Causality - Perspective

This manuscript (permalink) was automatically generated from slobentanzer/causality_perspective_2023@175ec21 on November 24, 2023.

Authors

- **Sebastian Lobentanzer** ✉
  0000-0003-3399-6695 · slobentanzer · slobentanzer
  Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany

- **Julio Saez-Rodriguez** ✉
  0000-0002-8552-8976 · saezrodriguez · saezlab
  Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany

✉ — Correspondence possible via GitHub Issues or email to Sebastian Lobentanzer <sebastian.lobentanzer@gmail.com>, Julio Saez-Rodriguez <pub.saez@uni-heidelberg.de>.

Abstract

Introduction

Correlation is not causation. As simple as this widely agreed-upon statement may seem, scientifically defining causality and using it to drive our modern biomedical research is immensely challenging. Since being described by Aristotle approximately 2500 years ago, causal reasoning (CR) remained virtually unchanged [1] until significant formal and mathematical advancements in the last decades [2,3,4]. In parallel, biomedicine has made major leaps in the past century, in particular in the development of high-throughput and large-scale methods.

Randomised clinical trials show that, in a lower-dimensional context, we can achieve the high levels of confidence needed to satisfy the ethical requirements of modern medicine. However, translating this mode of reasoning into the high-dimensional space of modern omics is met with enormous challenges. The dramatically increased parameter space of models at the molecular level leads to problems in the performance of methods and the identifiability of results, as well as in model explainability.

With this perspective, we want to encourage and guide the use of CR to inform biomedical problems and vice versa. We will elaborate on three main points:

- biases and what they mean for CR, particularly in the context of biomedical data
- the role of prior knowledge in CR and how to translate prior knowledge into suitable biases
- the role of foundation models in molecular systems biology and their relationship to CR

Background

Causal Discovery and Inference

To ultimately explain biases, we must briefly touch on the background of CR. The field of CR distinguishes between *causal discovery* - the process of building hypotheses from data on how agents interact causally - and *causal inference* - the process of predicting how a specific situation will turn out given data and the causal relationships known about the system. In the scientific process from unknowingness to inference on a specific event, the process of causal discovery is more data-intensive than the process of inference, which almost always relies on the prior knowledge from the discovery stage. As a result, most inference mechanisms perform better when including prior knowledge at some point of the process. This has also been observed in biomedical research, for instance in the DREAM challenges [5].

Causal discovery is computationally and statistically very expensive because it needs to account for the variability in data generation while isolating generalisable relationships between single measured species (cite). For modern systems biology, this means that methods for causal discovery typically require large amounts of measurements. Highly parameterised models such as neural networks increase this requirement even further. As such, many regard causal discovery in molecular biomedicine as a scaling problem.

Causal inference, on the other hand, only requires sufficient measurements (replicates) to confidently account for the state of measured species in any condition (which can still be expensive, given the many technical and biological parameters that can influence molecular biology measurements). However, inference is also very sensitive to the completeness of the prior knowledge that is applied; most biomedical prior knowledge is far from complete. For instance, the function of more than 95% of all the known phosphorylation events that occur in human cells is currently unknown [6,7]. In contrast to causal discovery, scaling therefore plays a smaller role in causal inference; here, the main problem is incompleteness and identifying the “right” biases to apply.

The Ladder of Causality

Orthogonally to the distinction between causal discovery and inference, we can also distinguish between different levels of causality. The framework of the *ladder of causality* [] roughly distinguishes three types of CR in increasing order of power: observation, intervention, and counterfactuals. While the inferences we wish to make in biomedical research are often of the counterfactual type (e.g., “would this patient have survived if they had received this treatment?”), the data we have available is typically observational (e.g., “this patient received this treatment and survived”) and sometimes interventional (e.g., clinical trials or perturbation screening). To generate interventional or even counterfactual inferences

from observational data is a major challenge at least, and impossible at most, depending on the characteristics of the system under study.

There are approaches to delineate interventional inference from observational data, such as the 'natural experiments' framework []. However, these approaches are by their nature even more data-hungry than when using interventional data, as they necessarily discard information that is not relevant to the intervention. Therefore, in biomedical research, there has been a push towards generating large-scale interventional data, for instance through the use of CRISPR/Cas9 screens with single-cell resolution []. Current developments of CR in the biomedical field therefore mostly focus on these types of data.

Deduction and Induction

Lastly, in CR, we can also distinguish between *deductive* and *inductive* reasoning. This is where certain biases are pivotal to the effectiveness of the CR method. Deductive reasoning is the process of deriving a conclusion from a set of fixed and known premises. "All men are mortal, Socrates is a man, therefore Socrates is mortal" is a classic example of deductive reasoning. In biomedical research, this is typically the process of deriving a conclusion from a set of prior knowledge. For instance, knowing about the causes of stroke (which include high blood pressure) and the consequences of an angiotensin receptor-blocking drug (lowering blood pressure) allows us to deduce that the drug can be used to prevent stroke.

Inductive reasoning, on the other hand, is the process of deriving a conclusion from a set of observations; in the biomedical case, this is often measurements. For instance, we would conduct a clinical trial on the preventative effect of the angiotensin receptor-blocking drug on high-risk patients. Since we cannot feasibly test the drug on all potential patients, we instead test on a *representative* sample of the population. Given a statistically significant effect on stroke prevention, we can then perform the inductive step that the drug is generally effective at preventing stroke.

The main difference between deduction and induction is that the former is more reliable, but also more limited in scope, than the latter. In biomedical research, we often have to rely on inductive reasoning because we cannot feasibly test all hypotheses in a deductive manner. In consequence, the *inductive* biases we introduce into our models are a pivotal part of performing CR in biomedical research.

Bias

Meaning and examples of biases

Biases, generally, are systematic prejudices of a model towards certain outcomes. Humans make frequent use of biases to function in a complex world with limited cognitive resources. Our brain seems predisposed to doing causal inference, a skill which we learn and hone from a very early age (cite). In fact, we may be over-eager to deduce causality from observation (i.e., "jump to conclusions"), which is indicative of a strong inductive bias. A good *heuristic* is the application of a suitable bias to a problem, such that the solution can be considered acceptable despite limited resources.

In machine learning (ML), we can distinguish between two types of biases: useful and harmful biases. Harmful biases are common issues the technical process of training models; they include, for instance, sampling bias, selection bias, confirmation bias, overfitting, and underfitting []. While addressing harmful biases is a crucial part of ML, we will not discuss them further in this perspective.

Useful biases, on the other hand, are biases that are introduced into a model to improve its performance. They can be relatively implicit, such as the choice of algorithm, architecture, or regularisation; or explicit, such as the choice of prior knowledge and how it is used. In the context of CR, useful biases are those that improve the performance of the model in terms of its ability to draw correct causal conclusions. Since most models developed in biomedical research and the broader ML community are inductive models, one of the most discussed useful biases is *inductive bias*.

Why do we need biases?

In biomedical research, we operate in a space that is very constrained in the amount and quality of data we can collect. This is due to the high cost of experiments, the limited availability of samples, and the high dimensionality of the data. These issues, in combination with the naturally high variablity of biological measurements, lead to a relatively low signal-to-noise ratio of our observations. In addition, we are often trying to "climb" the ladder of causality with our CR approaches, which comes with additional data requirements. Lastly, we also lack a ground truth for most contexts in which we perform measurements. As a result, we need to introduce biases into our models to make the most of the data we have.

Some central questions then arise:

- How do we choose the right biases to introduce?
- How explicit should we be in introducing biases (i.e., should the model determine its own biases, or do we force them on the model)?
- How do we evaluate the biases we introduce?

Bias from prior knowledge

The human mind will be the gold standard for reasoning for the foreseeable future. However, human reasoning is limited by our sensory and mnemonic capacity; we cannot reason about high-dimensional data since we can neither perceive nor keep in memory thousands of agents at the same time. Hence, to make the most of our immense wealth of data, we must elevate our algorithms' reasoning capabilities. A sensible approach is to look to our reference model, the human, to try and transfer some of our capabilities to the in silico reasoner. In particular, to be successful in developing algorithms which we can trust to reason in the high-dimensional space of molecular biomedicine, they must use the available prior knowledge effectively. This can be achieved through converting the prior knowledge into suitable inductive biases [8,9,10].

In combining prior knowledge with reasoning algorithms, we need to remain mindful that the aim is not just to increase performance based on some metric, which is known as the „bitter lesson" [11]. It has been argued that the intrinsic complexity of real-world systems does not obviate, but rather necessitate the integration of human insight into our learning frameworks [12,13]. The impressive performance of recent deep learning models is only made possible by the introduction of attention or convolution as architectural inductive biases [14]. Considering the shortcomings of prior knowledge on biomedical molecular interactions as well as the constraints on available data, the question thus is not whether to include prior knowledge in our reasoning, but which knowledge, when, and how [13].

Prior knowledge

To be able to effectively use our knowledge in reasoning, we must be able to represent it robustly and in a way that is conducive to the reasoning task. Biomedical entities and relationships must be clearly defined and represented unambiguously. Additionally, the diversity in our tasks and knowledge sources requires a flexible representation that can be adapted to the task at hand. Knowledge representation frameworks can aid in walking the line between robust and transparent, reproducible knowledge representation on the one hand, and flexible, task-specific workflows on the other [15]. In addition, they can ground the biological entities that are subject to reasoning using domain-specific ontologies, which can be another useful source of bias. For instance, knowing the directionality of the central dogma of biology can be a useful bias in reasoning about gene expression.

Modelling on prior knowledge

Statistical, causal, and mechanistic models

Why modelling needs biases and how to introduce them

Benchmarks

Causality in foundation models

There has been an enormous spike of interest in attention-based neural network models, in large part due to the success of transformers in natural language processing and the commercial acclaim of ChatGPT. While the high performance of Large Language Models (LLMs) is based on a myriad architectural improvements, the introduction of attention as an architectural bias has been a major contributor to their success [14]. This has led to the development of attention-based molecular models (most commonly for gene expression), which can also be considered "GPT" models: Generative Pre-trained Transformers [16,17,18].

Attention as a learning mechanism enables the integration of non-local information in a flexible manner. In a molecular model that reasons about gene expression, such as Geneformer, attention allows the integration of chromosomally distant regulatory elements [17]. For example, the angiotensin converting enzyme (ACE), which is responsible of converting angiotensin I to angiotensin II, is causally responsible for the activation of the angiotensin receptor II (AGTR2). However, the ACE gene is located on chromosome 17, while the AGTR2 gene is located on chromosome X. Thus, to learn

the causal relationship between ACE and AGTR2 in a self-supervised manner based purely on observational data, the model must be able to integrate information from distant genomic regions.

The generalist capabilities of LLMs have led to the designation of "foundation models," a term introduced by a group of Standford ML researchers. Foundation models are models that achieve high performance by the combination of large amounts of data and model parameters, a generic architecture without specific biases, and self-supervised training. They can be fine-tuned for more specific tasks, because they are thought to derive generalisable representations and mechanisms by training on an amount of data large enough to encapsulate the complexity of real-world systems. While this designation is not too far off the mark for LLMs, it is not yet clear whether the same can be said for molecular models.

Recent molecular foundation model benchmarks highlight clear discrepancies between the "foundational" aspirations of the pre-trained models and the real-world evaluation of their performance [19,20]. - Details here

Attention - and large amounts of data - is all you need?

Given enough data to train on - and ample funds for compute - is attention "all you need" to induce reliable biases in your model? While there are doubts as regards the reasoning capabilities of our most advanced LLMs, GPT arguably "understands" language very well already, to the point where it can flawlessly communicate and also synthesise information [21]. This is what the term "foundation model" implies: the model has derived a generalisable representation of language, a tool that can be fine-tuned for a variety of language-related tasks. This behaviour is not possible without assuming some form of causality, even if it is not explicitly encoded in the model. In this light, what are the reasons to be skeptical about the capacity of molecular foundation models to understand the "grammar" of the cell?

For one, large transformer models are not explainable due to their large number of parameters and non-linearities. As such, there is no way to scrutinise their reasoning beyond the output they produce. What seems simple in the case of language models - the famous Turing test can be performed by any human with a basic understanding of language - is exceedingly difficult in the molecular space, where many causal relationships are yet unknown [21]. Yet the only way to scrutinise and subsequently improve the reasoning capabilities of a model is precisely this explicit validation of its predictions in an interpretable setting.

While the creation of explicit molecular models (e.g., logic, structural causal, or ODE-based models) and the self-supervised training of molecular foundation models are methodically very different, both result in a hypothesis on causal structure that can be formulated as a network. Theodoris et al. explore the attention layers of their Geneformer foundation model to explain the model's reasoning [17]. While some layers show clear patterns of attention, such as attending to highly connected or highly expressed genes, other layers are not as readily interpretable. Whether these complex layers reflect the true complexity of the underlying biology or are rather evidence for overfitting to the training data is not clear. One argument in favour of overfitting is the poor generalisation of the model in independent benchmarks [19,20]. To determine whether molecular foundation models indeed capture generalisable causal representations of biology, dedicated benchmarks are needed.

Cite some benchmarks, point out the most important aspects of benchmarking causally.

What is the mathematical relationship between explicit (e.g. ODE) and implicit (transformers) models? How could this be studied empirically?

Back-of-the-envelope data requirements for molecular foundation models (parameters, data points, comparison with LLMs).

Causal latent spaces

Latent encodings of explicit prior knowledge (GEARS)

Do 'causal latent spaces' exist, and how would we prove it?

How do we explore these latent spaces and use them for inference?

Stefan's comment: newer architectures (self-supervised) do not decode; how important is it for biological insights, particularly compared with scaling? Exploring and explaining the latent space...

- https://scholar.google.de/citations?view_op=view_citation&hl=de&user=soxv0s0AAAAJ&sortby=pubdate&citation_for_view=soxv0s0AAAAJ:2osOgNQ5qMEC

(decoder important)
- https://proceedings.neurips.cc/paper_files/paper/2022/hash/87213955efbe48b46586e37bf2f1fe5b-Abstract-Conference.html (decoder not important)

References

1. **The Organon, Or Logical Treatises, Of Aristole, Vol. 1 Of 2**
   Aristotle
   Forgotten Books (2015)
   ISBN: 9781330267608

2. **Causality**
   Judea Pearl
   Cambridge University Press (2009-09-14) https://doi.org/ggd72q
   DOI: 10.1017/cbo9780511803161

3. **Identification of Causal Effects Using Instrumental Variables**
   Joshua D Angrist, Guido W Imbens, Donald B Rubin
   Journal of the American Statistical Association (1996-06) https://doi.org/gdz4f4
   DOI: 10.1080/01621459.1996.10476902

4. **Myth and measurement: the new economics of the minimum wage**
   David E Card, Alan B Krueger
   Princeton University Press (2016)
   ISBN: 9781400880874

5. **Inferring causal molecular networks: empirical assessment through a community-based effort**
   Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, ... Sach Mukherjee
   Nature Methods (2016-02-22) https://doi.org/f3t7t4
   DOI: 10.1038/nmeth.3773 · PMID: 26901648 · PMCID: PMC4854847

6. **Illuminating the dark phosphoproteome**
   Elise J Needham, Benjamin L Parker, Timur Burykin, David E James, Sean J Humphrey
   Science Signaling (2019-01-22) https://doi.org/gf8c3h
   DOI: 10.1126/scisignal.aau8645 · PMID: 30670635

7. **The functional landscape of the human phosphoproteome**
   David Ochoa, Andrew F Jarnuczak, Cristina Viéitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A Kleefeldt, Anthony Hill, Luz Garcia-Alonso, Frank Stein, ... Pedro Beltrao
   Nature Biotechnology (2019-12-09) https://doi.org/ggd8n7
   DOI: 10.1038/s41587-019-0344-3 · PMID: 31819260 · PMCID: PMC7100915

8. **Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations**
   Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, Olivier Bachem
   arXiv (2018) https://doi.org/grx79c
   DOI: 10.48550/arxiv.1811.12359

9. **Toward Causal Representation Learning**
   Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio
   Proceedings of the IEEE (2021-05) https://doi.org/gjhqrh
   DOI: 10.1109/jproc.2021.3058954

10. **Beyond Predictions in Neural ODEs: Identification and Interventions**
    Hananeh Aliee, Fabian J Theis, Niki Kilbertus
    arXiv (2021) https://doi.org/gszw4d
    DOI: 10.48550/arxiv.2106.12430

11. **The Bitter Lesson** http://www.incompleteideas.net/IncIdeas/BitterLesson.html

12. **A Better Lesson – Rodney Brooks** (2019-03-19) https://rodneybrooks.com/a-better-lesson/

13. **Thread by @shimon8282: "Rich Sutton has a new blog post entitled “The Bitter Lesson” (incompleteideas.net/IncIdeas/Bitte...) that I strongly disagree with. In it, he [...]"**
    https://twitter.com/shimon8282
    https://threadreaderapp.com/thread/1106534178676506624.html

14. **Attention Is All You Need**

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv (2017) https://doi.org/gpnmtv
DOI: 10.48550/arxiv.1706.03762

15. **Democratizing knowledge representation with BioCypher**
Sebastian Lobentanzer, Patrick Aloy, Jan Baumbach, Balazs Bohar, Vincent J Carey, Pornpimol Charoentong, Katharina Danhauser, Tunca Doğan, Johann Dreo, Ian Dunham, … Julio Saez-Rodriguez
Nature Biotechnology (2023-06-19) https://doi.org/gszqjr
DOI: 10.1038/s41587-023-01848-y · PMID: 37337100

16. **Effective gene expression prediction from sequence by integrating long-range interactions**
Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, David R Kelley
Nature Methods (2021-10) https://doi.org/gm2wv4
DOI: 10.1038/s41592-021-01252-x · PMID: 34608324 · PMCID: PMC8490152

17. **Transfer learning enables predictions in network biology**
Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, XShirley Liu, Patrick T Ellinor
Nature (2023-05-31) https://doi.org/gr9x63
DOI: 10.1038/s41586-023-06139-9 · PMID: 37258680

18. **scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI**
Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Bo Wang
Cold Spring Harbor Laboratory (2023-05-01) https://doi.org/gshk6p
DOI: 10.1101/2023.04.30.538439

19. **Assessing the limits of zero-shot foundation models in single-cell biology**
Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, Alex X Lu
Cold Spring Harbor Laboratory (2023-10-17) https://doi.org/gszxk9
DOI: 10.1101/2023.10.16.561085

20. **A Deep Dive into Single-Cell RNA Sequencing Foundation Models**
Rebecca Boiarsky, Nalini Singh, Alejandro Buendia, Gad Getz, David Sontag
Cold Spring Harbor Laboratory (2023-10-23) https://doi.org/gszxmb
DOI: 10.1101/2023.10.19.563100

21. **ChatGPT broke the Turing test — the race is on for new ways to assess AI**
Celeste Biever
Nature (2023-07-25) https://doi.org/gskd92
DOI: 10.1038/d41586-023-02361-7 · PMID: 37491395