Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Deep Mendelian Randomization: Identifying and Verifying Genomic Deep Learning Models' Causal Knowledge

Stephen Malina SDM2181@COLUMBIA.EDU

Daniel Cizin TODO@TODO.EDU

David A. Knowles DAK2173@COLUMBIA.EDU

Editor: N/A

©2000 Marina Meilă and Michael I. Jordan.

Deep learning models have achieved success predicting many genomic features such as transcription factor (TF) binding (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015), chromatin accessibility (Zhou and Troyanskaya, 2015; Kelley et al., 2016), histone marks (Yin et al., 2019), and RNA binding protein (RBP) binding (Alipanahi et al., 2015; Pan and 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 and Troyanskaya, 2015) and BPNet (Avsec et al., 2020) 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.

We tested Deep MR in three experiments. In the first experiment, we simulated data based on a model of directional cooperativity between TFs and tested whether Deep MR could recover the causal relationship between the two TFs. Deep MR gave good estimates of the 'true' global causal effect but didn't perform as well at estimating sequence-region level causal effects. Therefore, in the second and third experiments, we used Deep MR to examine what global relationships between features two published models, BPNet (Avsec et al., 2020) and DeepSEA (Zhou and Troyanskaya, 2015), learn. In the BPNet experiment, Deep MR reproduced the paper's finding that 2 of the 4 TFs modeled strongly influence each others' and the two other TFs' binding but not the reverse. In the DeepSEA experiment, Deep MR . . .

0.1 Related Work

Our work draws on four threads of work spanning machine learning and statistical genetics.

0.1.1 Deep Learning for Functional Genomics

Functional genomics research maps sequence-to-function relationships between genotype and molecular phenotypes by leveraging large-scale observational data from high-throughput assays such as ChIP-seq, DNase-seq, and ATAC-seq. Understanding this mapping enables better understanding of epigenomic regulation and more accurate prediction of downstream traits. However, achieving this goal requires decoding complex relationships between high-dimensional genomic sequence inputs and interrelated outputs from large, noisy datasets. Encouraged by deep learning models' ability to overcome similar challenges in the fields of computer vision and natural language processing, genomics researchers have trained deep learning models on functional genomics datasets with moderate success.

Early work on deep learning for functional genomics proved that deep learning models could predict sequence-to-function relationships accurately and showed their promise for identifying trait-associated variants. DeepBind (Alipanahi et al., 2015), one of the earliest deep learning sequence-to-function classification models, outperformed then state-of-the-art models at predicting TF binding and RBP binding from sequence. DeepBind and other classification models – e.g. DeepSEA (Zhou and Troyanskaya, 2015) and Basset (Kelley et al., 2016) – also performed well at classifying and annotating trait-associated variants using their own predictions, identifying variants with higher accuracy than existing methods. More recent work has applied deep learning models to deepening understanding by decoding epigenomic regulatory logic. Avsec et al. (2020) trained a regression model, BPNet, to predict 4 TFs and used it to dissect the motif-based regulatory logic that governs the binding of these TFs. Together, these papers illustrate the promise of deep learning models for not only predicting function from sequence but also improving our understanding of epigenomic regulation and ability to anticipate disease risk. In our work, we seek evidence that genomic deep learning models learn the high-level relationships between features we expect them to, thereby increasing confidence in their capabilities.

0.1.2 Model Interpretability

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 and 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 and Troyanskaya, 2015), and Basset (Kelley et al., 2016). Here, we use saturation mutagenesis (combined with MC-dropout (Gal and Ghahramani, 2016), see Section **??**) to generate a set of estimated variant *effect sizes* which we then provide as input to Mendelian randomization.

0.1.3 Uncertainty Estimates and Calibration of Deep Learning Models

0.1.4 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.

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. **Unconfoundedness**: Z is independent of U,
2. Z is not independent of X, and
3. **Exclusion Restriction**: 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 a robust variant of MR-Egger with the goal of being robust to invalid instruments.

1. Methods

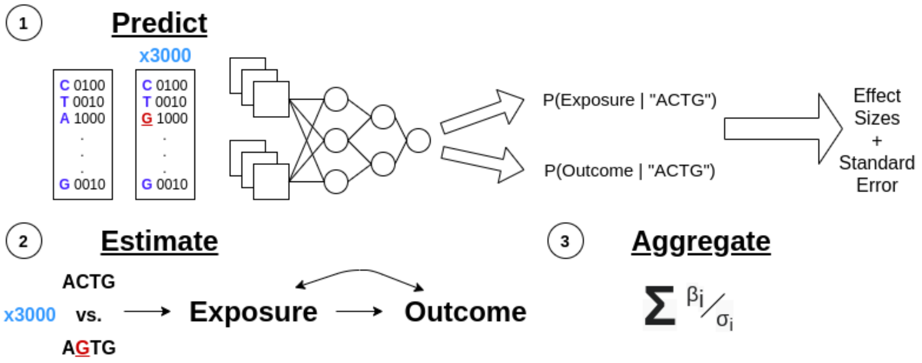

Figure 1: Graphical representation of Deep MR's high-level steps combining *in silico* mutagenesis and MR (see Section 1.1). Predict corresponds to steps 1 through 4. Estimate corresponds to step 5. Aggregate corresponds to step 6

1.1 Algorithm Overview

Deep MR estimates causal effect sizes between variables predicted by a multi-task model. It takes a trained, calibrated (regression or classification) model that outputs predictive means and standard errors 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/outcome-specific causal effects. It accomplishes this (see Figure 1 for a visual depiction) via the following steps for each exposure/outcome pair:

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 × alphabet size − 1) mutated sequences per reference sequence.
3. For each set of pairs of mutant and reference sequences, generate predictive means and standard errors for exposure and outcome features.
4. Generate (sequence length × 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. Filter instruments by effect size based on a z-score threshold to only include those that are strongly associated with the exposure.
6. Estimate a per-exposure, per-sequence region causal effect by running MR on the remaining effect sizes and their standard errors.
7. Estimate overall per-exposure causal effects using a random effects meta-analysis.

1.2 Key Assumptions

Deep MR relies on a few key assumptions regarding the DL model being used and the causal structure of the underlying data-generating process.

1.2.1 Underlying Model Performance

Because Deep MR uses variant effect estimates from a trained model rather than from real data as input to MR, the quality of its causal effect estimates is limited by the quality of the trained model. As a result, Deep MR works best when the trained model to which it's being applied achieves high accuracy on held-out test data.

1.2.2 Model Calibration

MR Egger, the Mendelian Randomization method we use, requires properly calibrated effect size and standard error estimates for each instrument. Proper calibration of standard errors matters for MR because over-confident effect size standard errors will produce over-confident causal effect confidence intervals and standard errors and vice versa for under-confidence. As mentioned, we derive our estimates of predictive mean and standard deviation, which determine effect size and standard error, from predictions generated with MC-dropout or an ensemble. In our experiments, both of these methods tend to produce un-calibrated, over-confident estimates as measured by the metrics proposed by Kuleshov et al. (2018). To remedy this, we apply and recommend isotonic regression for calibrating regression ensembles and Platt scaling for calibrating classifiers.

1.2.3 MR DAG Faithfulness

For MR to return unbiased causal effect estimates at the sequence region level, our underlying data-generating process and our model's proxy for it must both adhere to the three MR assumptions (0.1.4) and there must be a linear relationship between exposure and outcome features. In the statistical genetics setting, these assumptions can be justified in part by claims about the relationship between genotype, which is pre-natal, and potential confounders and phenotypes, both of which tend to be post-natal, assuming population structure is accounted for. We unfortunately cannot fall back on these justifications because they do not hold when looking at sequence-to-function relationships. Instead, we must re-examine each of these assumptions to determine whether they can be expected to hold. At a high level, assumption 2 is easy to satisfy because we can simply filter instruments based on their relationship to the exposure (see item 5 in Section 1.1), whereas the unconfoundedness (assumption 1), exclusion restriction (item 3), and linearity assumptions have the potential to be violated.

Instruments Independent of Unobserved Confounders Under classical MR assumptions, estimates will only be unbiased if all instruments are independent of any potential unobserved confounders. At the sequence-region level, potential unobserved confounders broadly fall into two categories: sequence-dependent and sequence-independent. Potential sequence-dependent unobserved confounders include other TFs or features which affect the exposure or outcome and which are affected by a mutation. In the case of an unobserved TF binding, this could manifest if a latent TF causally influences the exposure

and the outcome. In this scenario, mutations which impact the latent TF's binding would also affect the exposure and therefore prevent this specific mutation from being filtered out during step 5. Despite appearing as valid instruments, such mutations would clearly violate the instrument independent of unobserved confounders criterion and therefore bias classical MR estimates. Another potential sequence-dependent confounder would be an uncorrected assay bias such as GC-bias. GC-bias pushes estimates of read counts higher in regions with higher-than-average GC-content, thereby biasing models to systematically predict higher values for such regions. In MR terms, GC content acts as an unobserved confounder which causes mutations from A/T to C/G to incorrectly increase predicted exposure and/or outcome values. We don't have plausible examples of sequence-dependent confounding, but explore the effects of both types on the quality of our sequence-region level causal effect estimates in our simulation experiment nonetheless (TODO: ref).
TODO: add more to this section and decide whether to move some of it to discussion.
TODO: what are examples of potential non-sequence dependent unobserved confounders? We consider this sequence-independent because although it depends on sequence content, its effect should be almost entirely independent

Exclusion Restriction The exclusion restriction assumption requires all influence of an instrument, in our case a mutation, on the outcome to be mediated by the exposure. Thus, the latent TF confounder example given above violates the exclusion restriction assumption. The latent TF also influences the outcome variable, so mutations that influence the latent TF would also influence the outcome via the latent TF, i.e. through a pathway not mediated by the exposure. In classical MR, this violation would be enough to bias our causal effect estimates. However, the MR Egger method provides some additional conditions under which unbiased estimates are possible in the presence of one or more exclusion restriction violations. MR Egger is designed to furnish unbiased estimates of causal effects in the presence of exclusion-restriction violations as long as these violations satisfy the assumption that instrument strength for invalid instruments is independent of the magnitude of the direct effect (InSIDE assumption).

Local Linearity As discussed in Section 0.1.4, MR correctly estimates causal effects when all relationships – instrument to exposure and exposure to outcome – are linear. However, in functional genomics settings, the relationship between phenotypes can be non-linear. For example, in the case of strong TF binding cooperativity, knocking out one TF's binding will knock out the other's entirely, violating linearity. Fortunately, in our application, we only require the weaker condition of local linearity to hold in order for our method to work. Concretely, because each of our effect sizes is derived from the predicted effect of a single point mutation on the exposure/outcome phenotypes, our method works as long as the relationships between exposure and outcome predictions stays linear within the local neighborhood of the initial values.

2. Experiments

Each of our three experiments follows the same high-level process described in 1.1. We train or load an already trained model that outputs uncertainty estimates, run *in-silico mutagenesis* to obtain effect sizes and standard errors, and run MR to compute sequence-region level causal effects.

Code for all experiments can be found at `https://github.com/an1lam/deepmr`.

2.1 Simulation

2.1.1 Setup

In this section, we summarize how our simulation operates.[1] Our simulation is inspired by the one described in Finkelstein et al. (2020), but tailored towards to fit our goal of estimating causal relationships between features.

Our simulation models the binding of two TFs, an exposure TF and an outcome TF, to simulated 100-base pair DNA sequences. Binding strength (the output) in a given sequence is a noisy function of how well the sequence matches binding motifs and is measured in terms of counts, the typical output units of a high-throughput sequencing assay. Each sequence gets annotated with two output labels, exposure and outcome TF binding strength. We train convolutional neural network (CNN) models on the data produced in each scenario and use them, combined with held-out test sets, as inputs for Deep MR. Details on model parameters and training can be found in Appendix A.

We configured our simulation to support three different scenarios, no unobserved confounding between exposure and outcome, sequence-based unobserved confounding, and non-sequence-based unobserved confounding. Sequence-based unobserved confounding adds an additional confounder TF and motif which influences the binding strength of both exposure and outcome TFs to the simulation.

Sequence & Count Generative Process Sequences are simulated as follows. For a given sequence to-be-simulated, we:

1. Sample from a background distribution over the 4 DNA nucleotides.
2. With probability .75, insert motif-length sequence(s) sampled from one of the exposure motif, the outcome motif, or both motifs.
3. If a sequence-dependent confounder is being used, sample from the confounder motif and insert the sample in the sequence with probability .75.

Our counts generative process is designed based on our ultimate goal of estimating the strength of the causal relationship between our exposure and outcome TFs. To capture this, the exposure TF's binding strength is determined purely by the strength of the match between the overall sequence and the exposure's assigned motif, whereas the outcome TF's binding is a multiplicative function of both the strength of its own motif match and the strength of the exposure's. However, real assay datasets also inevitably include experimental noise. To remain faithful to this, our final counts are sampled from a Poisson

[1] For more details, see Appendix 2.3.

random variable with mean/variance equal to the 'raw' count value computed based on motif strength. Finally, following Finkelstein et al. (2020), we Anscombe transform the raw counts to produce the final output.

True Causal Effect Computation To assess the quality of our method, we need to compare its estimates to a ground truth. Deep MR estimates the effect of a unit change in the exposure on the outcome by using single point mutation that meaningfully affects the exposure as instruments. Our simulation can provide us with the true count value for any given mutated sequence, which we leverage to compute true sequence-region level causal effects. For a given sequence which contains the exposure motif, the true causal effect is the regression coefficient obtained from regressing the effect of all point mutations to bases within the exposure motif on the outcome on the corresponding effects on the exposure. This is similar to the two-stage least squares MR method (Angrist and Imbens, 1995) where all mutations within the region of the exposure motif are assumed to be valid instruments.

2.1.2 Results

Simulation Estimated Effects			
Confounding	True Global CE	Estimated Global CE	Coverage (%)
None			
Sequence-based			
Random 0.7257342 / 0.5672811 / 0.9432931	0.8164	0.9635866	

Deep MR estimates global CEs with high accuracy in . . . cases

Deep MR TODO in sequence-based and non-sequence-based confounding case

Deep MR has mediocre coverage at the sequence-region level

- Due to systematic underestimation of CEs/CI uppers

2.2 BPNet

2.3 DeepSEA

Appendix A.

Computing Counts

References

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

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription factor binding reveal soft motif syntax. *bioRxiv*, page 737981, 2020.

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

Stephen Burgess, Verena Zuber, Apostolos Gkatzionis, and Christopher N Foley. 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.

Mara Finkelstein, Avanti Shrikumar, and Anshul Kundaje. Look at the loss: Towards robust detection of false positive feature interactions learned by neural networks on genomic data. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, The 2020 ICML Workshop on Computational Biology, 2020.

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

Shreshth Gandhi, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan Frey. cdeepbind: A context sensitive deep learning model of rna-protein binding. *bioRxiv*, page 345140, 2018.

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

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

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

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

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

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

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Qijin Yin, Mengmeng Wu, Qiao Liu, Hairong Lv, and Rui Jiang. Deephistone: a deep learning approach to predicting histone modifications. *BMC genomics*, 20(2):193, 2019.

Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S Small. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652*, 2018.

Jinfang Zheng, Xiaoli Zhang, Xunyi Zhao, Xiaoxue Tong, Xu Hong, Juan Xie, and Shiyong Liu. Deep-rbppred: Predicting rna binding proteins in the proteome scale based on deep learning. *Scientific reports*, 8(1):1–9, 2018.

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