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

1. Introduction

- **Motivation**: interrogating genome-level causal relationships learned by multi-task sequence-to-function machine learning models.
- **Method summary**: Our method estimates causal effects for cause ('exposure') and effect ('outcome') feature pairs from a sequence-to-function model's predicted labels. It requires:
  - A trained multi-task (classification or regression) model.
  - Method for getting the model to output predictive means and standard errors or a full predictive distribution.
  - Sample sequence inputs.
- **Experiments summary**: To test our method, we conducted 3 experiments:
  1. Simulation experiment
  2. BPNet (Avsec et al. (2020)) experiment
  3. DeepSEA (TODO) (Zhou and Troyanskaya (2015)) experiment
- **Results summary**: our results suggest that backing out causal relationships from a sequence-to-function model is possible.
  - **Simulation experiment**: good at recovering global relationships, even in presence of confounding (TODO: double-check after re-running).
  - **BPNet experiment**: recovers expected global effect pattern.
  - **DeepSEA experiment**: TODO
- **Related work**:
  1. Builds on genomic DL literature

2. Leverage probabilistic DL model papers - MC dropout & deep ensembles
3. Complements interpretability work

2. Methods

2.1 Algorithm

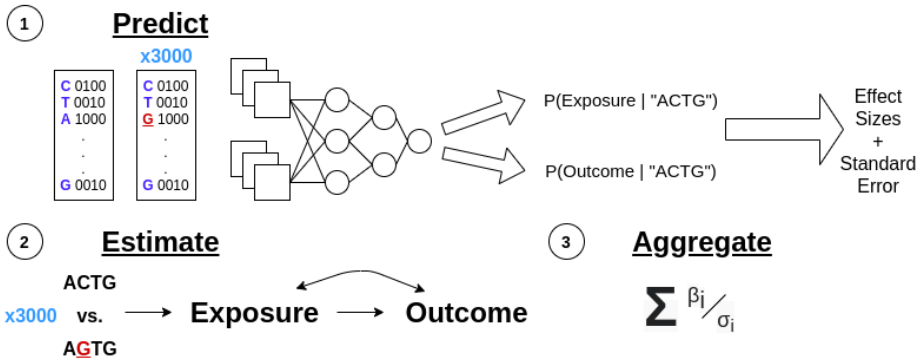


Figure 1: Graphical representation of Deep MR's high-level steps combining *in silico* mutagenesis and MR (see Section 2.1.1). Predict corresponds to steps 1 through 4. Estimate corresponds to step 5. Aggregate corresponds to step 6

Inputs Deep MR takes a calibrated, trained model 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.

Outputs Deep MR outputs local, sequence-specific causal effects and global, exposure-specific causal effects.

2.1.1 Overview

Deep MR accomplishes its goal via the following procedure:

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 $\times$ alphabet size $-$ 1) mutated sequences per reference sequence.
3. For each reference and set of mutated sequences, generate predictive means and standard errors for the reference and mutated sequences.
4. Generate (sequence length $\times$ 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 MR on the effect sizes and their standard errors.
6. Estimate overall per-exposure causal effects using a random effects meta-analysis.

2.1.2 Key Assumptions

Devote a paragraph(s) to discussing the assumptions Deep MR relies on and maybe 1 sentence just mentioning why we think the assumption is satisfied for each. These assumptions are:

- Local linearity
- MR DAG faithfulness
- Model performance
  - Accuracy upper bound and variant effect prediction warning
  - Inherited biases
  - Calibration (reference later section on dealing with this)

2.1.3 Components

Mendelian Randomization

- Introduce MR assumptions
- Discuss method we choose and its additional assumptions / features (what assumptions it allows us to weaken)

Calibrated probabilistic model

- We use deep learning models in all of our experiments
- Discuss two methods we use for making DL models probabilistic
- Reference how we calibrate them (Kuleshov et al. (2018))

3. Experiments

In which we describe the three experiments we conducted and their results.

3.1 Simulation

3.1.1 Setup

- Describe generative process (here or in appendix?)
- Three sub-experiments: no confounding, sequence-based confounding, and non-sequence-based confounding
- Questions we were trying to answer:

  Can Deep MR identify the "true" local and global causal effects?

3.1.2 Results

Simulation Estimated Effects			
Confounding	Global CE (True)	Global CE (Estimated)	Calibration
None			
Sequence-based			
Random			

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

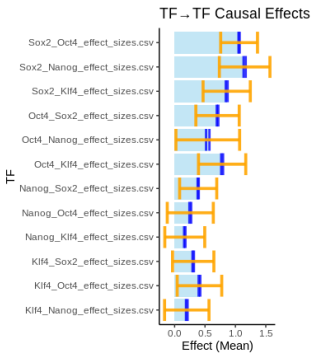
- Due to systematic underestimation of CEs/CI uppers

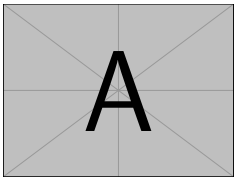
3.2 BPNet

3.2.1 Setup

- Introduce BPNet, mention ensemble and calibration method, and dataset / features
- Question we were trying to answer: can we correctly identify the key features of the TF-to-TF relationships discussed in the paper? Principally, Oct4/Sox2 strong effects on others vs. weak effect of others on others

3.2.2 Results





(a) Global CEs for all pairs of TFs predicted by BPNet. TODO: Annotate with predictions from BPNet paper.

(b) Example MR plots for 5 sequences and Oct4 → Sox2 and Klf4 → Nanog respectively.

Figure 2: BPNet results

Deep MR correctly identifies which TFs strongly do and do not influence others

3.3 DeepSEA

3.3.1 Setup

- Introduce DeepSEA, mention ensemble and calibration method, and dataset / features

- Question we were trying to answer: depends on which features we choose to use

4. Discussion

4.1 Strengths

- Deep MR recovers known patterns in both real model experiments
- Deep MR can generate new hypothesized relationships for experimental work to investigate
- Deep MR's global effect patterns can help validate and improve confidence in models
- Deep MR is compatible with existing, already trained models

4.2 Limitations

- Strong assumptions inherited from MR
- Quality of estimates depends on model quality
- Calibration of CE intervals
- Inability to determine correct direction from data

4.3 Future Work

- Network analysis
- Bi-directionality & weakening need for other assumptions
- Diagnostics for whether Deep MR can be safely applied

5. Conclusion

- Summarize most important results
- Repeat or emphasize framing of deep MR as exciting proof-of-concept for determining what causal relationships multi-task genomic deep learning models learn
- (Maybe) connect to larger context/project of trying to make multi-task genomic models more trustworthy

6. Questions

- What to test with DeepSEA? Options:
  - Stick with TFs → accessibility.
  - PolII occupancy but without doing the promoter enhancer thing?
- Should we be using the held-out set as opposed to the training set? In theory, we don't really care about generalization performance so much as just getting the most accurate sense of what the model's learned.
- What's our 'Theory of change' for the paper?
  - Inspire more work in the area
  - Provide another method for people's toolbox
- More separate discussion of local vs. global CE quality?
- Which MR method?
  - Egger has the problem with the intercept going crazy.
  - MBE seems pretty good for our use-case but is ery slow.
- Futz around with z-scores - is it worth it?
- Should we be including heterogeneous effects in the simulation?

References

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.

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

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