Deep Mendelian Randomization: Using Mendelian Randomization to Detect Learned Causal Relationships in Deep Learning Models

Anonymous Authors[1]

Abstract

1. Introduction

Recently, deep learning models have been used to classify genomic features such as transcription factor binding (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015), chromatin accessibility (Zhou & Troyanskaya, 2015; Kelley et al., 2016), the presence / absence of histone marks (Yin et al., 2019), and RNA binding protein binding (Alipanahi et al., 2015; Pan & Shen, 2017; Gandhi et al., 2018; Zheng et al., 2018). These models achieve high predictive accuracy on these tasks and recognize features that match those found in experiments . Furthermore, multi-task models such as DeepSEA achieve high accuracy simultaneously on multiple genomic feature prediction tasks. Given this, we would like to understand whether these multi-task models, through learning to predict multiple features jointly, gain an understanding of the causal relationships between these features.

That said, answering this question requires a methodology that can identify causal relationships in the presence of potential unobserved confounding between cause and effect. To that end, we employ Mendelian randomization, an instrumental variable approach for causal inference, to estimate learned causal effects in genomic deep learning models. Our algorithm obtains local (sequence level) and global (genome level) estimates of the linear causal relationship between two biological processes learned by a multi-task genomic prediction model. In this work, we apply our approach to estimating the learned causal effect of transcription factor binding on chromatin accessibility in a single cell type, but our method can in principle be applied to other processes that are believed to satisfy the instrumental variable assumptions.

2. Related Work

2.1. Deep Learning Model Interpretability

Local interpretability methods such as saliency maps (Simonyan et al., 2013), guided back-propagation (Springenberg et al., 2014), DeepLift (Shrikumar et al., 2017), and Deep SHAP (Lundberg & Lee, 2017), characterize how specific input features influence deep learning model predictions and, in some cases, intermediate layer activations. Even DeepLift, which was designed with genomic deep learning in mind, focus on interpreting individual model predictions rather than discovering higher-level properties and therefore complement rather than compete with Deep MR.

Saturation in-silico mutagenesis characterizes how a model's predictions for an input change as a result of all possible point mutations to the input. Saturation mutagenesis has been used to assess the learned representations of genomic deep learning models such as DeepBind (Alipanahi et al., 2015), cDeepBind (Gandhi et al., 2018), DeepSEA (Zhou & Troyanskaya, 2015), Basset (Kelley et al., 2016), and others. In our work, we use saturation mutagenesis (combined with MC-dropout, see section 2.3) to generate a set of *effect sizes* which we then provide as input to Mendelian randomization.

2.2. Mendelian Randomization (MR)

As alluded to above, Mendelian randomization is a technique for estimating linear causal effects in the presence of potential unobserved confounders. Mendelian randomization is a type of instrumental variable method in which the instrument(s) are genetic variants. While Mendelian randomization is typically used to estimate inter-phenotype causal effects from observational data, we use it in our work to estimate causal effects implied by model-generated data.

2.2.1. Mendelian Randomization Assumptions

As depicted in figure 1 under Estimate, Mendelian randomization only produces valid causal effect estimates under the following assumptions. Let Z be a variable we intend to use as an instrument (a genetic variant for example), X a

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

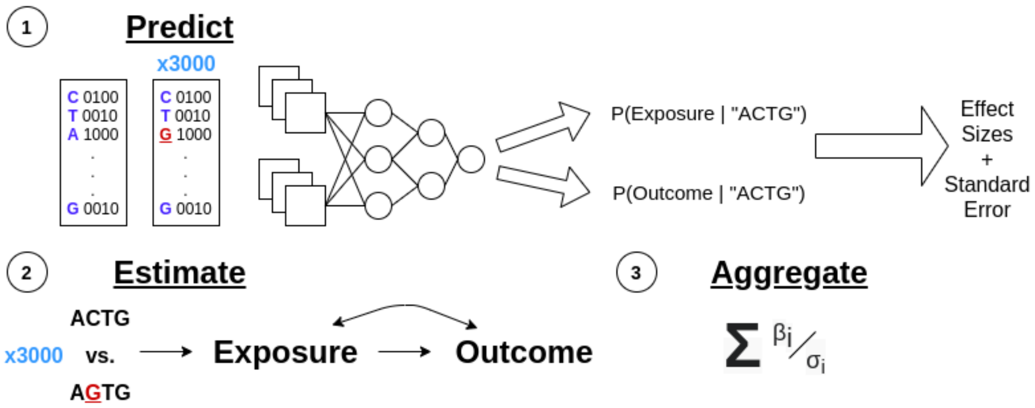

Figure 1. Graphical representation our algorithm's high-level steps. Predict corresponds to steps 1 through 4 in section 3.1. Estimate corresponds to step 5 in section 3.1. Aggregate corresponds to step 6 in 3.1.

purported cause (*exposure*), and Y a purported effect (*outcome*), and suppose that there may be unobserved confounding between X and Y, denoted by U. Then, Mendelian randomization's estimates are valid if:

1. Z is independent of U.
2. Z is not independent of X.
3. Z only influences Y through X.

However, recent MR methods such as Robust Adjusted Profile Score (Zhao et al., 2018), MR-Egger (Bowden et al., 2015), and the modal-based estimator (Burgess et al., 2018) seek to leverage multiple instruments to relax some of these assumptions without compromising the validity of results. In our work, we estimate causal effects using MR-Egger with the goal of being robust to weak instruments.

2.3. Uncertainty Estimates from Deep Learning Models

3. Methods

3.1. Algorithm Overview

Our algorithm attempts to estimate causal effect sizes between variables from predictions generated by a multi-task model. It requires as input a trained model[1] and a set of one-hot encoded sequences, representing a sequence of nucleotides in our case, for the model to make predictions on. Following the MR literature, we refer to purported causes as *exposures* and effects as *outcomes*.

Given these inputs, our algorithm outputs a set of local, sequence-specific and exposure-specific causal effects and set of global, exposure-specific causal effects. It accomplishes this (see figure 1 for a visual depiction) via the following steps for each exposure:

1. Randomly sample sequences to predict exposure and outcome values for ("reference sequences").
2. Perform *saturation in-silico mutagenesis* for each reference sequence to generate sequence length × number of nucleotides – 1 mutated sequences per original sequence.
3. For each reference and set of mutated sequences, use MC-dropout to generate predictive means and standard errors of binding probabilities for the (reference—mutated) sequences.
4. Generate $(\text{sequence length} \times \text{alphabet size} - 1)$ *effect sizes* by subtracting each reference sequence's predictive mean from the corresponding mutated sequences' predictive means. Also compute the standard errors of these differences.
5. Estimate a per-exposure, per-sequence region causal effect by running Mendelian randomization on the per-exposure, per-sequence effect sizes and their standard errors.
6. Estimate overall per-exposure factor causal effects using a meta-analysis.

This leaves us with estimates of local (transcription factor and sequence level) and global (transcription factor level) causal effects.

[1] The model could in principle be a regression or classification model, but we focus on classification in our experiments and discussion.

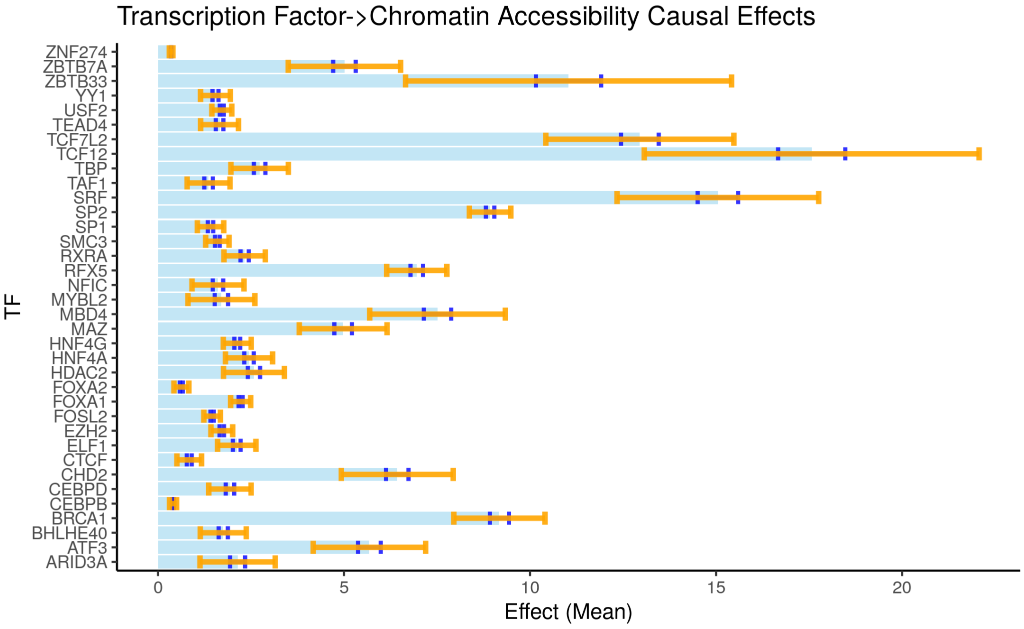

Figure 2. Per-transcription factor causaal effect estimates output by Deep MR's final step. The light blue bars show the magnitude of the overall causal effect estimated by the meta-analysis. Orange bars show τ's magnitude and dark blue the standard deviation of the mean's.

3.2. Exposure and Outcome Effect Size & Standard Error Estimation

As part of the above, we need the predicted difference in both the exposure and outcome value for every mutated, reference sequence pair. Saturation mutagenesis and MC-dropout together provide us with predicted exposure and outcome differences for each mutated sequence, reference sequence pair, but only estimates for the standard errors of the individual predictions not the difference between the two. To obtain the latter quantity, the variance of the differences between the predicted exposure and outcome values for the mutated and reference sequences, we apply the following well-known identity

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 2 \cdot \mathrm{Cov}(X, Y). \quad (1)$$

This provides us with a standard error value which we give to Mendelian randomization along with our effect size estimates.

4. Overall Causal Effect Estimation

To estimate overall causal effects at the per-exposure level, we used an inverse-variance weighted random effects meta-analysis.

TODO: Question for David - what to say here?

5. Experimental Results

To test our method, we used a pre-trained DeepSEA model provided by the Kipoi library to estimate the learned causal effect of 36 transcription factors on chromatin accessibility in the HepG2 cell type. We drew our sequence regions from DeepSEA's held-out test set, which was generated via processing the results of ChIP-seq (for transcription factors) and DNase-seq (for chromatin accessibility) experiments as part of the ENCODE project.

For each transcription factor, we randomly sampled 25 (1000 base pair) sequences on which binding was experimentally observed to occur and followed the process described above.

Causal effect estimates vary significantly across transcription factors

The results of our final meta-analysis step, shown in table 2, imply significant variation in the strength of causal relationships between different transcription factors and chromatin accessibility. While all causal effects are positive, certain transcription factors' binding seems to have a very large positive influence on chromatin accessibility. We intend to try and understand the degree to which this reflects modeling assumptions and matches experimental evidence in future work.

Sequence-level causal effect estimates vary significantly for individual transcription factors

To try and better understand how our results relate to experimental evidence, we inspected the sequence-level causal effect estimates for a known transcriptional enhancer, FOXA1, and transcriptional repressor, EZH2. Based purely on the coarse grained enhancer (FOXA1) and repressor (EZH2) (in the HepG2 cell type) classification, we'd expect FOXA1's causal effect estimates to be mostly positive and EZH2's estimates to be mostly negative. As figure shows, FOXA1 has all positive effects but, unexpectedly, all but one of EZH2's causal effect estimates are positive as well. Hypotheses for why this might occur include that EZH2 plays different roles in different positions or that selecting sequences on which binding occurs initially makes it harder to recover negative effects.

6. Discussion

In our experiment, Deep MR identifies a consistent positive effect of transcription factor binding on chromatin accessibility. Of course, obtaining true causal effect estimates from Mendelian Randomization requires relying on strong assumptions that we can't guarantee hold here. Nonetheless, we believe this provides preliminary evidence that DeepSEA partially recovers the relationship between binding certain transcription factors binding and changes in chromatin accessibility. In future work, we hope to verify our sequence-level predictions by comparing them to more fine-grained results from experiments, understand why our causal effect estimates are almost entirely positive, and better understand where the per-exposure sequence level causal effect estimate heterogeneity comes from.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

Bowden, J., Davey Smith, G., and Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

Burgess, S., Zuber, V., Gkatzionis, A., and Foley, C. N. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in mendelian randomization when a plurality of candidate instruments are valid. *International journal of epidemiology*, 47(4):1242–1254, 2018.

Gandhi, S., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. cdeepbind: A context sensitive deep learning model of rna-protein binding. *bioRxiv*, pp. 345140, 2018.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7): 990–999, 2016.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Pan, X. and Shen, H.-B. Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18(1):136, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153. JMLR. org, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Yin, Q., Wu, M., Liu, Q., Lv, H., and Jiang, R. Deephistone: a deep learning approach to predicting histone modifications. *BMC genomics*, 20(2):193, 2019.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652*, 2018.

Zheng, J., Zhang, X., Zhao, X., Tong, X., Hong, X., Xie, J., and Liu, S. Deep-rbppred: Predicting rna binding proteins in the proteome scale based on deep learning. *Scientific reports*, 8(1):1–9, 2018.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.