M. Cyrus Maher1\*, Ryan D. Hernandez2,3,4

1Department of Epidemiology and Biostatistics, University of California, San Francisco, 185 Berry Street, Lobby 5, Suite 5700 San Francisco, CA 94107

2Department of Bioengineering and Therapeutic Sciences,

3Institute for Human Genetics,

4Institute for Quantitative Biosciences (QB3),

University of California, San Francisco

San Francisco, California, USA

To whom correspondence should be addressed.

E-mail: [cyrusmaher@gmail.com](mailto:cyrusmaher@gmail.com)

Address: 191 Castro St. Mountain View, CA 94041

Phone: (858) 249-7500

# Introduction

Establishing health-related causal relationships is a pivotal objective in biomedical research. Yet, the interdependent non-linearity of biological systems often impedes a thorough understanding of causal dynamics. Existing and forthcoming time series data will likely play an important role in taming this complexity. Traditional cross-sectional sampling have the limitation that they may average out non-linear patterns by pooling heterogeneous signals across subjects. Long time series from a single source, on the other hand, can allow us to understand dynamic and context-specific patterns of change.

We are just beginning to grasp the biomedical relevance of such a dynamical systems perspective. Consider for example the human microbiome. Dysbiosis in the gut has been implicated in, e.g. irritable bowel disease (IBD), obesity, diabetes, asthma, anxiety, and depression (Foster & McVey Neufeld, 2013; Arrieta et al., 2014). Meanwhile, recent studies on microbiome dynamics have found that the ecological makeup of the human microbiome is dynamic and individual-specific (Caporaso et al., 2011; Gajer et al., 2012; Fisher & Mehta, 2014). These dynamics may also interact with pathogens in interesting and therapeutically important ways. For example, there is evidence that ecological time series dynamics within the body may play a role in the progression from HIV to AIDS (Vujkovic-Cvijin *et al.*, 2013).

Complex, dynamically evolving interdependent systems such as the microbiome pose a significant challenge to existing time series methods. Several metrics exist for detecting static non-linear relationships. These include: spearman correlation (Spearman, 1904), distance correlation (Székely, Rizzo & Bakirov, 2007), and mutual information content (Reshef *et al.*, 2011). Causal relationships, on the other hand, can be examined using methods such as time-lagged regression (Granger, 1969), instrumental variables (Bowden & Turkington, 1990), and dynamical Bayesian networks (Wu *et al.*, 2009).

These causal methods are heavily model-based, however. As a result, they often falter when examining arbitrary non-linear or context-dependent relationships. Furthermore, the approaches mentioned above cannot adequately handle feedback loops, and they frequently generate both false positives and false negatives due to the influence of unmeasured confounders (Vanderweele & Arah, 2011). These are significant liabilities, particularly in biomedicine, where relationships are usually embedded within a broad network of incompletely observed interactions.

In this paper, we present the first publicly available, open source implementation of convergent cross mapping (CCM), a model-free approach to detecting dependencies and inferring causality in complex non-linear systems (even in the presence of feedback loops and unmeasured confounding; Sugihara *et al.*, 2012). CCM derives this power from explicitly capturing time-dependent dynamics through a technique known as state-space reconstruction (SSR). SSR has demonstrated utility for problems as diverse as wildlife management (Dixon, Milicich & Sugihara, 1999; Deyle et al., 2013) and cerebral autoregulation (Heskamp *et al.*, 2013). In practice, this analysis typically requires at least 25 data points, measured with sufficient density to capture system dynamics.

CCM leverages the fact that time series can be viewed as projections of higher-dimensional system dynamics (Sugihara *et al.*, 2012). As a logical result of this property, the time series of individual variables must contain information about the full causal system. Causal dynamics (conceptualized as the state space, or manifold) can then be reconstructed using individual time series. These reconstructions can be thought of as shadows of the true causal system. If the shadows reconstructed from distinct variables can be used to predict points from each other’s time series, we can infer that these variables provide views of the same causal system and so are causally related. Since these relationships are fundamentally asymmetric, this test can also establish the directionality of causation.

Further details on CCM are available in the supplementary material of this paper, as well as in that of Sugihara *et al.* 2012. Additional explanatory resources can also be accessed through the project website (<http://cyrusmaher.github.io/CauseMap.jl>).

# Materials and Methods

## CauseMap is fast

CauseMap implements CCM in Julia, a high-performance programming language designed for facile technical computing (Bezanson *et al.*, 2012). Via intelligent JIT (just in time) compilation, Julia offers much of the speed of low-level, low-productivity languages like C, while also providing the ease of use and platform independence of much slower high-level languages like Python, R, or Matlab.

At the core of CauseMap is the calculation of distances between a large number of manifold points in potentially high dimensional spaces. To optimize efficiency, CauseMap precomputes all necessary manifolds and pairwise distances using a state-of-the-art, BLAS-based protocol (for benchmarks, see: https://github.com/JuliaStats/Distance.jl).

## Tuning parameter values aid causal interpretation

Beyond the speed and comparative simplicity resulting from cutting-edge JIT compilation, CauseMap offers a number of conveniences and performance enhancements. For CCM, it is particularly important to optimize two tuning parameters: *E* and τp. *E* is related to the assumed dimensionality of the full causal system. This quantity is used to determine the dimensions of the reconstructed manifolds. τp, on the other hand, denotes the time delay of the causal effect of interest. By examining the optimal values of these two parameters, we may place bounds on the number of variables involved in the full causal system, gain insight into the timeframe of causal effects, and obtain a built-in sensitivity analysis of the final results.

## CauseMap optimizes and visualizes tuning parameters

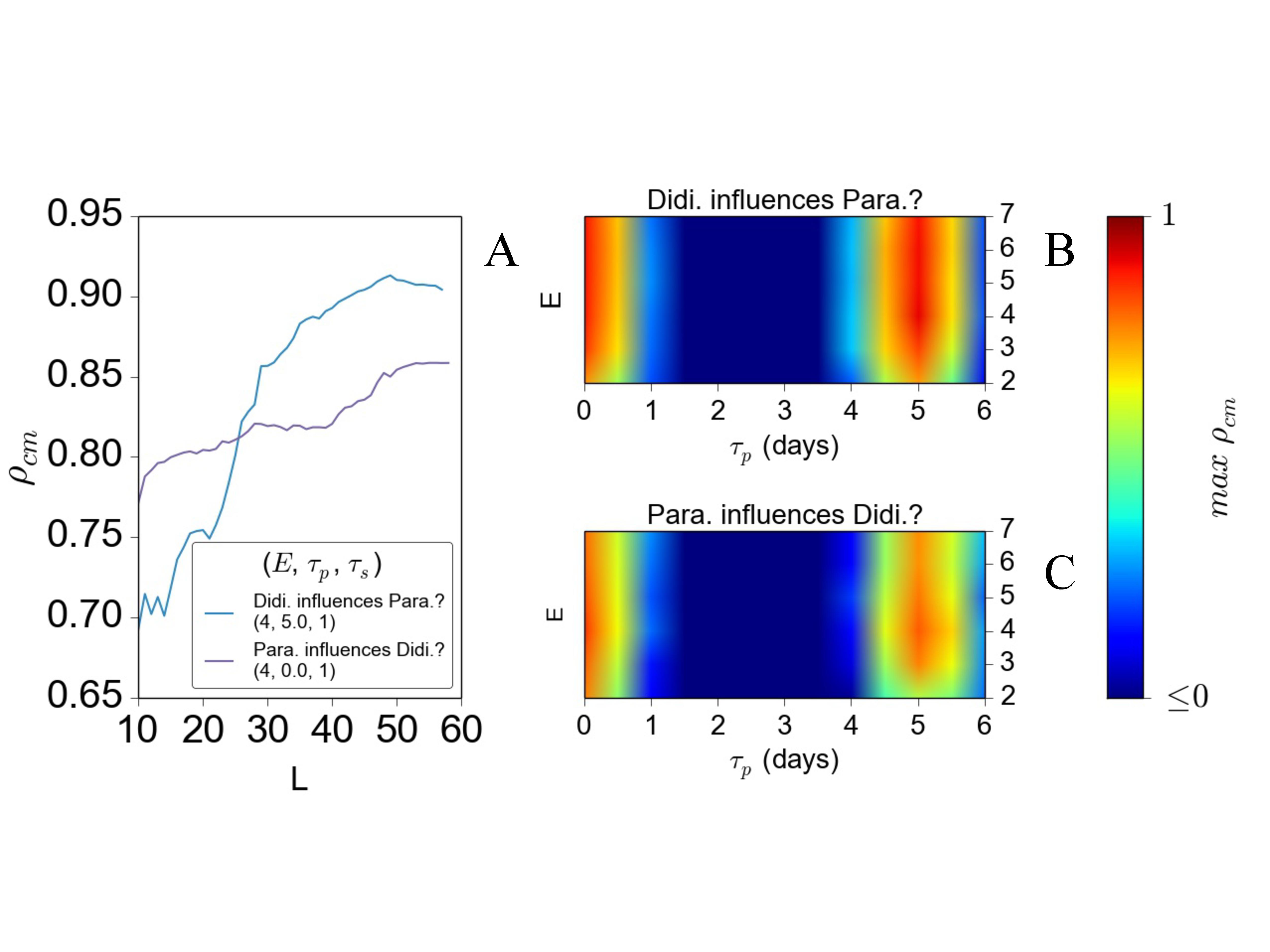
*E* and τp are then optimized by multiple iterations of cyclic coordinate descent (Bertsekas, 1999). Typically convergence of the cross map signal as a function of the time series length (*L*) alone is taken as the practical criterion for causality. However, measuring the dependence of this signal on *E* and τp is also useful for evaluating whether the result is suitably specific with respect to the assumed structure of the causal system. CauseMap therefore also includes a plotting function to visualize the dependence of the predictive skill (ρccm) on L, as well as on the joint values of E and τp.

## CauseMap is easy to use

Beyond the tuning parameters mentioned above, CCM requires one to specify a range of library sizes, as well as the window of time points for which cross mapping should be performed. Valid values for these parameters depend in turn on *E* and τp. To reduce complexity for the user, CauseMap calculates intelligent defaults for these parameters, while also offering the option of specifying them directly.

# Results and Discussion

To demonstrate CauseMap’s functionality and performance, we examined the predator-prey relationship between *Paramecium aurelia* and *Didinium nasutum* (Sugihara *et al.*, 2012). Observations were collected every 12 hours for 30 days, yielding a total of 60 data points. Plotted in Figure 1 is the CauseMap visualization of the dependence of predictive skill (ρccm) on *L, E,* and τp. In Figure 1A, we observe convergence in ρccm with respect to *L*, the number of data points used for prediction of held-out observations. This convergence is a practical criterion for causality and the source of the name *convergent* cross mapping. Figures 1B and 1C show the dependence of the max ρccm on E (proportional to the assumed dimensionality of the system), and the supposed time lag of the causal effect (τp). While the max ρccm is relatively insensitive to the assumed dimensionality, the best-performing τp values correspond to either immediate causal effects, or those delayed by five days. Note that τp=5 corresponds to the principal frequency of the *Paramecium aurelia* and *Didinium nasutum* time series, as determined by fourier transform analysis (see supplemental materials for further details).

**Fig. 1.** **An example visualization from CauseMap using abundances of *Paramecium aurelia* and *Didinium nasutum*** (see supplemental materials for more information on this system). A.) For optimal parameter values, the convergence of the cross-map correlation with library size. B-C.) The dependence of the maximum cross-map correlation on assumed dimensionality (measured by E) and the time lag of the causal effect (measured by τp). Note that the second maximum at τp=5 corresponds to the principal frequency of the *P. aurelia* and *D. nasutum* time series, as determined by fourier transform analysis.

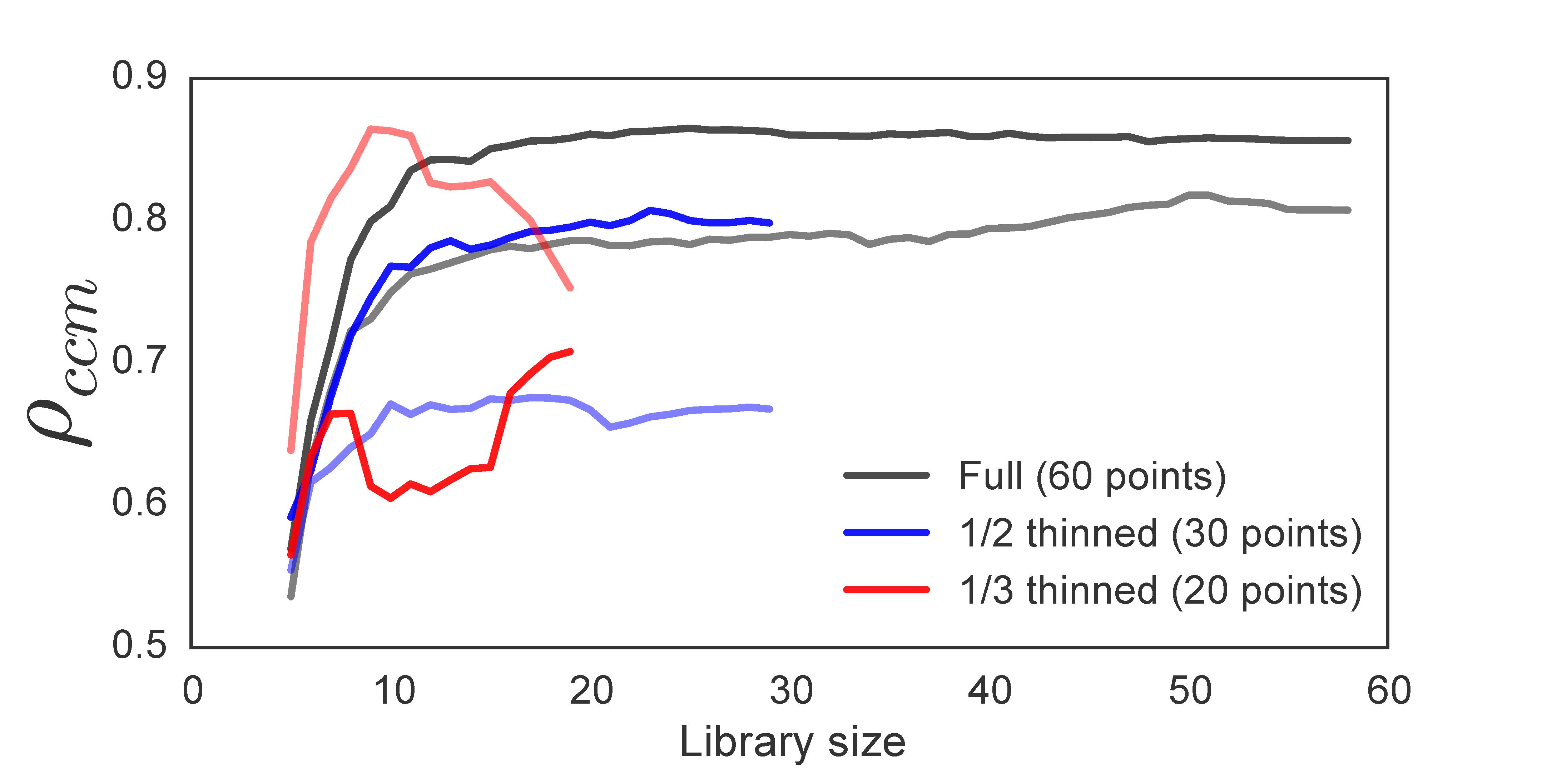
## Performance

Approximately 300 CCM evaluations were conducted to produce Figure 1. These calculations finished in less than 30 seconds on a single 2.6 GHz processor. Each of these evaluations involved the prediction of over 60,000 points, compiled across all sliding windows of libraries of varying lengths. At an average of 1.7 microseconds per prediction, this is a highly efficient implementation given the computational challenges.

## Dependence of predictive skill on time series length

CauseMap is designed to examine causal relationships in time series with 25 or more observations. In order to illustrate the effects of shorter time series, we thinned the *Paramedium-Didinium* data set by one-half and by one-third, yielding series of 30 and 20 observations, respectively. Figure 2 demonstrates the effect of this reduction on the convergence of predictive skill (ρccm). We see that the 1/2 thinned data set recapitulates the trends observed in the full series, including the relative magnitudes of ρccm between the mappings of *Didinium* to *Paramecium* and vice versa. The 1/3 thinned sample set, on the other hand, no longer demonstrates convergence. In addition, compared to the longer sets, it exhibits the opposite trend in relative predictive skillbetween the two mappings. Patterns in max ρccm versus E and τp are approximately conserved, however (fig. S1).

This example illustrates two phenomena. First, examining the dependence of predictive skill on E and τp gives us a more robust measure of causality than L alone. Second, CCM performance drops off sharply between 20 and 30 data points. This behavior is partially due to the fact that the predictive skill for a given library size is averaged across sliding windows of that size. As time series get shorter, there are fewer windows of appropriate size across which to average, so the estimate for predictive skill becomes much less reliable.



**Fig. 2. The effect of time series length on ρccm convergence.** Black, blue, and red lines illustrate **ρccm** for the full, 1/2 thinned, and 1/3 thinned datasets, respectively. For a given color, darker lines show ρccm for the test of whether *Didinium* abundance influences *Paramecium* abundance. Lighter lines examine the converse.

## Potential biomedical applications

Despite its requirement for relatively long time series (>25 observations), CauseMap has the advantage of requiring only a single time series for each variable. In dynamical systems with widely varying or context-specific behavior, this would allow researchers to draw conclusions that are tailored to, e.g. a given patient. Rather than acting on population averages, biomedical researchers would be free to fully personalize therapy to the unique biology and ecology of the patient. One example of this is in the treatment of microbiome dysbiosis. Imbalances in the microbiome have been implicated in, e.g. irritable bowel disease (IBD), obesity, diabetes, asthma, anxiety, and depression (Foster & McVey Neufeld, 2013; Arrieta et al., 2014). While fecal transplantation therapy is effective in treating specific types of dysbiosis (Khoruts & Weingarden, 2014), next generation therapeutics may offer a blend of purified strains, tailored to the gut ecology of the patient. We believe CauseMap has the potential to be a valuable tool for designing such breakthrough therapies.

Additional examples include understanding patient-to-patient variability in drug response using time series metabolomics, and examining the basis of e.g. influenza seasonality using global time series. We expect that such applications will continue to proliferate as the costs of data collection decrease over the coming years. For this reason, we believe it is vitally important that the biomedical research community have access to an efficient implementation of CCM that is user-friendly and available for immediate field testing.

## Planned future development

In future versions, we will include S-map calculations to evaluate the non-linearity of the causal system (Sugihara, 1994). We will also add a bootstrap-based procedure for library selection, as opposed to the current approach using sliding windows. This has been shown to reduce the effect of secular trends on the cross map correlation (Hao Ye, George Sugihara, *personal communication*). In addition, we will re-implement the plotting functionality in Julia, removing the requirements of Python and matplotlib for visualization. Finally, we will design Python and R wrappers for CauseMap functions so that our codebase can be easily leveraged from those environments as well. User suggestions will also be considered as we decide how best to develop the tool.

# Conclusions

CauseMap provides a fast, user-friendly implementation of CCM, a powerful new method for exploring dependencies and even establishing causality in complex, highly non-linear datasets with many unobserved variables. We believe that CCM holds a great deal of promise for a wide range of applications, including personalized microbiome therapy and metabolic dynamics analysis. As novel time series datasets continue to emerge, it is our hope that CauseMap will allow researchers to uncover interesting and biomedically actionable causal relationships using this next-generation time series method.

# Availability and Requirements

**Project name:** CauseMap

**Project home page:** http://cyrusmaher.github.io/CauseMap.jl/

**Operating system(s):** Platform independent

**Programming language:** Julia

**Other requirements:** Python and matplotlib (for graphing)

**License:** MIT

**Any restrictions to use by non-academics:** No

# List of abbreviations

Convergent cross mapping (CCM), State space reconstruction (SSR)

# Competing interests

The authors had no competing interest to declare.

# Author’s contributions

MCM and RDH conceived the project and drafted the manuscript. MCM implemented the algorithm and built the project website.

# Author’s Information

M. C. M. is a University of California, San Francisco (UCSF) graduate student with an emphasis in statistical computing. After graduation, he will be working as a Software Engineer for Human Longevity Inc., a San Diego-based biotechnology startup. R. D. H. is a Bioengineering & Therapeutics Sciences professor at UCSF. He is also the author of SFS\_CODE a popular program for flexible simulation of population genetic evolution.

# Acknowledgements

We would like to thank George Sugihara, Hao Ye, and Ethan Deyle for their invaluable help in understanding the core details of the CCM algorithm, and Lawrence Uricchio, Nicolas Strauli, and Raul Torres for comments on this manuscript.

References

Arrieta M-C, Stiemsma LT, Amenyogbe N, Brown EM, Finlay B. 2014. The Intestinal Microbiome in Early Life: Health and Disease. *Frontiers in Immunology* 5:427.

Bertsekas D. 1999. *Nonlinear Programming*. Athena Scientific; 2nd edition.

Bezanson J, Karpinski S, Shah VB, Edelman A. 2012. Julia: A Fast Dynamic Language for Technical Computing.

Bowden RJ, Turkington DA. 1990. *Instrumental Variables*. Cambridge University Press.

Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N et al. 2011. Moving pictures of the human microbiome. *Genome biology* 12:R50.

Deyle ER, Fogarty M, Hsieh C, Kaufman L, MacCall AD, Munch SB, Perretti CT, Ye H, Sugihara G. 2013. Predicting climate effects on Pacific sardine. *Proceedings of the National Academy of Sciences of the United States of America* 110:6430–5.

Dixon PA, Milicich MJ, Sugihara G. 1999. Episodic Fluctuations in Larval Supply. *Science* 283:1528–1530.

Fisher CK, Mehta P. 2014. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression.

Foster JA, McVey Neufeld K-A. 2013. Gut-brain axis: how the microbiome influences anxiety and depression. *Trends in neurosciences* 36:305–12.

Gajer P, Brotman RM, Bai G, Sakamoto J, Schütte UME, Zhong X, Koenig SSK, Fu L, Ma ZS, Zhou X et al. 2012. Temporal dynamics of the human vaginal microbiota. *Science translational medicine* 4:132ra52.

Granger CWJ. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods Title. *Econometrica* 37:424–438.

Heskamp L, Meel-van den Abeelen A, Katsogridakis E, Panerai R, Simpson D, Lagro J, Claassen J. 2013. Convergent cross mapping: a promising technique for future cerebral autoregulation estimation. *CEREBROVASCULAR DISEASES* 35:15–16.

Khoruts A, Weingarden AR. 2014. Emergence of fecal microbiota transplantation as an approach to repair disrupted microbial gut ecology. *Immunology letters*.

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. 2011. Detecting novel associations in large data sets. *Science (New York, N.Y.)* 334:1518–24.

Spearman C. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15:72–101.

Sugihara G. 1994. Nonlinear Forecasting for the Classification of Natural Time Series. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 348:477–495.

Sugihara G, May R, Ye H, Hsieh C, Deyle E, Fogarty M, Munch S. 2012. Detecting causality in complex ecosystems. *Science (New York, N.Y.)* 338:496–500.

Székely GJ, Rizzo ML, Bakirov NK. 2007. Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics* 35:2769–2794.

Vanderweele TJ, Arah OA. 2011. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)* 22:42–52.

Vujkovic-Cvijin I, Dunham RM, Iwai S, Maher MC, Albright RG, Broadhurst MJ, Hernandez RD, Lederman MM, Huang Y, Somsouk M et al. 2013. Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science translational medicine* 5:193ra91.

Wu J, Sinfield JL, Buchanan-Wollaston V, Feng J. 2009. Impact of environmental inputs on reverse-engineering approach to network structures. *BMC systems biology* 3:113.