ba, 2015), Nesterov-accelerated Adaptive Moment Estimation (Nadam) (Dozat, 2016) and AMSGrad (Reddi *et al.*, 2018).

The architecture of MtBNN is presented in Figure 1. The input genomic sequences related to chromatin-profiling tracks or non-coding SNPs are fed into a Bayesian convolutional neural network (BCNN) (Gal and Ghahramani, 2015) followed by a bidirectional Bayesian recurrent neural network (BRNN) (Fortunato *et al.*, 2017) layer (see the detailed parameters in Supplementary Table S1), the output of the BRNN is integrated at a self-attention module (Vaswani *et al.*, 2017). The architecture of the BRNN is a gate recurrent unit (Cho *et al.*, 2014). The task-specific feature vector is obtained after the gate, which is widely used in recurrent neural networks (Hochreiter and Schmidhuber, 1997), filters the information from the self-attention layer by $h'_\alpha(z_i)$. We finally obtain the predicted label, which represents the probability that the input sequence is a chromatin accessible region, a TFBS or a histone mark peak region, by reducing the dimensionality of the feature vector through some fully-connected (FC) layers. In this model, all network parameters of BCNNs, BRNN and FCs are denoted as the random variable α in the Equations 1–5, and $h_\alpha(z_i)$ is composed of a task-common part $h^l_\alpha(z_i) = \alpha$ in the left subnetwork, including all the layers before the output layer, and a task-specific part $h^r_\alpha(z_i) = h'_\alpha(z_i)$ in the right subnetwork.

The deleterious impact of the SNP contained in the input sequence can be estimated by transfer learning. After training the model on large amounts of chromatin-profiling data, the output of the self-attention layer or the task-specific feature vectors, depending on the property of the mutations under study, are fed to a new subnetwork (see Section 2.2). Finally, the distribution of the logit $L = \log \frac{p}{1-p}$, where p is the probability that the input SNP is deleterious, is estimated by the subnetwork.

2.2 Model fine-tuning

The pre-trained model was fine-tuned using known functional non-coding SNPs. In order to explore how feature selection would influence the final result, we attempted three different fine-tuning methods with prediction scores denoted as MtBNN_SINGLE, MtBNN_ALL and MtBNN_GENERIC, respectively. MtBNN_SINGLE scores were derived from the task-specific features, shown as the feature vector in Figure 1. The method is only applicated to dsQTL SNPs because they are defined based on DNase-seq signals. MtBNN_ALL was computed by combining feature vectors from all tasks (shown as F_t). MtBNN_GENERIC utilized the integrated information (shown as F_{nt}) from the self-attention layer to explore whether task-common information or task-specific features are more effective. All features were fed into a simple network of five ReLU-activated FC layers, which we call FNET here for abbreviation. Batch normalization (Ioffe and Szegedy, 2015) was used to constrain the distribution of the final result (see Supplementary Table S1 and Fig. S1).

In the fine-tuning process, features were sampled from the pre-trained model and fed into FNET. Then, the gradients were calculated and backpropagation was performed to update the parameters in FNET and the primary network. In the experiments to evaluate performance on regulatory variants, a 5-fold cross-validation was applied to each dataset. To avoid potential biases caused by sequences shared by the pre-training stage and the fine-tuning stage, when evaluating performance on regulatory variants, we excluded all peaks overlapping with the regulatory SNPs used in the cross-validation and pre-trained the model using the resulting 'clean' datasets. In the other experiments including the identification of TFBSs and prioritization causal or high-risk SNPs from candidate sets, the complete set of sequences was used. When utilizing MtBNN to infer causal SNPs from LD groups related to autoimmune diseases, locate high-risk SNPs in enhancer regions and prioritize GWAS SNPs, the model was first pre-trained on GM12878 chromatin profiles and then fine-tuned on dsQTLs in GM12878, as all these SNPs are related to B cells.

2.3 Sequence motifs recovery

We recovered motifs located in the input genomic sequences by following Basset (Kelley *et al.*, 2016). Briefly, each kernel of the first BCNN layer was converted into a position weight matrix (PWM) by scanning the input sequences, locating activated positions of the kernel and then calculating PWM from the involved genomic regions. TomTom (Bailey *et al.*, 2009) v5.0.4 was used with an false discovery rate (FDR) *P*-value threshold of 0.1 to match PWMs identified by our method to 719 known motifs from JASPAR database (Khan *et al.*, 2018).

2.4 Comparison to baseline models

Five existing methods, CAPE, GWAVA, CADD (Rentzsch *et al.*, 2018), DeltaSVM and Deopen, were chosen as the baseline models here. Both Deopen and DeltaSVM are able to identify regulatory signal tracks such as TFBSs or chromosome accessibilities, and prioritize SNPs. We trained Deopen with the same sequences used in our pre-trained MtBNN, but with one task each time since it does not support multi-task learning. DeltaSVM was trained on 300 bp sequences with the same center as the 1000 bp sequences that were used to train the other models. But due to its limited scalability, the training data concerning DNase-seq signals and histone mark peak regions were downsampled to 80 000 sequences for DeltaSVM. The methods, CAPE, GWAVA, CADD and DeltaSVM, were compared in the SNP analysis. DeltaSVM scores were derived following Lee *et al.* (2015) with the default parameters. CAPE was run with full-length sequencing data and the control sets were sampled by the code offered in its companion paper (Li *et al.*, 2016b). We tried several parameter settings for CAPE and picked the best one for further analysis. GWAVA and CADD scores were directly downloaded from their websites (GWAVA v1.0 and CADD v1.4 were used in our experiments).

2.5 Data preparation

The chromatin profiles, including DNase-seq data, chromatin immunoprecipitation sequencing (ChIP-seq) data of chromatin histone marks and some key TFs in GM12878 and HepG2, were downloaded from the website of Encyclopedia of DNA elements (ENCODE) project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/ and https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/. The files are listed in Supplementary Tables S2 and S3). The ATAC-seq data was obtained from Gate *et al.* (2018). The 1000 bp sequences centered at each signal peak were selected as the positive sequences, and the control sets were randomly sampled from the whole genome (GRCh37/hg19), excluding the peak regions, with the same length and number of sequences. 90% sequences in each cell-type-specific data were used to pre-train the model and the rest were used for validating the performance.

dsQTL and eQTL benchmark data for GM12878 and HepG2 are available in Li *et al.* (2016b). More specifically, 574 dsQTLs in GM12878, 1948 eQTLs in GM12878, 1044 eQTLs in HepG2 and 3-fold control sets with similar minor allele frequency distributions were collected. The positive ATAC-QTL data was obtained from Gate *et al.* (2018) after removing indel variants, and a 3-fold control set was randomly sampled from benign SNPs within ATAC-seq peak regions (see Supplementary Table S4). Sequences of lengths 1000 bp centered at each mutation site were extracted for the MtBNN fine-tuning and validation.

To verify the ability of MtBNN in identifying causal SNPs, several SNP groups with strong LD effects were collected. Candidate causal SNPs of myeloma (Li *et al.*, 2016a), pan-autoimmune (McGovern *et al.*, 2016) and chronic lymphocytic leukemia (CLL) (Kandaswamy *et al.*, 2016) were obtained from the variants listed in the corresponding literature (see Supplementary Tables S7–S9).

The GWAS catalog was downloaded from https://www.ebi.ac.uk/gwas/ (Buniello *et al.*, 2019). Following CAPE, SNPs related to B-cell traits were used to validate high-risk enhancer SNPs. To further explore the downstream applications of MtBNN, B-cell related diseases (see Supplementary Table S12) with more than 10 non-coding GWAS SNPs were selected for a finer analysis. The three SNPs with the highest scores in each disease assuming their scores are higher than the threshold (0 for MtBNN and 0.5 for CAPE), were considered as the candidate lead SNPs recognized by the corresponding method. These SNPs were further validated via a literature search (see Supplementary Tables S10 and S11). More specifically, the SNPs that were found to alter TF bindings or to be deleterious by methods other than GWAS are considered as having literature support.

3 Results

3.1 MtBNN exhibits high performance in identifying TFBSs

Before applying our method to the study of non-coding variants, the performance of the pre-trained MtBNN is evaluated on the identification of TFBSs. Since the test data is balanced, the area under the receiver operating characteristic curve (AUC) is used to measure the performance. MtBNN outperforms DeltaSVM and Deopen in 19 of the 28 TFBS datasets, with a mean AUC of 0.970 compared to 0.962 of DeltaSVM (*P*-value of 0.0037 in the Mann–Whitney *U*-test) and 0.961 of Deopen (*P*-value of 0.2086) (Fig. 2 and Supplementary Figs S2 and S3). Moreover, MtBNN achieves lower AUC variance (0.00056) than DeltaSVM (0.00076) and Deopen (0.00112) across different TF datasets.

We further visualize the kernels of the first BCNN layer (Fig. 2) to illustrate what the model learned from sequences. Using the motif comparison tool, TomTom with an FDR *P*-value threshold of 0.1, 42% of the calculated PWMs match the known motifs in the JASPAR database (Khan *et al.*, 2018) (see Section 2). Interestingly, partially due to the use of chromatin accessibility information, we also recovered motifs of some well-known important TFs not in the training data, such as BATF3, E2F1, EST1 and FOXP3 (Fig. 2). They have been proven to be related to genetic diseases (Kurts *et al.*, 2010; Engler *et al.*, 2014). As sequence perturbation on TFBSs has been found to be an important mechanism of diseases caused by non-coding mutations, the capability of MtBNN to identify TF motifs lays the foundation to quantify the functional impact of SNPs residing in non-coding genomic regions.

In order to evaluate the contribution of the Bayesian framework, a deep neural network (DNN) with the same architecture as MtBNN is trained and tested on the same datasets, and the classification results are summarized in Supplementary Figures S2 and S3. With a mean AUC score of 0.936, the performance of DNN is noticeably worse than MtBNN and other baseline methods, indicating the effectiveness of Bayesian neural networks. The performance of MtBNN and baseline models to predict chromatin accessibility and histone mark peak locations is also shown in Supplementary Figures S4–S7. MtBNN achieves lower AUC scores than DeltaSVM and Deopen but better or comparable performance as DNN except in the identification of histone marks (Supplementary Fig. S6). Considering the high performance of MtBNN in identifying TFBSs and recovering motifs, the deficiency in identifying chromatin accessibility and histone mark peak locations has little influence on its ability in quantifying SNP functional effects as shown in the results below.

To sum up, the pre-trained MtBNN performs slightly better than the state-of-the-art methods in identifying TFBSs and achieves lower but acceptable AUC scores in the chromatin accessibility and histone mark prediction tasks. The sequence features extracted by MtBNN provide a basis for quantifying the impact of SNPs on gene regulation.

3.2 MtBNN outperforms existing methods in prioritizing functional regulatory SNPs

Since our goal is to quantify the functional impact of non-coding SNPs, we fine-tune the pre-trained model on cell-type-specific functional regulatory SNPs to prioritize other SNPs. More specifically, dsQTLs and eQTLs within regulatory regions in GM12878, eQTLs

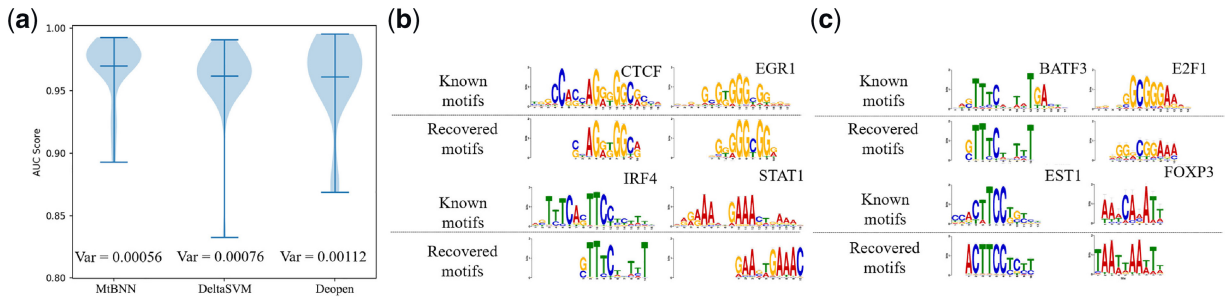

Fig. 2. The performance of the pre-trained MtBNN in identifying TFBSs. (**a**) MtBNN achieves slightly higher AUC scores than DeltaSVM and Deopen in 19 of the 28 TFBS datasets and is more robust across different TFBS datasets, with a mean AUC score of 0.97 compared to 0.961 of Deopen and 0.962 of DeltaSVM. (**b**) Some of the discovered TF binding motifs. The ChIP-seq signals of these TFs are included in the training data. (**c**) MtBNN also recovers some important TF binding motifs beyond those contained in the TFBS datasets, owing to the contribution of DNase-seq signals. These recovered motifs suggest that the pre-trained MtBNN has extracted useful information from genomic sequences

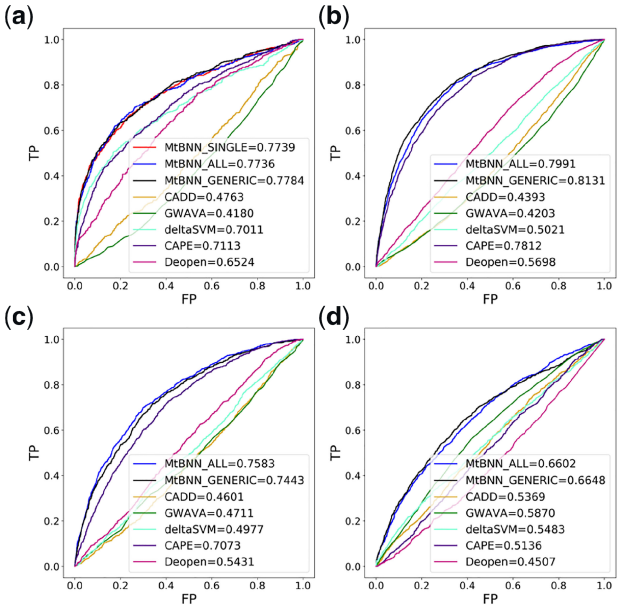

Fig. 3. Performance comparison of MtBNN and existing methods in prioritizing cis-regulatory SNPs. (**a**) dsQTLs in GM12878, (**b**) eQTLs in GM12878, (**c**) eQTLs in HepG2 and (**d**) ATAC-QTLs in T cell

in HepG2 and ATAC-QTLs in T-cell were used for 5-fold cross-validation of MtBNN on each of these datasets. In order to explore how the selection of the features would affect fine-tuning effectiveness, three different types of MtBNN scores, MtBNN_ALL, MtBNN_SINGLE and MtBNN_GENERIC are evaluated (see Section 2). For the ATAC-QTL dataset, the MtBNN_ALL score is exactly the same as the MtBNN_SINGLE score because the ATAC-QTL model contains only one task.

On the dsQTL dataset, MtBNN achieves an AUC score of around 0.77 when all task information, only chromatin accessibility information or the integrated information is considered (Fig. 3a). The best baseline models on this dataset are CAPE and deltaSVM, which achieve AUC scores of 0.71 and 0.70, respectively. MtBNN significantly outperforms baseline methods even though CAPE uses longer sequences and gene expression profiles. CADD and GWAVA, mainly designed for whole-genome variants, show almost no ability to distinguish deleterious SNPs from the control set. Deopen, which is the state-of-the-art model in predicting chromatin accessibility, achieves an AUC score of 0.65. Due to the fact that the sequences in the control set are three times of those in the positive set, the area under precision-recall curve (AUPRC) is also calculated to measure the performance (Supplementary Fig. S8). MtBNN outperforms other methods with an advantage of around 0.1 in AUPRC scores.

MtBNN also achieves significantly higher performance in identifying eQTLs and ATAC-QTLs in different cell types (Fig. 3 and Supplementary Fig. S8). When the MtBNN_GENERIC score is used, MtBNN obtains an AUC score of 0.81 in identifying non-coding eQTLs in GM12878 and 0.74 in identifying eQTLs in HepG2. CAPE, the state-of-the-art model, achieves AUC scores of 0.78 and 0.71 in identifying the above two types of eQTLs, respectively. The AUC scores of the other baseline models are close to random guessing. MtBNN also exhibits superiority when the AUPRC score is used as the performance measure (Supplementary Fig. S8), with an improvement of around 0.1 in AUPRC scores. Furthermore, all the four baseline methods fail to distinguish ATAC-QTLs from the background, while our approach achieves an AUC score of 0.66, further demonstrating the broad applicability of MtBNN in dealing with different types of SNPs. Based on the performance in different MtBNN scores on these SNP datasets, we select MtBNN_ALL for further applications and analysis.

To further test, the possibility that the functional regulatory SNPs from the active genomic regions shared between the training and testing stages might have caused any bias in the above comparison, we also perform a chromosome hold-out experiment similar to the partition of data in (Zhang *et al.*, 2018). SNPs on chromosomes 1–17 were used to pre-train and fine-tune the model, and the SNPs on chromosomes 18–22 were prioritized in the testing stage. The results shown in Supplementary Fig. S10 demonstrate that MtBNN still outperforms all baseline methods similarly as above and achieves significantly higher scores than random guessing.

3.3 MtBNN pinpoints causal SNPs from LD groups

To demonstrate how to locate deleterious SNPs in non-coding genomic regions, three sets of experimentally validated disease-related SNPs are selected to be prioritized by MtBNN and CAPE (Fig. 4 and Supplementary Fig. S9), including 10 myeloma risk variants at 7p15.3 that alter IRF4 binding (Li *et al.*, 2016a), seven pan-autoimmune genetic susceptibility variants at 6p23 that influence the binding of eight TFs (McGovern *et al.*, 2016) and 27 CLL risk variants at 15q15.1 that disrupt the binding of RELA (Kandaswamy *et al.*, 2016) (see Section 2). Since the variants in each of these groups have strong LD with each other ($r^2 > 0.8$), it is almost impossible to determine causal SNPs by association studies.

After being trained on the chromatin-profiling data of GM12878 and the dsQTL dataset mentioned before, MtBNN and CAPE are used to prioritize these SNPs and the highest scored SNP in each set is assumed to be the causal one found by the corresponding method (Fig. 4 and Supplementary Fig. S9). Both CAPE and MtBNN successfully identify causal SNPs in myeloma risk variants (rs4487645) and CLL risk variants (rs539846). Moreover, the causal SNPs are assigned significantly higher scores compared to other variants having LD with the GWAS traits. Nevertheless, CAPE fails to correctly prioritize the causal SNP, rs6927172, among the seven pan-autoimmune genetic susceptibility candidate SNPs, while MtBNN is able to pick the right one with high confidence. The SNP (rs11757201) assigned the highest score by CAPE has never been reported to have a causal effect on diseases based on a literature search, although it has strong LD with the GWAS traits (LD $r^2 = 1$) and alters the motifs of Mrg and Sp4 (McGovern *et al.*, 2016). In addition, MtBNN assigns a much lower score to rs11757201 than to rs6927172 in spite of the motif disruption caused by the former, which indicates that MtBNN is able to look beyond simple motif analysis.

Overall, our results suggest that MtBNN is capable of recognizing the causal non-coding SNPs in high precision and reliability from candidate SNPs from the same LD groups.

3.4 MtBNN contributes to the fine-mapping of GWAS SNPs

In practice, analyzing GWAS trait-associated variants is more challenging than locating a single causal SNP in an LD group since a trait is possibly associated with multiple functional variants. Furthermore, not only deleteriousness but also the impact degree of SNPs matter in GWAS. Here, we apply MtBNN to prioritize non-coding SNPs associated with B-cell related GWAS traits to verify whether our approach is capable of fine-mapping the existing GWAS SNPs.

By following CAPE (Li *et al.*, 2016b), we first apply MtBNN to prioritize SNPs in GM12878 enhancer regions (see Section 2) in order to evaluate its ability to identify unlabeled deleterious non-coding SNPs. We trained MtBNN and CAPE with the chromatin profilings of GM12878 and the dsQTL data. About 56 497 enhancer SNPs are prioritized by MtBNN or CAPE and the top 5% scored SNPs, denoted as high-risk SNPs here, are compared to B-cell-related traits in the GWAS catalog. We observe that 4.6% of the high-risk SNPs found by MtBNN are in GWAS catalog. As a comparison, only 3.0% of the top 5% functional SNPs found by CAPE are in the GWAS catalog (Li *et al.*, 2016b). This result is consistent with the previous conclusion that our approach is capable of pinpointing deleterious regulatory variants with a higher precision than the baseline methods.

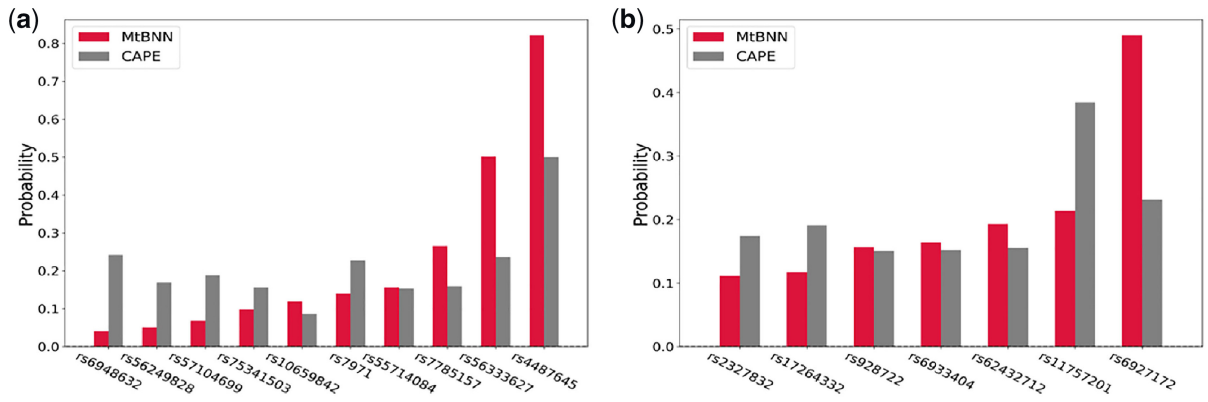

Fig. 4. MtBNN successfully pinpoints causal disease-related SNPs from LD groups and exhibits high precision and stability across different datasets. (**a**) Both MtBNN and CAPE distinguish the SNP altering IRF4 binding from ten myeloma risk variants. The causal variant, rs4487645, receives a notably higher MtBNN score and CAPE score compared with the rest. (**b**) MtBNN successfully identifies the causal SNP, rs6927172, from the candidates correlated with pan-autoimmune, while CAPE fails. The previous work in literature showed that rs6927172 alters the binding motifs of eight TFs, including NFKB and BCL3, while the variant identified by CAPE, rs11757201, only has LD with the GWAS SNPs

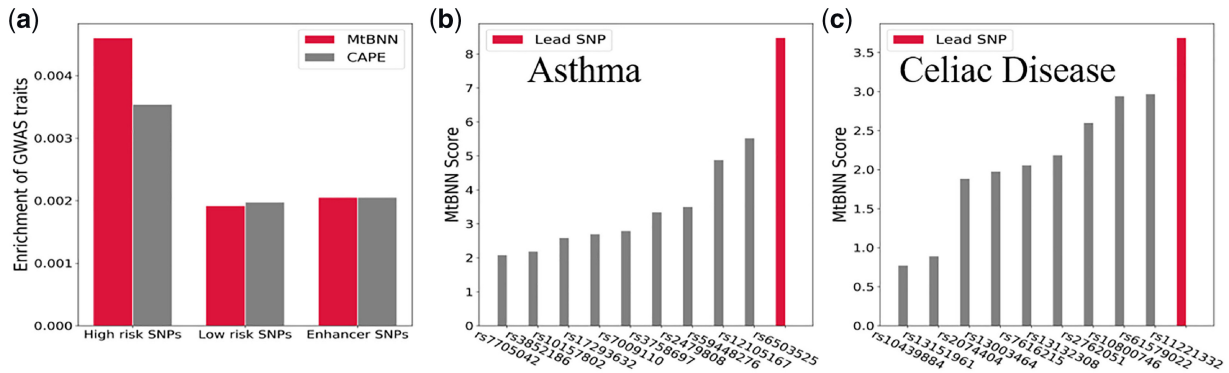

Fig. 5. MtBNN helps to fine-map known GWAS SNPs residing in non-coding regions. (**a**) In high-risk SNPs identified by MtBNN, the GWAS variants are more highly enriched compared to the background. Moreover, with the same threshold, MtBNN exhibits more enrichment than CAPE. (**b** and **c**) MtBNN pinpoints potentially functional non-coding SNPs associated with the GWAS traits of two B-cell related genetic diseases, asthma (rs6503525) and celiac disease (rs11221332). This suggests a potential method to speed up the inference of causal SNPs with the help of MtBNN

We then search for evidence from the literature for the prioritized non-coding SNPs associated with the GWAS traits of B-cell-related diseases (Supplementary Table S12). For each disease, we select top three SNPs with positive MtBNN scores and validate their functional impact via literature (Supplementary Table S10). It turns out that 6 out of 37 selected SNPs have literature support. More specifically, 15 of the 22 diseases have positively scored SNPs and 9 of the 15 candidate SNP sets have at least one SNP suspected of being deleterious in previous studies. We apply a similar process to CAPE. With a threshold of 0.5, CAPE selects 13 candidate SNP sets but only two of these sets contain SNPs supported by the literature.

We further study the top 10 SNPs in 2 of the 15 diseases, asthma and celiac disease, presented in Figure 5. The lead non-coding SNPs for asthma recognized by MtBNN, rs6503525, have been previously reported (Ferreira *et al.*, 2011; Ghani *et al.*, 2017; Shahid *et al.*, 2015) to have strong and reproducible effects on clinically diagnosed asthma. It shows strong association with asthma in multiple populations (Ghandi *et al.*, 2014; Shahid *et al.*, 2015) and is determined as the most-associated variant in the Australian population (Ferreira *et al.*, 2011), with functional impact on the locus MED24 (Pouladi *et al.*, 2016). rs11221332 is determined as the riskiest SNP for the other autoimmune disorder disease, celiac disease, by our method (Fig. 5). This variant is recognized in many previous studies to play a key role in the susceptibility of rheumatoid arthritis (Chatzikyriakidou *et al.*, 2013), idiopathic inflammatory myopathy (Svitalkova *et al.*, 2013), celiac disease (Dubois *et al.*, 2010) and other autoimmune disorder diseases (Ellinghaus *et al.*, 2016; Garrett-Sinha, 2013; Garrett-Sinha *et al.*, 2016) by altering the binding site of TF ETS1. It has also been proven to be the most strongly celiac disease-associated SNP in the European population (Dubois et al., 2010). These results provide further evidence about MtBNN's ability in finding potentially functional SNPs.

In conclusion, MtBNN still performs better in selecting lead SNPs than the existing methods in situations when such SNPs are more challenging to identify. Since the majority of SNPs recognized by GWAS lie in non-coding genomic regions, MtBNN will be quite useful in fine-mapping GWAS trait-associated SNPs and speeding up experimental studies.

4 Discussion

Interpreting non-coding sequence variation has been one of the greatest challenges owing to the limited comprehension of non-coding genetic regions. Utilizing large amounts of non-coding sequencing data, several machine learning methods have been proposed to prioritize these variants. However, the performance is far from satisfactory. Here, we developed a multi-task BNN based model, MtBNN, to quantify the functional impact of non-coding SNPs. We applied MtBNN to identify causal SNPs that impact gene expression via altering TF binding and modulating chromatin accessibilities. The multi-task framework provides an effective way to integrate sequence features of different chromatin-profiling signals, and the Bayesian method guarantees robustness and generalization capability of the

model. The reliability of the pre-trained model to predict TF binding affinity was validated to be comparable to the state-of-the-art models in terms of AUC scores and exhibits higher stability across multiple datasets. Furthermore, the fine-tuned MtBNN significantly surpassed the existing methods in prioritizing functional cis-regulatory SNPs including eQTLs, dsQTLs and ATAC-QTLs, although we used shorter sequences without gene expression information. In case studies, MtBNN showed superior performance in finding causal SNPs among LD groups, while the existing methods failed or partially failed to correctly prioritize them. Finally, MtBNN was able to identify potentially deleterious variants among non-coding trait-associated SNPs, which is critical for the study of genetic diseases.

To the best of our knowledge, our work is the first to apply multi-task BNN to genetic studies. As demonstrated by the performance of MtBNN, this framework is effective and reliable in distinguishing functional SNPs from harmless ones. MtBNN can be applied to various SNP-prioritizing applications, such as identifying the causal one in a group of SNPs having strong LD with each other, pinpointing lead variants from GWAS traits and determining effects of sequence variants on altering TF bindings.

There are several directions worth further exploration. One is to quantify what extra information that the distribution obtained by the BNN might contain. Previous theoretical and experimental research has provided several ways to understand and utilize the distribution (Kendall and Cipolla, 2016; Lakshminarayanan *et al.*, 2017), in particular, the uncertainty derived by the BNN. How this uncertainty can be utilized to help construct a more interpretable model remains an interesting problem. Another direction is to improve the performance of MtBNN by integrating more data such as gene expression profiles, sequence conservation between different species and other sequencing data such as Hi-C since these data have been proven to help understand functional mutations. Finally, one may consider integrating different types of QTLs given the experimentally validated high correlation between many types of such variants.

Acknowledgments

The authors thank Rui Jiang for the support of computational resources and Dongfang Wang for helpful comments.

Funding

This work has been supported in part by the National Science Foundation [grant IIS-1646333], the National Natural Science Foundation of China [grant 61772197] and the National Key Research and Development Program of China [grants 2018YFC0910404].

Conflict of Interest: none declared.

References

Bailey,T.L. *et al.* (2009) Meme suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Barrera,L.A. *et al.* (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, **351**, 1450–1454.

Bonder,M.J. *et al.* (2017) Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.*, **49**, 131.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Chatzikyriakidou,A. *et al.* (2013) Altered sequence of the ETS1 transcription factor may predispose to rheumatoid arthritis susceptibility. *Scand. J. Rheumatol.*, **42**, 11–14.

Cho,K. *et al.* (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, http://arxiv.org/abs/1406.1078.

Collobert,R. and Weston,J. (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Cohen,W.W. *et al.* (eds.) *Proceedings of the 25th International Conference on Machine Learning*, International Conference Proceeding Series. ACM, Helsinki, Finland, pp. 160–167.

Degner,J.F. *et al.* (2012) DNase I sensitivity QTLS are a major determinant of human expression variation. *Nature*, **482**, 390.

Del Rosario,R.C.-H. *et al.* (2015) Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nat. Methods*, **12**, 458.

Dozat,T. (2016) Incorporating Nesterov momentum into Adam. In: *4th International Conference on Learning Representations (ICLR), Workshop Track Proceedings*. OpenReview.net, San Juan, Puerto Rico.

Dubois,P.C. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295.

Engler,D.B. *et al.* (2014) Effective treatment of allergic airway inflammation with Helicobacter pylori immunomodulators requires BATF3-dependent dendritic cells and IL-10. *Proc. Natl. Acad. Sci. USA*, **111**, 11810–11815.

Ellinghaus,D. *et al.* (2016) Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.*, **48**, 510.

Ferreira,M.A. *et al.* (2011) Association between ORMDL3, IL1RL1 and a deletion on chromosome 17q21 with asthma risk in Australia. *Eur. J. Hum. Genet.*, **19**, 458.

Fortunato,M. *et al.* (2017) Bayesian recurrent neural networks. *CoRR*, http://arxiv.org/abs/1704.02798.

Gal,Y. and Ghahramani,Z. (2015) Bayesian convolutional neural networks with Bernoulli approximate variational inference. *CoRR*. http://arxiv.org/abs/1506.02158.

Garrett-Sinha,L.A. (2013) Review of ETS1 structure, function, and roles in immunity. *Cell. Mol. Life Sci.*, **70**, 3375–3390.

Garrett-Sinha,L.A. *et al.* (2016) The role of the transcription factor ETS1 in lupus and other autoimmune diseases. *Crit. Rev. Immunol.*, **36**, 485.

Gate,R.E. *et al.* (2018) Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.*, **50**, 1140.

Ghandi,M. *et al.* (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.*, **10**, e1003711.

Ghani,M.U. *et al.* (2017) A report on asthma genetics studies in Pakistani population. *Adv. Life Sci.*, **4**, 33–38.

Glorot,X. *et al.* (2011a) Deep sparse rectifier neural networks. In: Gordon,G.J. *et al.* (eds.) *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, *JMLR Proceedings*. JMLR.org, Fort Lauderdale, pp. 315–323.

Glorot,X. *et al.* (2011b) Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Getoor,L. and Scheffer,T. (eds.) *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, Omnipress, Bellevue, Washington, pp. 513–520.

Hernández-Lobato,J.M. and Adams,R. (2015) Probabilistic backpropagation for scalable learning of Bayesian neural networks. In: Bach,F.R. and Blei,D.M. (eds.) *Proceedings of the 32nd International Conference on Machine Learning (ICML) 2015, JMLR, Workshop and Conference Proceedings*, JMLR.org, Lille, France, pp. 1861–1869.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*, **106**, 9362–9367.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

Ioffe,S. and Szegedy,C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach,F.R. and Blei,D.M. (eds.) *Proceedings of the 32nd International Conference on Machine Learning (ICML) 2015, JMLR Workshop and Conference Proceedings*. JMLR.org, Lille, France, pp. 448–456.

Kandaswamy,R. *et al.* (2016) Genetic predisposition to chronic lymphocytic leukemia is mediated by a BMF super-enhancer polymorphism. *Cell Rep.*, **16**, 2061–2067.

Kelley,D.R. *et al.* (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.

Kendall,A. and Cipolla,R. (2016) Modelling uncertainty in deep learning for camera relocalization. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Stockholm, Sweden, pp. 4762–4769.

Khan,A. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.

Khurana,E. *et al.* (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, **17**, 93.

Kingma,D.P. and Ba,J. (2015) Adam: a method for stochastic optimization. In: *Conference Track Proceedings, 3rd International Conference on*

Learning Representations (ICLR) 2015. OpenReview.net, San Diego, CA. http://arxiv.org/abs/1412.6980.

Kurts,C. *et al*. (2010) Cross-priming in health and disease. *Nat. Rev. Immunol*., **10**, 403.

Lacoste,A. *et al*. (2018) Uncertainty in multitask transfer learning. *CoRR*. http://arxiv.org/abs/1806.07528.

Lakshminarayanan,B. *et al*. (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon,I. *et al*. (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Neural Information Processing Systems Foundation, Inc., Long Beach, CA, pp. 6402–6413.

Lee,D. *et al*. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Gene*., **47**, 955.

Li,N. *et al*. (2016a) Multiple myeloma risk variant at 7p15. 3 creates an IRF4-binding site and interferes with CDCA7L expression. *Nat. Commun*., **7**, 13656.

Li,S. *et al*. (2016b) Quantifying deleterious effects of regulatory variants. *Nucleic Acids Res*., **45**, 2307–2317.

Liu,Q. *et al*. (2018) Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics*, **34**, 732–738.

McGovern,A. *et al*. (2016) Capture HI-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol*., **17**, 212.

Orre,R. *et al*. (2000) Bayesian neural networks with confidence estimations applied to data mining. *Comput. Stat. Data Anal*., **34**, 473–493.

Pouladi,N. *et al*. (2016) Complex genetics of pulmonary diseases: lessons from genome-wide association studies and next-generation sequencing. *Transl. Res*., **168**, 22–39.

Quang,D. and Xie,X. (2016) DANQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res*., **44**, e107–e107.

Reddi,S.J. *et al*. (2018) On the convergence of adam and beyond. In: *6th International Conference on Learning Representations (ICLR), Conference Track Proceedings*. OpenReview.net, Vancouver, BC, Canada.

Rentzsch,P. *et al*. (2018) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*., **47**, D886–D894.

Ritchie,G.R. *et al*. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294.

Shahid,M. *et al*. (2015) Sequence variants on 17q21 are associated with the susceptibility of asthma in the population of Lahore, Pakistan. *J. Asthma*, **52**, 777–784.

Silver,D. *et al*. (2016) Mastering the game of go with deep neural networks and tree search. *Nature*, **529**, 484.

Svitalkova,T. *et al*. (2013) A7.24 the pentanucleotide insertion in HSPA1B gene is associated with idiopathic inflammatory myopathy. *Ann. Rheum. Dis*., **72**, A56.2–A56.

Tehranchi,A.K. *et al*. (2016) Pooled chip-seq links variation in transcription factor binding to complex disease risk. *Cell*, **165**, 730–741.

Vaswani,A. *et al*. (2017) Attention is all you need. In: Guyon,I. *et al*. (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Neural Information Processing Systems Foundation, Inc., Long Beach, CA, pp. 5998–6008.

Zhang,Y. *et al*. (2018) Enhancing HI-C data resolution with deep convolutional neural network HiCplus. *Nat. Commun*., **9**, 750.

Zheng,Y. *et al*. (2010) Role of conserved non-coding DNA elements in the FOXP3 gene in regulatory T-cell fate. *Nature*, **463**, 808.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931.