




# GRAPL: A computational library for nonparametric structural causal modelling, analysis and inference

Max A. Little<sup>1,2</sup>

<sup>1</sup> School of Computer Science, University of Birmingham, UK <sup>2</sup> MIT, Cambridge, MA, USA

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 29 March 2022

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

*Structural causal models* (SCMs) provide a probabilistic language for describing directed relationships between random variables. SCMs are widely used in science, engineering and statistical modelling to capture causal relationships between quantitative, measured phenomena in the real world. Two SCM formalisms, *directed acyclic graphs* (DAGs) and *acyclic directed mixed graphs* (ADMGs) have been extensively studied. In these formalisms, the conditions under which causal dependence between variables occurs is well understood. Furthermore, analytical techniques have been developed which allow manipulation of the model so as to perform *nonparametric causal adjustment*, that is, the isolation of desired causal relationships from the SCM. The GRAPL library described in this paper brings together the most important and useful of such algorithms in one convenient Python package. Using this library it is possible to represent, analyze and manipulate DAGs and ADMGs of arbitrary complexity.

## Statement of need

There exist a large number of techniques for statistical estimation of *causal* rather than merely *associational relationships* from data ([Hernan & Robins, 2020](#)). Under special causal relationships such as the existence of *observed confounders*, classical *causal inference* methods such as stratification, propensity scoring and IPW are widely used to *adjust* for these spurious associations. However, there are countless other physical scenarios where e.g. the confounders are *unobserved* or there are multiple, more intricate, cause-effect relationships between measured variables. In these circumstances, the classical techniques are not valid and more generalized *adjustment* methods are required ([Pearl, 2009](#)).

Where all measured variables are observed (so that the SCM is a DAG), it can be shown that it is always possible to compute the adjusted nonparametric cause-effect (*interventional*) distribution which can then be used to derive a suitable adjustment technique ([Bareinboim et al., 2020](#)). Doing these computations by hand for complex DAGs can be laborious and error-prone; this package automates the construction of such distributions, additionally enabling the determination of important DAG properties such as all *local Markov conditional dependence* relationships between variables ([Koller & Friedman, 2009](#)). These relationships encode for meaningful statistical dependence relationships which can, for instance, be tested against real data to validate any proposed DAG model.

When there are non-observed variables and the SCM is e.g. an ADMG (such as the existence of *unmeasured common causes* of pairs of variables) then the above guarantee no longer applies and in some cases it will not be possible, even in principle, to compute an interventional distribution in order to adjust for these spurious associations ([Richardson et al., 2012](#)). The computation of exactly when adjustment is possible, and the derivation of a nonparametric, interventional adjustment distribution when it is possible, is addressed by several recently developed, generic algorithms ("[Complete Identification Methods for the Causal Hierarchy](#),"

2008; Richardson et al., 2012). These computations are complex and too tedious and/or difficult to apply by hand, and require automation to be practical. However, the relevant algorithms are fairly complex and (to the author's knowledge) there are no publicly-available implementations written in widely used languages such as Python which are accessible to non-statisticians. (At the time of writing, the one possible exception to this, the R package `causaleffect`, does implement one of these algorithms, but does not provide the constituent, fine-grained analytical methods required to understand the topological causal structure of a given SCM).

The design of GRAPL was intended to bring together all the most general algorithms for handling all associational and causal distribution computations in DAGs and ADMGs, of arbitrary complexity, into one convenient package. The library is structured so as to expose all layers of computations such as the necessary fine-grained analysis of SCM graph topology and symbolic computations with arbitrary nonparametric marginalized conditional distributions on those graphs. Distributions are output in Latex format for convenient inclusion in publications, and SCMs can be represented using the built-in domain-specific language for specifying directed relationships between variables. It is chiefly aimed at researchers wanting to carry out causal inference analysis for a wide array of disciplines including machine learning, AI, data science, mathematical statistics and also in practical application areas such as bioinformatics or medical statistics, but can be used by any interested researcher wanting to understand the full topological, causal structural implications of specific SCMs, for both theoretical and practical reasons.

It is already in regular use as a tool for developing specialized causal inference methods, in particular *causal bootstrapping* techniques for general machine learning applications (Little & Badawy, 2019), and for constructing and analyzing SCMs representing the complex causal processes behind neurodegeneration (Sturchio & Espay, 2020).

## Acknowledgements

This work partially funded by NIH grant UR-Udall Center, award number P50 NS108676.

## References

- Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2020). *On Pearl's Hierarchy and the Foundations of Causal Inference*. ACM Books. <https://doi.org/10.1145/3501714.3501743>
- Complete identification methods for the causal hierarchy. (2008). *Journal of Machine Learning Research*, 1941–1979.
- Hernan, M. A., & Robins, J. M. (2020). (2020). *Causal Inference: What If*. Chapman; Hall/CRC.
- Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
- Little, M. A., & Badawy, R. (2019). *Causal Bootstrapping*. arXiv:1910.09648.
- Pearl, J. (2009). *Causality: Models, reasoning and inference*. Cambridge University Press. <https://doi.org/10.1017/S0266466603004109>
- Richardson, T. S., Evans, R. J., Robins, J. M., & Shpitser, I. (2012). Nested markov properties for acyclic directed mixed graphs. *UAI'12: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 13.
- Sturchio, A., & Espay, A. J. (2020). Phenotype-Agnostic Molecular Subtyping of Neurodegenerative Disorders: The Cincinnati Cohort Biomarker Program (CCBP). *Frontiers in Aging Neuroscience*, 12, 324. <https://doi.org/10.3389/fnagi.2020.553635>