Determining causal interactions learned by genomic DL models with in silico mutagenesis and Mendelian randomization

Stephen Malina [1] **Daniel Cizin** [1] **David Knowles** [1 2]

Abstract

Deep learning models predict genomic features well but their ability to learn inter-feature causal relationships remains unknown. In this work, we develop Deep MR, which estimates local (sequence level) and global (feature level) inter-feature causal effects learned by genomic deep learning models. We test Deep MR using data from ENCODE and a pre-trained DeepSEA model. Deep MR finds heterogeneous causal effects across sequences of transcription factor binding on chromatin accessibility, suggesting partial recovery of true causal relationships. We propose follow-up experiments such as comparing our predicted effects to those determined by knockdown experiments to verify and extend our results.

1. Introduction

Deep learning models have achieved success predicting many genomic features such as transcription factor binding (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015), chromatin accessibility (Zhou & Troyanskaya, 2015; Kelley et al., 2016), 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) from genomic sequence. These models achieve high predictive accuracy and recognize sequence features that match those found by orthogonal experiments. In particular, multi-task models such as DeepSEA (Zhou & Troyanskaya, 2015) can accurately predict multiple genomic features simultaneously. Do such multi-task models, through learning to predict multiple features jointly, gain an implicit understanding of mechanistic, causal relationships between features?

We attempt to answer this question by developing Deep Mendelian Randomization (Deep MR), a method that can identify causal relationships in the presence of potential unobserved confounding. Deep MR combines *in silico* mutagenesis with Mendelian randomization (Lawlor et al., 2008), an instrumental variable approach for causal inference, to estimate learned causal effects in genomic deep learning models. Deep MR obtains local (sequence level) and global (genome level) estimates of (an assumed) linear causal relationship between pairs of features learned by a multi-task genomic prediction model. In this work, we apply Deep MR to estimating the implicitly learned causal effect of transcription factor (TF) binding on chromatin accessibility (CA) in one cell type, but our method can in principle be applied to other processes that might reasonably satisfy the instrumental variable assumptions.

[1]Department of Computer Science, Columbia University, New York, NY [2]New York Genome Center, New York, NY. Correspondence to: Stephen Malina <sdm2181@columbia.edu>.

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

2. Related Work

2.1. Interpreting Deep Learning Model

Local interpretation methods characterize how specific input (sequence) features influence predictions and in some cases, intermediate layer activations (e.g., saliency maps (Simonyan et al., 2013), guided back-propagation (Springenberg et al., 2014), DeepLIFT (Shrikumar et al., 2017), and DeepSHAP (Lundberg & Lee, 2017)). Even DeepLIFT, which was designed with genomic deep learning in mind, focuses on interpreting individual model predictions for a single output rather than discovering relationships between outputs and is therefore complementary to our work.

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), and Basset (Kelley et al., 2016). Here, we use saturation mutagenesis (combined with MC-dropout (Gal & Ghahramani, 2016), see Section 2.3) to generate a set of estimated variant *effect sizes* which we then provide as input to Mendelian randomization.

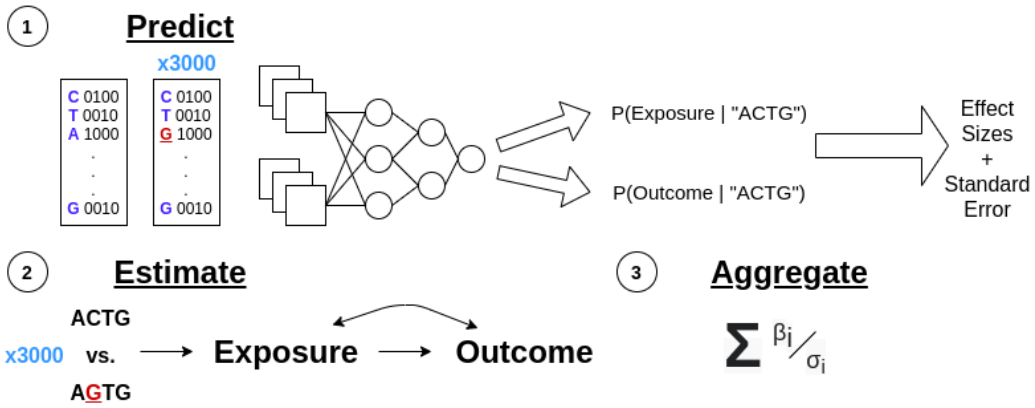

Figure 1. Graphical representation of Deep MR's high-level steps combining *in silico* mutagenesis and MR (see Section 3.1). Predict corresponds to steps 1 through 4. Estimate corresponds to step 5. Aggregate corresponds to step 6

2.2. Mendelian Randomization

Mendelian randomization (MR) is a technique for estimating linear causal effects in the presence of potential unobserved confounders. MR is an instrumental variable method where the instrument(s) are genetic variants. While MR is typically used to estimate inter-phenotype causal effects from population-scale observational data (i.e., genome-wide association studies, GWAS), here we explore its application to estimating causal effects implied by model-generated data.

2.2.1. Mendelian Randomization Assumptions

MR only produces valid causal effect estimates under the following assumptions (Figure 1 under Estimate) (Lawlor et al., 2008). Let Z be a variable we intend to use as an instrument (a genetic variant for example), X a 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, MR gives an unbiased estimate of the causal effect of X on Y if:

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

Recently developed 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) leverage multiple instruments to relax some of these assumptions without compromising the validity of results. In this work, we estimate causal effects using MR-Egger with the goal of being robust to invalid instruments.

2.3. Uncertainty Estimates from Deep Learning Models

For MR, we need standard error estimates for each predicted variant effect size. To acquire these, we use Monte Carlo Dropout (MC-dropout) (Gal & Ghahramani, 2016). MC-dropout is motivated by showing that a typical deep learning model that uses dropout can be thought of as a variational approximation to Gaussian process regression/classification. In practice, MC-dropout involves enabling dropout at test time, making repeated predictions for each sequence (50 times in our case), and then computing the predictive mean and variance as described in Gal & Ghahramani (2016) (assuming classification).

3. Methods

3.1. Method Overview

Deep MR estimates causal effect sizes between variables predicted by a multi-task model. It takes a trained model[1] and a set of one-hot encoded sequences as input. In our case, the one-hot encoded sequence represents a sequence of nucleotides for the model to make predictions on.

Deep MR outputs local, sequence-specific causal effects and 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 $(\text{sequence length} \times \text{alphabet size} - 1)$ mutated sequences per reference sequence.
3. For each reference and set of mutated sequences, use MC-dropout (Gal & Ghahramani, 2016) to generate predictive means and standard errors of binding probabilities for the reference and 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.

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

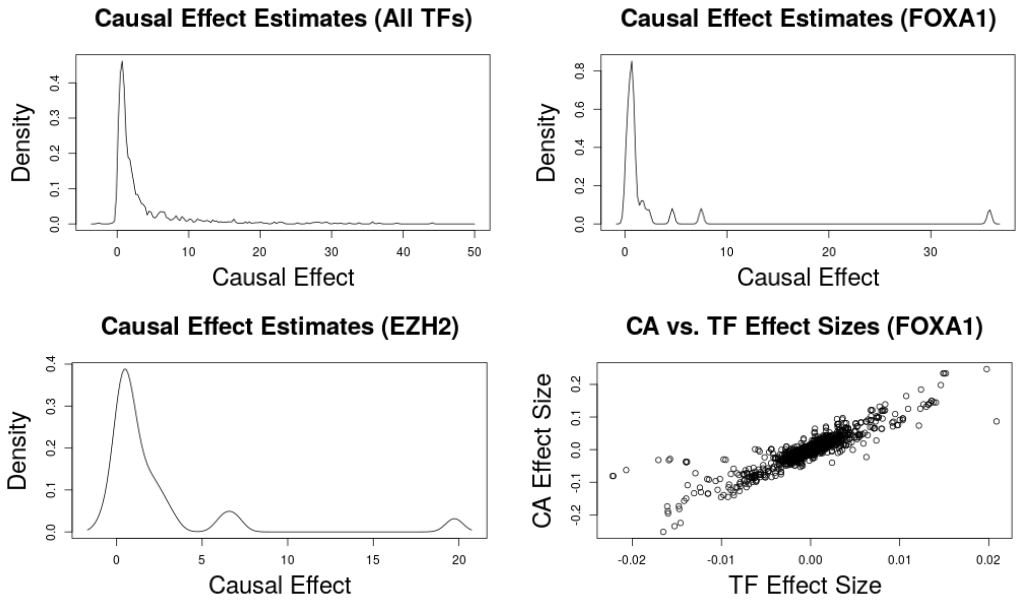

Figure 2. Kernel density estimate of per-region causal effect estimates for all TFs (a) (top left), FOXA1 (b) (top right), and EZH2 (c) (bottom left). Example scatter plot showing the effect sizes for the median causal effect estimate sequence for FOXA1 (d) (bottom right).

5. Estimate a per-exposure, per-sequence region causal effect by running MR on the effect sizes and their standard errors.
6. Estimate overall per-exposure causal effects using a random effects meta-analysis.

3.2. Exposure and Outcome Effect Size & Standard Error Estimation

MR requires variant effect estimates for each mutation for both the exposure X and outcome Y. Let $P(X = 1 \mid Z, \theta) = f_X(Z, \theta)$ and $P(Y = 1 \mid Z, \theta) = f_Y(Z, \theta)$ be the DL model for X and Y respectively with input sequence Z and parameters θ. Appealing to the interpretation of MC-dropout as approximate variational inference, the prediction for X is $P(X = 1 \mid Z) \approx \frac{1}{N}\sum_{n=1}^{N} f_X(Z, \theta^{(n)})$ where n corresponds to different random dropout masks (analogously different draws from the variational posterior). Calculating this MC estimate for both the mutant sequence m and reference r we can obtain an unbiased estimate of the variant effect $\hat{\beta}_{ZX} = P(X = 1 \mid Z = m) - P(X = 1 \mid Z = r)$. We proceed analogously for the outcome Y.

A naive estimate of the standard errors (s.e.) would use $\text{var}[\hat{\beta}_{ZX}] = \text{var}[P(X = 1 \mid Z = m)] + \text{var}[P(X = 1 \mid Z = r)]$ with the variances estimated by MC. However, this would give inflated s.e. since it ignores statistical dependence resulting from θ. We therefore instead use

$$\text{var}[\hat{\beta}_{ZX}] = \text{var}[P(X = 1 \mid Z = m) - P(X = 1 \mid Z = r)]$$
$$= \frac{1}{N}\sum_{n=1}^{N}\left[f_X(m, \theta^{(n)}) - f_X(r, \theta^{(n)})\right]^2 - \hat{\beta}_{ZX}^2.$$

In practice this requires ensuring the dropout mask used for each mutated sequence is the same as for its corresponding reference sequence for each n. The s.e. for β_{ZX} is then the square root of this expression. An analogous computation is performed to obtain the s.e. of β_{ZY}.

4. Experimental Results

To test Deep MR, we used a pre-trained DeepSEA (Zhou & Troyanskaya, 2015) model provided by the Kipoi library (Avsec et al., 2019) to estimate the learned causal effects of 36 TFs on CA in the HepG2 cell type. For each TF we randomly sample 25 1000 base pair sequence regions (labelled as having binding of the TF) from DeepSEA's held-out test set[2]. This data was generated via processing the results of ChIP-seq (for TFs) and DNase-seq (for CA) experiments as part of the ENCODE project (Consortium et al., 2004).

Linear relationship between variant effects on TF binding and accessibility. We found TF and CA variant effect estimates were approximately linearly related across sequence regions and TFs (see e.g. effects for one sequence region for FOXA1 in Fig 2d). That we see a similar pattern

[2]Full list found in supplementary Table 1 here.

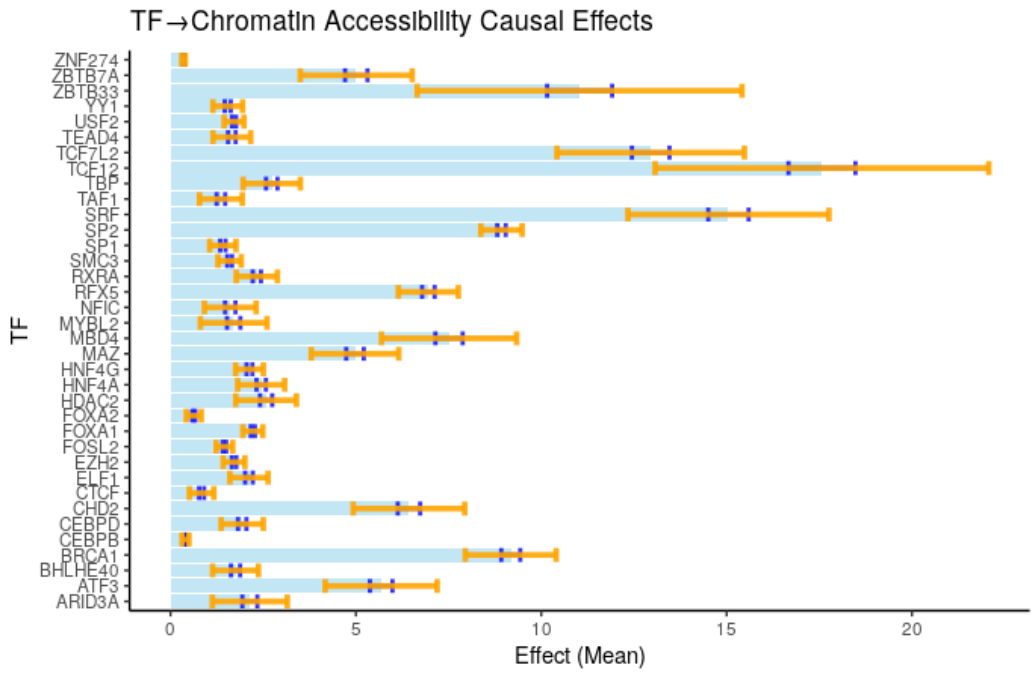

Figure 3. Per-TF causal effect estimates output by Deep MR's final step. The light blue bars show the magnitude of the overall causal effect estimated by the meta-analysis. Orange bars show τ's magnitude and dark blue the standard deviation of the mean's.

for many sequences and TFs is suggestive that MR's linear effect assumption is valid in this setting.

Causal effect estimates are mostly positive. All of the exposure-specific causal effect estimates are positive with τ intervals that never include 0 (Figure 3). This defies our initial expectation that increased probability of binding of transcription repressors such as EZH2 would lower the probability of chromatin being accessible. Furthermore, the 25% to 75% quartile range for sequence-specific causal effect estimates is .64 to 3.54 (mode: 0.41, median: 1.36, see Figure 2a's density plot for the whole distribution), and out of the 900 region-specific causal effect estimates, only 3 are negative. Together, these results suggest that increased binding of even repressive factors such as EZH2 locally increases CA (see Discussion).

Sequence-specific causal effect variance. We inspected the sequence-level causal effect estimates for a known transcriptional activator, FOXA1, and transcriptional repressor, EZH2, in HepG2. The majority of causal effect estimates for both FOXA1 and EZH2 (Figures 2b and 2c respectively) are significantly non-zero but with absolute value < 1 (medians .70 and .75 respectively) with a few outliers greater than 5. Taken at face value, this implies that many sequence regions can be mutated in a way that impacts CA via effects on TF binding, some much more dramatically than others.

Exposure-specific causal effect variance. The variance in exposure-specific causal effects (Figure 3) implies that the binding of certain TFs impacts accessibility throughout the genome much more than that of other TFs. An alternative hypothesis is that the varying impact on accessibility is specific to *the sequences we sampled for each TF*. In future work we intend to distinguish these alternative explanations by more comprehensive sampling of bound sequence regions.

5. Discussion

In our experiment, Deep MR identified a consistent positive effect of TF binding on CA. Obtaining true causal effect estimates from MR requires relying on specific assumptions that we cannot guarantee hold here. Nonetheless, we believe this provides preliminary evidence that DeepSEA at least partially recovers meaningful relationships. We intend to follow up with work to:

- Verify Deep MR's sequence-level causal effect estimates by comparing to results from TF knockdown experiments.
- Apply Deep MR to outcomes other than accessibility that are more directly associated with active transcription (such as H3K27ac) to test whether negative causal effects are then detected for repressors.
- Investigate whether causal effect estimate heterogeneity comes from binding of co-factors/interaction effects with other TFs by estimating networks of TF interactions rather than just pairwise effects.

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.

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., Banerjee, A., Kim, D. S., Beier, T., Urban, L., et al. The kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature biotechnology*, 37(6):592–600, 2019.

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.

Consortium, E. P. et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

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.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

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.