Cell Systems

Transparent Peer Review Record

Transcriptome data are insufficient to control false discoveries in regulatory network inference

Eric Kernfeld, Rebecca Keener, Patrick Cahan, Alexis Battle

Summary

Initial Submission: Received May 24, 2023
Scientific editor: Ernesto Andrianantoandro, Ph.D.
Preprint: https://doi.org/10.1101/2023.05.23.541948

First round of review: Number of reviewers: Three
Three confidential, Zero signed
Revision invited August 02, 2023
Major changes anticipated
Revision received May 31, 2024

Second round of review: Number of reviewers: Two
Two original, Zero new
Two confidential, Zero signed
Accepted July 22, 2024

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Cahan,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript.

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns that can be addressed by:

1. Improving the presentation by clarifying the driving question, justification for the approach and description of the methods, main advance presented, and interpretation of how the data support the claims. In particular, the conceptual advance needs to be illustrated in graphical form, in a figure, and the existing figures need revision to more clearly present the data.
2. Providing a convincing gut-check for the output of the Model X knockoffs (if it were wrong, how would you know?) and adequate benchmarking.
3. Providing a clear recommendation for best practices and a broader context for the implications of the findings - where else is inflation of FDR a problem? Where else is causal sufficiency an issue? What about other inference approaches?

As-is, the manuscript focuses too much on presenting Model X knockoffs as a tool or approach for addressing false positives in TRN inference. This leads readers to expect a TRN inference tool that avoids false positives or a formalism that controls FDR in the process of TRN inference. However, it is not clear that the Model X Knockoffs approach presented here would in fact mitigate false positives with real data, or that this is its actual utility at all. Rather, the stronger point (as alluded to in the title) seems to be in what the analysis itself reveals about FDR, causal sufficiency, and TRN inference in general. Therefore the main advance of the paper would need to be in explaining the deep relationship between these things in a rigorous way that provides guidance for future practitioners.

Reviewer #2 has the strongest grasp on the strengths of the paper and provides good guidance for improving the presentation. But addressing the concerns of Reviewer #1 is imperative - they ask for viable solutions to the problem of false positives - but if false positives are categorically unavoidable, this needs to be proven with rigor, at least theoretically. Relatedly, in the absence of having an inference method that avoids false positives, with the major contribution here being conceptual, the question of why lack of causal sufficiency inflates FDR needs to be addressed (or if this is not answerable, the reasoning for why it is not answerable needs further explication). Since all of this hinges on how well the approach performs and user confidence in the output, addressing the concerns of Reviewer #3 is also vitally

important. Reviewer #3 has also provided a PDF with annotations of some of your figures (please let me know if you cannot access this PDF). In addition to the concerns I've detailed above, I've highlighted portions of the reviews that strike me as particularly critical.

As you address these concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by Zoom, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: In this manuscript, the authors focus on the two causes of false positives when inferring transcriptional regulatory networks (TRNs): faulty hypothesis testing, or causal sufficiency. The authors propose the use of Model-X knockoffs to control false positives during conditional independence tests (CIT). Although the use of knockoffs was effective for managing false positives in CITs, the authors found the false discovery rate (FDR) still inflated due to the lack of causal sufficiency. While the analysis of false positives' causes is insightful, the manuscript falls short on several fronts, including a lack of a viable solution to the problem of false positives.

Major Comments:

1. The manuscript suffers from several formatting issues, the most crucial ones include:
a. Figures in the paper are of low resolution, making them hard to read.
b. The absence of panel labels in figures is confusing, for instance, Figure 1 lacks a clear indication of which panel is Figure 1A.
c. In several panels, the line thickness obstructs clear differentiation of overlapping lines.
d. The pattern in Figure 2E seems unnatural, requiring explanation. Additionally, terms like M3Dknockout and RegulonDB are not explained in the main text.
e. There's inconsistency in figure legends; some figures use "Expected FDR", others use "expected_fdr".

f. Algorithm 1 is not explicit, each step should be clearly defined to prevent confusion.
g. Page 13 is blank, lacking any content.
h. The "sample" method, considered one of the best benchmarking methods, lacks adequate definition in the main text.
These issues obstruct comprehension and interpretation of the manuscript, impeding an accurate evaluation.

2. The manuscript fails to clearly define the goal of the TRN analysis. The authors introduce several statistical models and assumptions, but neglect to provide formal definitions or mathematical formulations for TRN analysis or conditional independence testing, leading to ambiguities.

3. Most results compare expected FDR vs. observed FDR across an FDR threshold range from 0 to 100%. However, most analyses only focus on small FDR thresholds; it is suggested to focus more on a target FDR from 0 to 20%, as it would be more meaningful than the full range.

4. The top right panel in Figure 1 suggests the knockoff-based method fails to control FDR in toy networks from the BEELINE benchmarking framework. While the rest of the article explains that knockoffs cannot effectively control FDR due to a lack of causal sufficiency, it does not clearly articulate why this absence inflates FDR.

5. The authors do not provide adequate recommendations to avoid false positives when the Model-X knockoff approach fails. This shortfall detracts from the overall significance of the manuscript's conclusions.

Reviewer #2: OVERVIEW
This manuscript investigates why computational methods to learn transcriptional regulatory networks (TRNs) from gene expression data are generally recognized to be inaccurate despite decades of methodological research on the problem. More specifically, it focuses on the issue of false positive predictions: incorrectly predicting that a gene is under the direct control of a transcriptional regulator. The analyses within explore two possible causes for rampant false positives in existing TRN inference algorithms. One is the hypothesis testing of conditional independence between regulators and targets. The other is causal sufficiency, which relates to all factors influencing a gene's expression being observed in the experimental data available for modeling. Using the statistical technique model-X knockoffs to study false discovery rates of conditional independence testing, the manuscript concludes that causal sufficiency is the root cause for excessive false positives in TRN inference applications.

Even though the underlying issues - the abundance of false positives and invalidity of the causal sufficiency assumption - are known in the TRN inference community, this work puts a fresh take on them through the model-X knockoff formalism. One important takeaway is that permuting the gene expression data, which is a practice used in prior work to eliminate false positives, is a poor approach due to its incorrect independence assumptions and how unrealistic the resulting simulated datasets appear. This finding alone could be impactful for future TRN modeling. Although all simulations have limitations, the simulations here emphasize how modeling RNA plus protein-level information much better controls the

false discovery rate compared to simulated RNA data alone. This supports the later analyses with real data from E. coli and mouse in which unobserved factors regulating gene expression are determined to be the driver of high false discovery rates. Overall, the computational analyses are rigorous and very well-documented in supplementary GitHub repositories. Those resources include an R package that makes new technical contributions to knockoff modeling, such as improved computational scalability.

One of the main limitations of the manuscript in its present form is in the presentation. The manuscript is written with the assumption that the reader is familiar with the model-X knockoff framework, which is not introduced until the supplement. The consequence is that it can be challenging to relate these analyses to those that a computational biology or regulatory genomics audience is accustomed to in a typical TRN inference methods manuscript.

REQUIRED MAJOR REVISIONS
1) Expanding on what was noted above, the results and methods section discuss knockoffs without introducing them. The knockoff filter is not introduced until supplementary file S1, which many readers will not read. The writing quality is quite good, but the global organization of the manuscript could be improved by assuming the reader knows nothing about model-X knockoffs and needs a basic introduction to the concept before reading the results. Similarly, many readers interested in this work may be coming a TRN inference background and expect an "algorithm" for predicting a TRN. They would benefit by making it more explicit how the knockoff filter and the approach for controlling the false discovery rate across all target genes (supplementary info S5) combine to produce the predicted TRN for some specified false discovery rate.

2) Algorithm 1 and its results are unclear. It requires true regulators and dose-response curves as input. Where do these come from?

3) There are many TRN methods that attempt to acknowledge the incorrectness of the causal sufficiency assumption. The Discussion mentions ARMADA, and this is a good example of methods that do not assume regulator activity is represented by mRNA abundance. There are many other approaches in this family that estimate transcription factor activities before inferring the network structure whose influence perhaps extends back to Network Component Analysis (Liao 2003 doi:10.1073/pnas.2136632100) if not farther. In addition, there have been various types of probabilistic graphical models for TRN inference that accommodate latent variables. Discussing how others in the field recognize the invalidity of the causal sufficiency assumption and have been exploring alternatives would better place this work into context.

4) It is difficult to interpret the result in Figure 3F that the global activities of JASPAR mouse motifs in the ATAC matrix do not improve the false discovery rates. This approach for summarizing ATAC-seq data provides only a coarse summary quite different from current methods that integrate RNA-seq and ATAC-seq for TRN inference (as reviewed recently in Badia-i-Mompel 2023 doi:10.1038/s41576-023-00618-5). Current sentiment in the field is that combining these two data modalities is one of the most promising ways to overcome the severe limitations of TRN inference from RNA-seq data alone.

MINOR REVISIONS, SUGGESTIONS & COMMENTS
5) Some of the results and approaches may be familiar to a biostatistics audience but less familiar to a

computational biology audience. Additional narrative walking the reader through figure panels and expected results would help ensure their message is not lost. For instance, Q-Q plots are a core part of the results. Not all readers will be familiar with this visualization. Explaining how to interpret the different shapes in these plots of observed versus expected false discovery rate will make the results more accessible. Similar explanation and motivation would improve the rationale for including principal components in the E. coli analysis (e.g., examples of unobserved confounders) and the KNN-based exchangeability diagnostic.

6) I question some of the overall conclusions in the Conclusions section. False discovery rate control would indeed be nice to have for TRN inference, but all of the results presented here on real data show that it is not achievable with the model-X knockoffs framework. In addition, the concluding sentence is "Methods controlling FDR in TRN inference must either explicitly check the assumption of causal sufficiency, or avoid it." Based on this work and prior work on the field, it is reasonable to believe that the causal sufficiency assumption will never be met in real data. Should the recommendation be to firmly move away from TRN inference methods that make that assumption?

7) In the mouse SHARE-seq case study, Gaussian knockoffs cannot be constructed because there are only expression profiles from 57 clusters. However, this is a self-imposed limitation. The original dataset had expression profiles from 34,774 cells. Why reduce the dimensionality so greatly if it limits the type of knockoffs that are attainable?

8) BETS (Lu 2021 doi:10.1371/journal.pcbi.1008223) is a TRN inference algorithm that emphasizes controlling FDR. It is based on permutations but uses time series expression data so it permutes the temporal profiles. That strategy is relevant related work, even if it only applies to special cases.

9) Figure 3 has "naive" instead of "permuted" in the legend.

10) Supplementary file S2 contains the reference "Unable to find information for 13741696".

DATA & CODE AVAILABILTY
The supplemental repositories reflect a serious effort to make the research resources broadly available and the computational methodology transparent in a manner that follows best practices and goes beyond the typical manuscript in the field. Overall this is a major strength of the submission.

The resources are organized across one Zenodo data repository and eight GitHub repositories, which are outlined at https://github.com/ekernf01/knockoffs_paper. I looked through most of the GitHub repositories and tested code from the core repositories. Overall, I recommend archiving releases of the GitHub repositories through Zenodo, figshare, or Software Heritage. Most of the repositories have a license file, but some are missing one.

https://doi.org/10.5281/zenodo.6573413
The data repository is well-organized with readme files describing original sources of the files. The only exception is share_seq.zip, which lacks such a readme.

https://github.com/ekernf01/knockoffs_ecoli
This repository documents in detail the E. coli experiments in the manuscripts and provides a Dockerfile with the goal of creating a container that can rerun the analyses. I was able to pull the Docker image but was unsure of its intended use. Running the R scripts directly produced errors:

```
root@930a93de5e24:/# cd knockoffs_ecoli/
root@930a93de5e24:/knockoffs_ecoli# Rscript dream5_ecoli_genets.R
Loading required package: corpcor
Loading required package: longitudinal
Loading required package: fdrtool
Error in ecoli_tf_expression %>% sweep(2, colMeans(ecoli_tf_expression), :
could not find function "%>%"
Execution halted
```

There are also instructions about needing to modify the datalake. Running the bash script also installs many dependencies that have seemingly been installed by the Dockerfile already. It also produced many errors:

```
./run_on_aws.sh: line 9: sudo: command not found
./run_on_aws.sh: line 20: wget: command not found
./run_on_aws.sh: line 21: wget: command not found
unzip: cannot find or open modern_ecoli.zip, modern_ecoli.zip.zip or modern_e
coli.zip.ZIP.
unzip: cannot find or open dream5.zip, dream5.zip.zip or dream5.zip.ZIP.
Error in eval(ei, envir) :
Datalake not found. Place it in '~/datalake' or modify `dream5_ecoli_setup.R
`.
Calls: source -> withVisible -> eval -> eval
Execution halted
AWS Access Key ID [None]:
AWS Secret Access Key [None]:
Default region name [None]:
Default output format [None]:
Partial credentials found in shared-credentials-file, missing: aws_secret_acce
ss_key
```

If the goal is to have this repository serve as extremely detailed documentation of the E. coli analyses that is not necessarily runnable, that is still valuable. However, the readme should set expectations. If the goal is to build a runnable container, more of the installation should be isolated to the Dockerfile so that the bash script focuses on downloading data and running the R code. That is not a requirement though, only a suggestion to pick one path or the other and clarify the readme. A minor point is that it was unclear why the readme discusses Ubuntu 18.04 and the Dockerfile is built from ubuntu:20.04. The Dockerfile also runs the same Rscript command twice.

https://github.com/ekernf01/knockoffs_shareseq
This repository is similar to the E. coli repository but for the SHARE-seq data. I only examined the files and expect similar feedback applies.

https://github.com/ekernf01/knockoffs_BEELINE
https://github.com/ekernf01/knockoffs_boolode
These repositories document modifications to the BEELINE software and its simulator. I found this to be a helpful way to track those changes while retaining the original commit history from the BEELINE authors.

https://github.com/ekernf01/knockoffs_quick_demo
This repository provides example scripts for using the new rlookc package and reproducing additional figures from the manuscript. If the goal is to transparently document the analyses in the manuscript, it serves that purpose well. If the goal is to provide instructions for using the package on new transcriptional data, much more documentation would be needed.

https://github.com/ekernf01/rlookc
This is the core R package for model-X knockoffs. I was able to install and run the package on Linux with R version 4.1.0 but not Windows with R version 3.6.0 because RcppEigen would not compile successfully. I ran into problems running vignette_calibration.md initially because the rlookc and knockoff libraries are not loaded, but I ultimately was able to follow the vignette to completion. Automated testing of the vignettes would help catch small issues like this. Overall, the vignettes provide good instructions for someone looking to execute knockoff modeling. The vignettes that are still R files would be more beneficial in the Markdown format. However, the vignettes do not provide guidance to a user who would like to use the package for TRN inference or to model a gene expression dataset in the manner shown in the manuscript. Doing so could be pieced together based on the examples shared in the other repositories, but there are no clear instructions for a user with only basic abilities in R and a new expression dataset.

https://github.com/ekernf01/jlookc
https://github.com/ekernf01/pylookc
These are Julia and Python packages that load saved knockoffs from the R package. Due to the sparse documentation, it is unclear what the use case is. If a user has gone through the trouble of generating the knockoffs in R, it seems likely they would continue with downstream analysis in R. If the Julia and Python packages directly called rlookc (e.g. with rpy2) that would enable different workflows, but that is out of scope (which is fine).

I only tested the Python package and faced some problems. I installed it in a fresh conda environment
$ conda create -n pylookc python=3.9
$ conda activate pylookc
$ pip install -e .
Installing collected packages: peppercorn, pylookc
Not all required dependencies were installed, for example, numpy was missing. The LICENSE.txt also needs to be updated. There are also boilerplate files like package_data.dat and simple.py that could be removed.

Reviewer #3: Manuscript summary
In the manuscript, the authors tackle the problem of transcriptional regulatory network (TRN) inference,

where existing methods are known to be plagued by false positives. The authors propose a new approach to this problem based on the model-X knockoffs methodology, and evaluate this approach on a combination of fully simulated, partially simulated, and real data. The conclusion is that model-X knockoffs generally improves the false discovery rate (FDR) in fully and partially simulated data compared to the permutation approach, but is unable to control the FDR on real data. The authors hypothesize that unmeasured confounders are to blame for this persistent FDR inflation.

Manuscript strengths
According to the authors, "empirical tests of advertised FDR rates have not been reported for any category of TRN inference method." If this is the case (I can't confirm this due to insufficient familiarity with the TRN literature), then this manuscript is an important step towards moving the field towards rigorous calibration checks. The various calibration checks proposed by the authors can serve as useful templates for future efforts to this end. In addition, the authors have proposed the novel leave-one-out-knockoff method (rlookc) and accompanying software, which may be of independent interest. Compared to the existing brute-force solution, the authors show that rlookc provides substantial computational acceleration. Finally, the authors evaluate their method on several real datasets, which enhances the robustness of their conclusions.

Manuscript weaknesses
Narrow scope of benchmarking studies
The authors note in their introduction that "dozens of TRN inference methods have been invented" but that "empirical tests of advertised FDR rates have not been reported for any category of TRN inference method." These two facts, taken together, underscore the importance of benchmarking the FDR control of a broad range of existing methods. On the other hand, the only comparison the authors make throughout their manuscript is the permutation-based method. The only other methods compared to is GeneNets, which appears in one panel of one figure. The authors state that "popular TRN inference methods based on tree ensembles or mutual information can account for both nonlinearity and indirect relationships, but they do not provide finite-sample FDR control." What is the authors basis for this claim? If it is theoretical, then they should note that model-X knockoffs also do not provide provable finite-sample FDR control in the case when the joint distribution of the features is unknown a priori (the case considered in this manuscript). Therefore, it is worthwhile at least to compare the authors' methodological proposal also to "popular TRN inference methods based on tree ensembles or mutual information."

Insufficient justification for the conclusion that "Model-X knockoffs control FDR in conditional independence testing"
The title of the first Results section is "In biochemical simulations, model-X knockoffs control FDR in TRN inference without using the true data-generating distribution." However, Figure 1B shows nontrivial FDR inflation for small target FDR levels (the important range) for the mixture knockoffs method, annotated with red ovals below. Saying that "model-X knockoffs approximately control FDR in TRN inference" would be more accurate.

In the semi-synthetic experiments, the authors found decent FDR control for the knockoffs method. They stated that "Since FDR control with simulated targets does not translate to FDR control with real targets, this microarray dataset must not meet the causal sufficiency criterion." However, another explanation that

cannot be ruled out is that the mechanism by which they simulated their targets (Algorithm 1) does not match that in the real data. In this simulation mechanism (which unfortunately is not described very clearly), it appears that each gene has exactly one regulator, and that each simulation involves just one target gene. If either of these are correct, then the simulation mechanism is not particularly realistic, and the fact that knockoffs does well at the conditional independence task on this particular semi-synthetic data does not imply that it will continue to do so with other simulated target mechanisms.

Insufficient justification for the conclusion that "relative to gold standards, conditional independence testing via the knockoff filter outperforms permutation-based testing"
In Figure 2D (reproduced below), the permutation test outperforms the knockoffs-based tests for the "chip and M3Dknockout" setting for low target FDRs (the most important range); this is annotated in the figure with a red oval. The authors should acknowledge this, while noting that the knockoffs-based methods (especially glasso_1e-04) does outperform the permutation method in the "chip and RegulonDB knockout" setting.

Additional points needing clarification or correction
1. The authors state that "With RNA only, static methods cannot infer directionality, so for RNA only, FDR was calculated with backwards edges counted as correct (Fig. 1B)." This sentence suggests that, if given the protein data, knockoffs can infer directionality. However, knockoffs is a conditional independence testing method rather than a causal inference method, so it cannot infer directionality even when the causal sufficiency assumption is satisfied.
2. The authors state that "Regarding causal sufficiency, this criterion was satisfied given RNA + protein data, which led to better FDR control." How do the authors conclude this? Is this based on knowledge of the data-generating model? In general, the causal sufficiency assumption is uncheckable based on data alone. Even informally, why do we expect that protein expression is a confounding variable, which when included in the analysis helps with causal sufficiency?
3. Should Figure 1A include a lack of causal sufficiency as an obstacle to FDR control for all three classes of methods?
4. The authors use three values for the regularization parameter for the graphical lasso. A more standard approach would be to use cross-validation.
5. In some figures, the expected FDR for some methods is bounded from below. The authors state that in cases of low power they were "unable to assess observed FDR for sets of hypotheses with low expected FDR." However, the FDR for an empty rejection set is defined by convention as zero. Perhaps the authors find that having FDRs of zero will make their plots cluttered or misleading; in this case, they should say so.
6. The authors state that "To adjust for confounders, knockoffs were computed after appending columns (features) to the TF expression matrix containing either non-genetic perturbations or non-genetic perturbations and the top principal components (Fig. 2E)." Are the authors creating knockoffs also for these additional features as well? Note that creating knockoffs for features not used for testing unnecessarily induces higher correlations among features that are used for testing and their knockoffs.
7. The authors should clean up their figures, including labeling panels with letters.

Cell Systems

Transparent Peer Review Record

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Cahan,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication. Please note that Reviewer #1 was not available to re-review your manuscript, but we're moving forward based on the comments of the other reviewers.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager. ***We hope to receive your files within 5 business days. Please email me directly if this timing is a problem or you're facing extenuating circumstances.***

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

Transparent Peer Review: Thank you for electing to make your manuscript's peer review process transparent. As part of our approach to Transparent Peer Review, we ask that you add the following sentence to the end of your abstract: "A record of this paper's Transparent Peer Review process is included in the Supplemental Information." Note that this ***doesn't*** count towards your 150 word total!

Also, if you've deposited your work on a preprint server, that's great! Please drop me a quick email with your preprint's DOI and I'll make sure it's properly credited within your Transparent Peer Review record.

Manuscript Text:

- We do not allow appendices. Please incorporate the Mathematical Appendix into STAR Methods.
- Please only use the word "significantly" in the statistical sense.

Figures and Legends:

Please look over your figures keeping the following in mind:

- When data visualization tools are used (e.g. UMAP, tSNE), please ensure that the dataset being visualized is named in the figure legend and, when applicable, its accession number is included.
- When color scales are used, please define them, noting units or indicating "arbitrary units," and specify whether the scale is linear or log.
- Please ensure that every time you have used a graph, you have defined "n's" specifically and listed statistical tests within your figure legend.
- Please ensure that all figures included in your point-by-point response to the reviewers' comments are present within the final version of the paper, either within the main text or within the Supplemental Information.

Thank you!

Reviewer comments:

Reviewer #2: OVERVIEW
The revisions address the major comments from my initial review. The new Box 2 and Appendix 4 are strong additions for readers who do not have this technical background.

REQUIRED MAJOR REVISIONS
None

MINOR REVISIONS, SUGGESTIONS & COMMENTS

1) The Figure 6c caption states "The top margin shows cell count cutoffs" but these counts are in the legend at the bottom.

2) Some of the text and figures use possessives instead of plurals, for example, TF's, DOI's, and PC's

3) There are broken data links in the text and key resources table:
- https://regulondb.ccg.unam.mx/menu/download/full_version/files/10.9/regulonDB10.9_Data_Dist.tar.gz
- https://regulondb.ccg.unam.mx/highthroughputdatasetssearch?term=all
- https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-no-cell-sorting-standard-2-0-0

DATA & CODE AVAILABILTY
The authors made the changes I suggested to their GitHub and Zenodo repositories. All the new Zenodo archives of their software repositories are excellent. In addition to listing the Zenodo archives for their software in the key resources table, they may still want to link to the main GitHub repo https://github.com/ekernf01/knockoffs_paper in the Data and code availability statement.

However, even with the clarified instructions for running the Docker container in https://github.com/ekernf01/knockoffs_ecoli I still encountered errors:

```
root@e12340da77a9:/# cd knockoffs_ecoli/
root@e12340da77a9:/knockoffs_ecoli# ./run_in_docker.sh
Warning: invalid package 'rlookc'
Error: ERROR: no packages specified
Warning message:
In install.packages("rlookc", repos = NULL, type = "source", lib = Sys.getenv("R_LI
BS_USER")) :
installation of package 'rlookc' had non-zero exit status
Fetching data...
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 265 100 265 0 0 311 0 --:--:-- --:--:-- --:--:-- 311
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 253 100 253 0 0 435 0 --:--:-- --:--:-- --:--:-- 434
Archive: modern_ecoli.zip
End-of-central-directory signature not found. Either this file is not
a zipfile, or it constitutes one disk of a multi-part archive. In the
latter case the central directory and zipfile comment will be found on
the last disk(s) of this archive.
unzip: cannot find zipfile directory in one of modern_ecoli.zip or
modern_ecoli.zip.zip, and cannot find modern_ecoli.zip.ZIP, period.
Archive: dream5.zip
End-of-central-directory signature not found. Either this file is not
a zipfile, or it constitutes one disk of a multi-part archive. In the
```

latter case the central directory and zipfile comment will be found on
the last disk(s) of this archive.
unzip: cannot find zipfile directory in one of dream5.zip or
dream5.zip.zip, and cannot find dream5.zip.ZIP, period.

Prepping data.
Error in setwd(dir = new) : cannot change working directory
Calls: source ... withVisible -> eval -> eval -> -> setwd
Execution halted

Prepping data.
Error in setwd(dir = new) : cannot change working directory
Calls: source ... withVisible -> eval -> eval -> -> setwd
Execution halted
root@e12340da77a9:/knockoffs_ecoli# cd ..
root@e12340da77a9:/# knockoffs_ecoli/run_in_docker.sh
* installing *source* package 'rlookc' ...
** using staged installation
** R
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (rlookc)
Fetching data...
mkdir: cannot create directory '/root/datalake': File exists
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 265 100 265 0 0 264 0 0:00:01 0:00:01 --:--:-- 264
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 253 100 253 0 0 354 0 --:--:-- --:--:-- --:--:-- 353
Archive: modern_ecoli.zip
End-of-central-directory signature not found. Either this file is not
a zipfile, or it constitutes one disk of a multi-part archive. In the
latter case the central directory and zipfile comment will be found on
the last disk(s) of this archive.
unzip: cannot find zipfile directory in one of modern_ecoli.zip or
modern_ecoli.zip.zip, and cannot find modern_ecoli.zip.ZIP, period.
Archive: dream5.zip

End-of-central-directory signature not found. Either this file is not
a zipfile, or it constitutes one disk of a multi-part archive. In the
latter case the central directory and zipfile comment will be found on
the last disk(s) of this archive.
unzip: cannot find zipfile directory in one of dream5.zip or
dream5.zip.zip, and cannot find dream5.zip.ZIP, period.
mkdir: cannot create directory 'v28': File exists

Prepping data.
Error in setwd(dir = new) : cannot change working directory
Calls: source ... withVisible -> eval -> eval -> -> setwd
Execution halted

Prepping data.
Error in setwd(dir = new) : cannot change working directory
Calls: source ... withVisible -> eval -> eval -> -> setwd
Execution halted

Reviewer #3: Thank you for addressing my comments. I have a few more comments and questions:

- Why did you omit the other methods (GeneNet, Gaussian Mirror, BINCO) from the BEELINE simulation in Figure 2?
- Do the axis labels or orderings in the x-axes in panels D and E of Figure 2 have any meaning? If not, perhaps another visualization would be appropriate.
- On p. 6, you state that "Per-target FDR control and global FDR control are not equivalent (Appendix 3, Methods S1), and simulations indicate that pooling improves global FDR control (Fig. S1C)." In Appendix 3, you state that "Joint choice of threshold is not mathematically guaranteed to our knowledge, but in simulations, it seems to resolve this issue." It would be helpful to clarify also in the main text that choosing a threshold jointly does not come with theoretical guarantees.
- On p. 7, you state "When protein levels and transcription rates are revealed, proteins are assumed to regulate transcripts and not vice versa, so backwards edges are counted as false positives." Since knockoffs tests conditional independence, its output is not directional. Therefore, there is no distinction between backwards edges and forwards edges in the output of knockoffs.
- Currently, the figure titles in the captions for Figures 4 and 5 are the same. I assume this should not be the case.

Summary of revisions made in response to reviews

Reviewer #1:

In this manuscript, the authors focus on the two causes of false positives when inferring transcriptional regulatory networks (TRNs): faulty hypothesis testing, or causal sufficiency. The authors propose the use of Model-X knockoffs to control false positives during conditional independence tests (CIT). Although the use of knockoffs was effective for managing false positives in CITs, the authors found the false discovery rate (FDR) still inflated due to the lack of causal sufficiency.

> **We thank the reviewer for their thoughtful and thorough comments and suggestions.**

While the analysis of false positives' causes is insightful, the manuscript falls short on several fronts, including a lack of a viable solution to the problem of false positives.

> **This comment indicates that we did not clearly articulate and explain the most important message of our study. The main finding of our study is that the absence of causal sufficiency *prohibits* FDR control of TRNs inferred from transcriptomic data. Prior work on FDR-controlled TRNs has not systematically compared predicted FDR against any gold standard (1–7), leaving this gap undiagnosed. Prior work also required strong technical assumptions such as linearity or absence of indirect effects, so that even if excess false discoveries had been detected, their root causes would remain unclear. Our primary contribution is to study FDR calibration on real data under flexible, testable modeling assumptions. In doing so, we have found that not only is FDR control highly unlikely to be achievable, but we have pinpointed why this is the case. We have fully rewritten the Introduction to clarify the intent and main findings of our study (pg5).**

Major Comments:

1. The manuscript suffers from several formatting issues, the most crucial ones include:

> **We apologize for these formatting issues, and have addressed then as follows:**

a. Figures in the paper are of low resolution, making them hard to read.

> **We apologize for this issue. We maintain scalable vector versions of all figures, and we have confirmed legibility in the PDF generated by the editorial manager's automated build process.**

b. The absence of panel labels in figures is confusing, for instance, Figure 1 lacks a clear indication of which panel is Figure 1A.

We have added panel labels throughout.

c. In several panels, the line thickness obstructs clear differentiation of overlapping lines.

For improved clarity, we have re-generated Fig. 2B-D (now labeled Fig. 2B, Fig. 3, and Fig. 4A), Fig. S1A, and all of Figure 3 (now labeled Figure 5).

d. The pattern in Figure 2E seems unnatural, requiring explanation. Additionally, terms like M3Dknockout and RegulonDB are not explained in the main text.

Many thanks to the reviewer for catching this: we have found and corrected a programming error that was obfuscating figures 2E and 2F (which are now figures 4A,B). We also added explanation for these terms in the figure legend, and we have clarified that the observed FDRs are volatile because there are so few discoveries at low cutoffs (pg17-18).

e. There's inconsistency in figure legends; some figures use "Expected FDR", others use "expected_fdr".

We have updated all relevant axis labels to “Reported FDR” and “Observed FDR”.

f. Algorithm 1 is not explicit, each step should be clearly defined to prevent confusion.

We have added explanations to Algorithm 1 explicitly defining how regulators and dose-response curves are selected (pg10).

g. Page 13 is blank, lacking any content.

We have removed the page break following Algorithm 1.

h. The "sample" method, considered one of the best benchmarking methods, lacks adequate definition in the main text.

We have clarified in the text that the “sample” method selects regulators for each target by using model-X knockoffs generated from a Gaussian whose covariance equals the sample covariance (pg7).

These issues obstruct comprehension and interpretation of the manuscript, impeding an accurate evaluation.

We hope that the changes listed above help to make it easier to evaluate our main contributions.

2. The manuscript fails to clearly define the goal of the TRN analysis. The authors introduce several statistical models and assumptions, but neglect to provide formal definitions or mathematical formulations for TRN analysis or conditional independence testing, leading to ambiguities.

We thank the reviewer for pointing this out. We agree and have now provided explicit definitions and mathematical formulations for the models and assumptions that we introduce as follows:

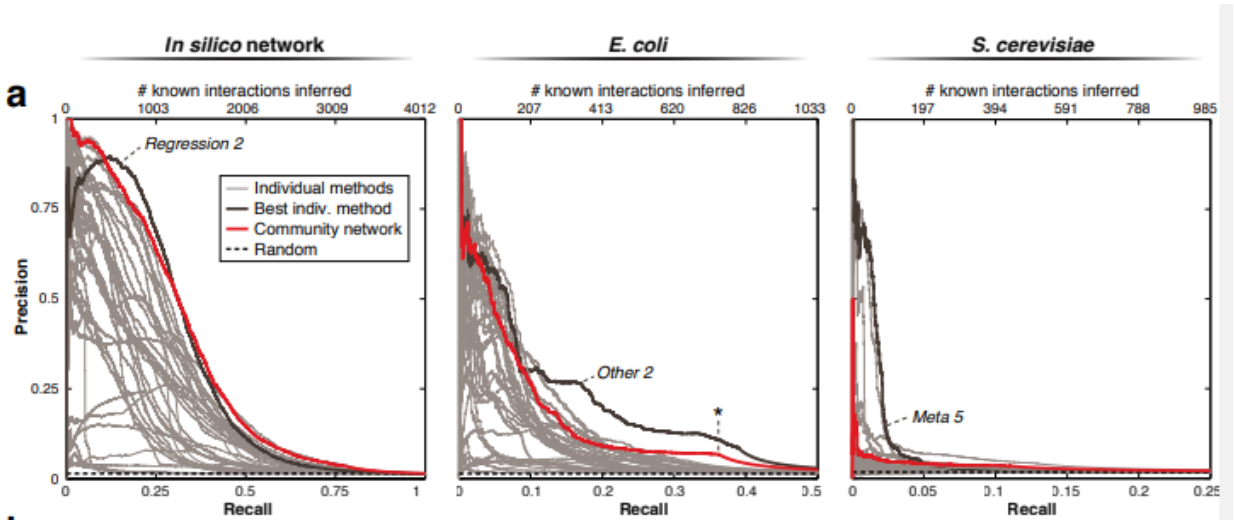
- **Regarding TRNs: we now start with an explicit definition of TRNs, pg3: "A transcriptional regulatory network (TRN) is the set of direct regulatory relationships between transcription factors and their target genes in a given biological system". We also provide a formal definition of "direct regulators" and connect to mathematical causal inference theory in a new appendix (supplemental file S4).**
- **Regarding conditional independence testing: we now outline the knockoff filter and cite literature with detailed descriptions of the assumptions underlying knockoff generation in Box 2 (pg 5-6). We also define FDR in Box 1 (pg 4-5).**

We hope that these changes will resolve any ambiguities.

3. Most results compare expected FDR vs. observed FDR across an FDR threshold range from 0 to 100%. However, most analyses only focus on small FDR thresholds; it is suggested to focus more on a target FDR from 0 to 20%, as it would be more meaningful than the full range.

This is an excellent point: TRN predictions in the 0 to 20% FDR range are the subject of more broad interest because, by definition, they should be more reliable and experimentally verifiable. We have added reports of performance on all gold standards at 20% expected FDR in table S6 (*E. coli*) and table S7 (human and mouse).

However, prior benchmarks have demonstrated that achieving FDRs of 20%, which corresponds to an observed precision of 80%, can be accomplished only with drastic loss of sensitivity. As an example, we have pasted figure S3A from the DREAM 5 paper below. The observed FDR on this plot is one minus the precision (20% FDR is 80% on the y-axis). The number of hypotheses returned is on the top x-axis. At 20% FDR, it is hard to read, but across all methods it looks like 0 to 1000 hypotheses *in silico*, 0 to 40 in *E. coli*, and 0 in *S. cerevisiae*. In many of our *E. coli* analyses, there are no testable hypotheses at 20% FDR (Table S6). Thus, we continue to include the full range of expected FDR's in our figures.



4. The top right panel in Figure 1 suggests the knockoff-based method fails to control FDR in toy networks from the BEELINE benchmarking framework. While the rest of the article explains that knockoffs cannot effectively control FDR due to a lack of causal sufficiency, it does not clearly articulate why this absence inflates FDR.

We thank the reviewer for raising this important point. Lack of causal sufficiency occurs when unobserved factor(s) are causally linked to more than one observed variable. For example, if a TF and a gene that is not its direct target are both controlled by retinoic acid levels, unmeasured variation in retinoic acid levels could lead to a correlation between the TF and the non-target. We explain this in more detail in a new appendix (S4), including simulations to illustrate how incomplete observations or unmeasured confounding can lead to violations of causal sufficiency and inflate false positive rates in TRN inference.

5. The authors do not provide adequate recommendations to avoid false positives when the Model-X knockoff approach fails. This shortfall detracts from the overall significance of the manuscript's conclusions.

Please see our introductory remarks above where we clarified that the main finding of our study is that the absence of causal sufficiency prohibits FDR control of TRNs inferred from transcriptomic data. While our findings may be perceived as a 'negative result', it is nonetheless important because it will channel the field's collective efforts into other, more productive methods to mapping TRNs. For example, we discuss TRN inference methods that allow for the presence of unmeasured confounding (Discussion, pg20) (8,9). Equipping these methods with finite-sample FDR control would be a potentially valuable future research direction, but is outside the scope of current study.

Reviewer #2:

OVERVIEW
This manuscript investigates why computational methods to learn transcriptional regulatory networks (TRNs) from gene expression data are generally recognized to be inaccurate despite decades of methodological research on the problem. More specifically, it focuses on the issue of false positive predictions: incorrectly predicting that a gene is under the direct control of a transcriptional regulator. The analyses within explore two possible causes for rampant false positives in existing TRN inference algorithms. One is the hypothesis testing of conditional independence between regulators and targets. The other is causal sufficiency, which relates to all factors influencing a gene's expression being observed in the experimental data available for modeling. Using the statistical technique model-X knockoffs to study false discovery rates of conditional independence testing, the manuscript concludes that causal sufficiency is the root cause for excessive false positives in TRN inference applications.

Even though the underlying issues - the abundance of false positives and invalidity of the causal sufficiency assumption - are known in the TRN inference community, this work puts a fresh take on them through the model-X knockoff formalism. One important takeaway is that permuting the gene expression data, which is a practice used in prior work to eliminate false positives, is a poor approach due to its incorrect independence assumptions and how unrealistic the resulting simulated datasets appear. This finding alone could be impactful for future TRN modeling. Although all simulations have limitations, the simulations here emphasize how modeling RNA plus protein-level information much better controls the false discovery rate compared to simulated RNA data alone. This supports the later analyses with real data from E. coli and mouse in which unobserved factors regulating gene expression are determined to be the driver of high false discovery rates. Overall, the computational analyses are rigorous and very well-documented in supplementary GitHub repositories. Those resources include an R package that makes new technical contributions to knockoff modeling, such as improved computational scalability.

One of the main limitations of the manuscript in its present form is in the presentation. The manuscript is written with the assumption that the reader is familiar with the model-X knockoff framework, which is not introduced until the supplement. The consequence is that it can be challenging to relate these analyses to those that a computational biology or regulatory genomics audience is accustomed to in a typical TRN inference methods manuscript.

> **We thank the reviewer for the overview of our paper and its contributions and limitations. Below, we attempt to address these concerns in detail, including the presentation.**

REQUIRED MAJOR REVISIONS
1) Expanding on what was noted above, the results and methods section discuss knockoffs without introducing them. The knockoff filter is not introduced until supplementary file S1, which

many readers will not read. The writing quality is quite good, but the global organization of the manuscript could be improved by assuming the reader knows nothing about model-X knockoffs and needs a basic introduction to the concept before reading the results. Similarly, many readers interested in this work may be coming a TRN inference background and expect an "algorithm" for predicting a TRN. They would benefit by making it more explicit how the knockoff filter and the approach for controlling the false discovery rate across all target genes (supplementary info S5) combine to produce the predicted TRN for some specified false discovery rate.

> **We agree this is a crucial shortcoming, and we have heavily revised the presentation to improve readability for a broad audience. We now describe the knockoff filter in the introduction (Box 2, pg 5-6) and we discuss how a network is constructed in the first section of the results (pg 6).**

2) Algorithm 1 and its results are unclear. It requires true regulators and dose-response curves as input. Where do these come from?

> **These can be provided to our software by the user, but we have clarified how we choose regulators and dose-response curves in algorithm 1 (pg10).**

3) There are many TRN methods that attempt to acknowledge the incorrectness of the causal sufficiency assumption. The Discussion mentions ARMADA, and this is a good example of methods that do not assume regulator activity is represented by mRNA abundance. There are many other approaches in this family that estimate transcription factor activities before inferring the network structure whose influence perhaps extends back to Network Component Analysis (Liao 2003 doi:10.1073/pnas.2136632100) if not farther. In addition, there have been various types of probabilistic graphical models for TRN inference that accommodate latent variables. Discussing how others in the field recognize the invalidity of the causal sufficiency assumption and have been exploring alternatives would better place this work into context.

> **Many thanks for this interesting reference! We are aware of only a limited range of TRN structure inference strategies accommodating latent variables, and to our understanding, network structure is a required input to the NCA algorithm, not something NCA can predict. But, we are certainly open to citing additional work, and based on this feedback, we have added some new references.**
>
> - **We now cite Liao et al. 2003 (pg 17-18).**
> - **We now cite the DoRothEA paper (10), which benchmarks a wide variety of methods for estimating TF activity (pg 17-18).**
> - **We cite the CLR paper (11), which was inspired by robustness to confounding or inadequate normalization (pg 20).**
> - **We cite a new work on identifiability of latent causal models (9) from Caroline Uhler's group. (pg20)**

4) It is difficult to interpret the result in Figure 3F that the global activities of JASPAR mouse motifs in the ATAC matrix do not improve the false discovery rates. This approach for summarizing ATAC-seq data provides only a coarse summary quite different from current methods that integrate RNA-seq and ATAC-seq for TRN inference (as reviewed recently in Badia-i-Mompel 2023 doi:10.1038/s41576-023-00618-5). Current sentiment in the field is that combining these two data modalities is one of the most promising ways to overcome the severe limitations of TRN inference from RNA-seq data alone.

> **We agree this is relevant, and we have added new multi-omics analyses accordingly.**
>
> - **We added analyses using genome-wide accessibility of all motifs as an alternate measure of TF activity (Fig 5E, Results pg17-18).**
> - **We added analyses using motif occurrences as a filter for specific hypotheses (Fig. 5G, Results pg17-18).**
> - **We added a new dataset with simultaneous RNA and ATAC on human PBMC's, which is analyzed in the same way as the mouse skin SHARE-seq data throughout Figure 5 (Results pg16-18).**

MINOR REVISIONS, SUGGESTIONS & COMMENTS
5) Some of the results and approaches may be familiar to a biostatistics audience but less familiar to a computational biology audience. Additional narrative walking the reader through figure panels and expected results would help ensure their message is not lost. For instance, Q-Q plots are a core part of the results. Not all readers will be familiar with this visualization. Explaining how to interpret the different shapes in these plots of observed versus expected false discovery rate will make the results more accessible. Similar explanation and motivation would improve the rationale for including principal components in the E. coli analysis (e.g., examples of unobserved confounders) and the KNN-based exchangeability diagnostic.

> **Thank you for this suggestion to improve the accessibility of our manuscript to a broader audience. We have updated the legends to more thoroughly describe the expected results of the initial Q-Q plots, the KNN diagnostic, and the use of PCA as a surrogate for confounders (pg 21-22).**

6) I question some of the overall conclusions in the Conclusions section. False discovery rate control would indeed be nice to have for TRN inference, but all of the results presented here on real data show that it is not achievable with the model-X knockoffs framework. In addition, the concluding sentence is "Methods controlling FDR in TRN inference must either explicitly check the assumption of causal sufficiency, or avoid it." Based on this work and prior work on the field, it is reasonable to believe that the causal sufficiency assumption will never be met in real data. Should the recommendation be to firmly move away from TRN inference methods that make that assumption?

> **Our results are dependent on specific datasets, and we are reluctant to make sweeping generalizations about all (present and future) datasets and**

technologies. We now phrase this as a conditional: "If careful analysis of data from improved experimental methodologies continue to indicate lack of causal sufficiency, then the field should pursue analytical approaches that do not require causal sufficiency." (pg 20).

7) In the mouse SHARE-seq case study, Gaussian knockoffs cannot be constructed because there are only expression profiles from 57 clusters. However, this is a self-imposed limitation. The original dataset had expression profiles from 34,774 cells. Why reduce the dimensionality so greatly if it limits the type of knockoffs that are attainable?

We do this because constructing valid knockoffs is not a limiting factor in this analysis, but measurement error in TF expression is a limiting factor. By all our diagnostics, shrinkage-based covariance estimate leads to valid knockoffs on these data. However, measurement error in TF expression can cause intractable false positives because causal sufficiency is violated. We demonstrate this with new simulations (Supplementary file S4).

8) BETS (Lu 2021 doi:10.1371/journal.pcbi.1008223) is a TRN inference algorithm that emphasizes controlling FDR. It is based on permutations but uses time series expression data so it permutes the temporal profiles. That strategy is relevant related work, even if it only applies to special cases.

We agree and we appreciate this recommendation. We have cited it at appropriate points in the revised manuscript (pg 3,4,5,7,19).

9) Figure 3 has "naive" instead of "permuted" in the legend.

This has been corrected.

10) Supplementary file S2 contains the reference "Unable to find information for 13741696".

This has been corrected.

DATA & CODE AVAILABILTY

The supplemental repositories reflect a serious effort to make the research resources broadly available and the computational methodology transparent in a manner that follows best practices and goes beyond the typical manuscript in the field. Overall this is a major strength of the submission.

The resources are organized across one Zenodo data repository and eight GitHub repositories, which are outlined at https://github.com/ekernf01/knockoffs_paper. I looked through most of the GitHub repositories and tested code from the core repositories. Overall, I recommend archiving releases of the GitHub repositories through Zenodo, figshare, or Software Heritage. Most of the

repositories have a license file, but some are missing one.

https://doi.org/10.5281/zenodo.6573413
The data repository is well-organized with readme files describing original sources of the files. The only exception is share_seq.zip, which lacks such a readme.

https://github.com/ekernf01/knockoffs_ecoli
This repository documents in detail the E. coli experiments in the manuscripts and provides a Dockerfile with the goal of creating a container that can rerun the analyses. I was able to pull the Docker image but was unsure of its intended use. Running the R scripts directly produced errors:
root@930a93de5e24:/# cd knockoffs_ecoli/
root@930a93de5e24:/knockoffs_ecoli# Rscript dream5_ecoli_genets.R
Loading required package: corpcor
Loading required package: longitudinal
Loading required package: fdrtool
Error in ecoli_tf_expression %>% sweep(2, colMeans(ecoli_tf_expression), :
could not find function "%>%"
Execution halted
There are also instructions about needing to modify the datalake. Running the bash script also installs many dependencies that have seemingly been installed by the Dockerfile already. It also produced many errors:
./run_on_aws.sh: line 9: sudo: command not found
./run_on_aws.sh: line 20: wget: command not found
./run_on_aws.sh: line 21: wget: command not found
unzip: cannot find or open modern_ecoli.zip, modern_ecoli.zip.zip or modern_e
coli.zip.ZIP.
unzip: cannot find or open dream5.zip, dream5.zip.zip or dream5.zip.ZIP.
Error in eval(ei, envir) :
Datalake not found. Place it in '~/datalake' or modify `dream5_ecoli_setup.R
`.
Calls: source -> withVisible -> eval -> eval
Execution halted
AWS Access Key ID [None]:
AWS Secret Access Key [None]:
Default region name [None]:
Default output format [None]:
Partial credentials found in shared-credentials-file, missing: aws_secret_acce
ss_key
If the goal is to have this repository serve as extremely detailed documentation of the E. coli analyses that is not necessarily runnable, that is still valuable. However, the readme should set expectations. If the goal is to build a runnable container, more of the installation should be isolated to the Dockerfile so that the bash script focuses on downloading data and running the R code. That is not a requirement though, only a suggestion to pick one path or the other and

clarify the readme. A minor point is that it was unclear why the readme discusses Ubuntu 18.04 and the Dockerfile is built from ubuntu:20.04. The Dockerfile also runs the same Rscript command twice.

https://github.com/ekernf01/knockoffs_shareseq
This repository is similar to the E. coli repository but for the SHARE-seq data. I only examined the files and expect similar feedback applies.

https://github.com/ekernf01/knockoffs_BEELINE
https://github.com/ekernf01/knockoffs_boolode
These repositories document modifications to the BEELINE software and its simulator. I found this to be a helpful way to track those changes while retaining the original commit history from the BEELINE authors.

https://github.com/ekernf01/knockoffs_quick_demo
This repository provides example scripts for using the new rlookc package and reproducing additional figures from the manuscript. If the goal is to transparently document the analyses in the manuscript, it serves that purpose well. If the goal is to provide instructions for using the package on new transcriptional data, much more documentation would be needed.

https://github.com/ekernf01/rlookc
This is the core R package for model-X knockoffs. I was able to install and run the package on Linux with R version 4.1.0 but not Windows with R version 3.6.0 because RcppEigen would not compile successfully. I ran into problems running vignette_calibration.md initially because the rlookc and knockoff libraries are not loaded, but I ultimately was able to follow the vignette to completion. Automated testing of the vignettes would help catch small issues like this. Overall, the vignettes provide good instructions for someone looking to execute knockoff modeling. The vignettes that are still R files would be more beneficial in the Markdown format. However, the vignettes do not provide guidance to a user who would like to use the package for TRN inference or to model a gene expression dataset in the manner shown in the manuscript. Doing so could be pieced together based on the examples shared in the other repositories, but there are no clear instructions for a user with only basic abilities in R and a new expression dataset.

https://github.com/ekernf01/jlookc
https://github.com/ekernf01/pylookc
These are Julia and Python packages that load saved knockoffs from the R package. Due to the sparse documentation, it is unclear what the use case is. If a user has gone through the trouble of generating the knockoffs in R, it seems likely they would continue with downstream analysis in R. If the Julia and Python packages directly called rlookc (e.g. with rpy2) that would enable different workflows, but that is out of scope (which is fine).

I only tested the Python package and faced some problems. I installed it in a fresh conda environment
$ conda create -n pylookc python=3.9

$ conda activate pylookc
$ pip install -e .
Installing collected packages: peppercorn, pylookc
Not all required dependencies were installed, for example, numpy was missing. The LICENSE.txt also needs to be updated. There are also boilerplate files like package_data.dat and simple.py that could be removed.

We are profoundly grateful to the reviewer for this detailed, conscientious review of our codebase. We have:

- **Added "sources.txt" to the share-seq data describing data provenance.**
- **Added explicit instructions for how to use the Docker containers.**
- **Migrated all *E. coli* and multi-omics analyses to Ubuntu 20.04.**
- **Explained in more detail how our limited Python and Julia packages are meant to be used. The explanations are in the web documentation for those tools.**
- **Cleaned up the packaging for the Python code, updating LICENSE and removing stray boilerplate files.**
- **Created Github releases and minted Zenodo DOI's for all code.**

Reviewer #3:

Manuscript summary
In the manuscript, the authors tackle the problem of transcriptional regulatory network (TRN) inference, where existing methods are known to be plagued by false positives. The authors propose a new approach to this problem based on the model-X knockoffs methodology, and evaluate this approach on a combination of fully simulated, partially simulated, and real data. The conclusion is that model-X knockoffs generally improves the false discovery rate (FDR) in fully and partially simulated data compared to the permutation approach, but is unable to control the FDR on real data. The authors hypothesize that unmeasured confounders are to blame for this persistent FDR inflation.

Manuscript strengths
According to the authors, "empirical tests of advertised FDR rates have not been reported for any category of TRN inference method." If this is the case (I can't confirm this due to insufficient familiarity with the TRN literature), then this manuscript is an important step towards moving the field towards rigorous calibration checks. The various calibration checks proposed by the authors can serve as useful templates for future efforts to this end. In addition, the authors have proposed the novel leave-one-out-knockoff method (rlookc) and accompanying software, which may be of independent interest. Compared to the existing brute-force solution, the authors show that rlookc provides substantial computational acceleration. Finally, the authors evaluate their method on several real datasets, which enhances the robustness of their conclusions.

We thank the reviewer for this summary and the remarks on the strengths of our work.

Manuscript weaknesses

Narrow scope of benchmarking studies

The authors note in their introduction that "dozens of TRN inference methods have been invented" but that "empirical tests of advertised FDR rates have not been reported for any category of TRN inference method." These two facts, taken together, underscore the importance of benchmarking the FDR control of a broad range of existing methods. On the other hand, the only comparison the authors make throughout their manuscript is the permutation-based method. The only other methods compared to is GeneNets, which appears in one panel of one figure. The authors state that "popular TRN inference methods based on tree ensembles or mutual information can account for both nonlinearity and indirect relationships, but they do not provide finite-sample FDR control." What is the authors basis for this claim? If it is theoretical, then they should note that model-X knockoffs also do not provide provable finite-sample FDR control in the case when the joint distribution of the features is unknown a priori (the case considered in this manuscript). Therefore, it is worthwhile at least to compare the authors' methodological proposal also to "popular TRN inference methods based on tree ensembles or mutual information."

We agree that additional benchmarks would be informative. In the *E. coli* application, we now include benchmarks of knockoffs, permutation tests, GeneNet, the Gaussian Mirror, and BINCO (6,7,12) (pg13-16). Regarding information-theoretic and tree-ensemble TRN methods: typical and popular examples are ARACNe (13) and GENIE3 (14) respectively. We did not intend to claim that these methods produce inaccurate p-values or q-values. Rather, it is important to point out that these methods do not provide p-values or q-values at all, so they cannot be benchmarked in terms of expected versus observed FDR. For example, the original GENIE3 paper states "The question of the choice of an optimal confidence threshold, although important, will be left open." We have clarified this in the text (pg 4, Box 1).

Insufficient justification for the conclusion that "Model-X knockoffs control FDR in conditional independence testing"

The title of the first Results section is "In biochemical simulations, model-X knockoffs control FDR in TRN inference without using the true data-generating distribution." However, Figure 1B shows nontrivial FDR inflation for small target FDR levels (the important range) for the mixture knockoffs method, annotated with red ovals below. Saying that "model-X knockoffs approximately control FDR in TRN inference" would be more accurate.

We agree, and we have updated the subsection title to "Model-X knockoffs approximately control FDR in TRN inference from simulated data".

In the semi-synthetic experiments, the authors found decent FDR control for the knockoffs method. They stated that "Since FDR control with simulated targets does not translate to FDR control with real targets, this microarray dataset must not meet the causal sufficiency criterion." However, another explanation that cannot be ruled out is that the mechanism by which they simulated their targets (Algorithm 1) does not match that in the real data. In this simulation mechanism (which unfortunately is not described very clearly), it appears that each gene has exactly one regulator, and that each simulation involves just one target gene. If either of these are correct, then the simulation mechanism is not particularly realistic, and the fact that knockoffs does well at the conditional independence task on this particular semi-synthetic data does not imply that it will continue to do so with other simulated target mechanisms.

> **We regret that we did not describe the simulation mechanics clearly, and have now clarified exactly how target genes were simulated. We have also updated the simulations to include a more realistic mechanism. Each simulation includes 1,000 target genes, and each target gene has max(R, 1) regulators where R is Poisson with mean 2. The target gene's expression is set to 1 if all regulators exceed their mean expression, and it is set to 0 otherwise. Please see updated Algorithm 1 (pg10) and the Methods (pg26).**
>
> **More generally, we think that our simulation design is sufficient for our purposes because guarantees on model-X knockoffs depend only on the distribution of the regulators P(X), and are valid for any distribution of the target P(Y|X). Similar semi-synthetic data appear in several prior applications of knockoffs to genomic data: (15) supplement K.1; (16) section 6.1; (17) section 3.2; (18) section "Performance in simulations"; (19) section 3.2). Similar simulations with real X and synthetic Y have even been used to validate statistical methods that make explicit assumptions about P(Y|X), for example (20) section "Simulations to evaluate COLOC performance", (21) section 2.3 "Simulation data"; (22) section 4; and (23) section 4.**

Insufficient justification for the conclusion that "relative to gold standards, conditional independence testing via the knockoff filter outperforms permutation-based testing"
In Figure 2D (reproduced below), the permutation test outperforms the knockoffs-based tests for the "chip and M3Dknockout" setting for low target FDRs (the most important range); this is annotated in the figure with a red oval. The authors should acknowledge this, while noting that the knockoffs-based methods (especially glasso_1e-04) does outperform the permutation method in the "chip and RegulonDB knockout" setting.

> **We agree, and we have made changes in order to be more appropriately conservative in making claims of improved performance. This figure panel (now numbered Fig. 4A) was updated in response to errors noticed by reviewer 1, and we have also removed any observed FDR based on fewer than 10 testable discoveries. With only a few positives expected, these estimates will be highly volatile. We have re-written the description of these results (pg15).**

Additional points needing clarification or correction

1. The authors state that "With RNA only, static methods cannot infer directionality, so for RNA only, FDR was calculated with backwards edges counted as correct (Fig. 1B)." This sentence suggests that, if given the protein data, knockoffs can infer directionality. However, knockoffs is a conditional independence testing method rather than a causal inference method, so it cannot infer directionality even when the causal sufficiency assumption is satisfied.

> **In the examples where protein levels and transcription rates are revealed, protein is assumed to regulate transcription rates and not vice versa, based on the way BoolODE works. We have clarified the text on this point (pg7).**

2. The authors state that "Regarding causal sufficiency, this criterion was satisfied given RNA + protein data, which led to better FDR control." How do the authors conclude this? Is this based on knowledge of the data-generating model? In general, the causal sufficiency assumption is uncheckable based on data alone. Even informally, why do we expect that protein expression is a confounding variable, which when included in the analysis helps with causal sufficiency?

> **These data are fully simulated, and given how BoolODE works, we can assert this *a priori*. More information on BoolODE is given in the BEELINE paper (24), and we have clarified the text on this point (pg7).**

3. Should Figure 1A include a lack of causal sufficiency as an obstacle to FDR control for all three classes of methods?

> **Yes. We have revised this figure considerably in order to clarify our main contribution. It is now the graphical abstract (pg1).**

4. The authors use three values for the regularization parameter for the graphical lasso. A more standard approach would be to use cross-validation.

> **Similar to cross-validation, we begin with a grid of values and select the best from the data. However, cross-validation is optimized for prediction, not inference, and to check for validity of knockoff-based inferences, our evaluations based on the KNN exchangeability test and the simulated target genes are more relevant than predictive performance.**

5. In some figures, the expected FDR for some methods is bounded from below. The authors state that in cases of low power they were "unable to assess observed FDR for sets of hypotheses with low expected FDR." However, the FDR for an empty rejection set is defined by convention as zero. Perhaps the authors find that having FDRs of zero will make their plots cluttered or misleading; in this case, they should say so.

> **Indeed, we would rather omit the points than claim an observed FDR of 0 based on zero discoveries. In fact, when fewer than 10 testable hypotheses are returned below a given expected FDR, the observed FDR is highly uncertain and we believe**

it is best to leave the corresponding part of the plot blank. We have clarified this point in the legend of Figure 4 (formerly Fig. 2 D-F), pg 22.

6. The authors state that "To adjust for confounders, knockoffs were computed after appending columns (features) to the TF expression matrix containing either non-genetic perturbations or non-genetic perturbations and the top principal components (Fig. 2E)." Are the authors creating knockoffs also for these additional features as well? Note that creating knockoffs for features not used for testing unnecessarily induces higher correlations among features that are used for testing and their knockoffs.

We were unaware of this issue, and it is an interesting point. Unfortunately, the software we use does not allow for only some knockoffs to be constructed, and we defer this task to future work.

7. The authors should clean up their figures, including labeling panels with letters.

Thank you for pointing out this issue. We have revised all figures for improved clarity and consistent style. All panels are now labeled except in single-panel figures.

Bibliography

1. **Morgan D, Tjärnberg A, Nordling TEM, Sonnhammer ELL. A generalized framework for controlling FDR in gene regulatory network inference. Bioinformatics. 2019 Mar 15;35(6):1026–32.**
2. **Lu J, Dumitrascu B, McDowell IC, Jo B, Barrera A, Hong LK, et al. Causal network inference from gene transcriptional time-series response to glucocorticoids. PLoS Comput Biol. 2021 Jan 29;17(1):e1008223.**
3. **Chasman D, Iyer N, Fotuhi Siahpirani A, Estevez Silva M, Lippmann E, McIntosh B, et al. Inferring Regulatory Programs Governing Region Specificity of Neuroepithelial Stem Cells during Early Hindbrain and Spinal Cord Development. Cell Syst. 2019 Aug 28;9(2):167-186.e12.**
4. **Kimura S, Fukutomi R, Tokuhisa M, Okada M. Inference of Genetic Networks From Time-Series and Static Gene Expression Data: Combining a Random-Forest-Based Inference Method With Feature Selection Methods. Front Genet. 2020 Dec 15;11:595912.**
5. **Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol. 2007 Aug 6;1:37.**
6. **Schäfer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene**

association networks. Bioinformatics. 2005 Mar;21(6):754–64.

7. Li S, Hsu L, Peng J, Wang P. BOOTSTRAP INFERENCE FOR NETWORK CONSTRUCTION WITH AN APPLICATION TO A BREAST CANCER MICROARRAY STUDY. Ann Appl Stat. 2013 Mar 1;7(1):391–417.
8. Verny L, Sella N, Affeldt S, Singh PP, Isambert H. Learning causal networks with latent variables from multivariate information in genomic data. PLoS Comput Biol. 2017 Oct 2;13(10):e1005662.
9. Zhang J, Squires C, Greenewald K, Srivastava A, Shanmugam K, Uhler C. [2307.06250] Identifiability Guarantees for Causal Disentanglement from Soft Interventions. arXiv. 2023 Jul 12;
10. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019 Aug;29(8):1363–75.
11. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007 Jan;5(1):e8.
12. Xing X, Zhao Z, Liu JS. Controlling False Discovery Rate Using Gaussian Mirrors. arXiv. 2019;
13. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1(Suppl 1):S7.
14. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS ONE. 2010 Sep 28;5(9).
15. Candes E, Fan Y, Janson L, Lv J. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2018;80(3):551–77.
16. Sesia M, Sabatti C, Candès EJ. Gene hunting with hidden Markov model knockoffs. Biometrika. 2019 Mar;106(1):1–18.
17. Shen A, Fu H, He K, Jiang H. False discovery rate control in cancer biomarker selection using knockoffs. Cancers (Basel). 2019 May 29;11(6).
18. Sesia M, Katsevich E, Bates S, Candès E, Sabatti C. Multi-resolution localization of causal variants across the genome. Nat Commun. 2020 Feb 27;11(1):1093.
19. Li S, Ren Z, Sabatti C, Sesia M. Transfer Learning in Genome-Wide Association Studies with Knockoffs. Sankhya B. 2022 Nov 15;
20. Arvanitis M, Tayeb K, Strober BJ, Battle A. Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. Am J Hum Genet. 2022 Feb 3;109(2):223–39.
21. Liu L, Chandrashekar P, Zeng B, Sanderford MD, Kumar S, Gibson G. TreeMap: a

structured approach to fine mapping of eQTL variants. Bioinformatics. 2021 May 23;37(8):1125–34.

22. **Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016 May 15;32(10):1493–501.**
23. **Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Series B Stat Methodol. 2020 Dec;82(5):1273–300.**
24. **Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020 Feb;17(2):147–54.**