Prototyping CRISP: A Causal Relation and Inference Search Platform applied to Colorectal Cancer Data

1st Samuel Budd§
BioMedIA, Dept. of Computing
Imperial College London
London, UK
sfb113@imperial.ac.uk

2nd Arno Blaas§
Dept. of Engineering Science
University of Oxford
Oxford, UK
arno@robots.ox.ac.uk

3rd Adrienne Hoarfrost§
Dept. of Marine and Coastal Sciences
Rutgers University
New Brunswick, NJ, USA
adrienne.hoarfrost@rutgers.edu

4th Kia Khezeli§
Center for Individualized Medicine
Department of Surgery, Mayo Clinic
Rochester, USA
khezeli.kia@mayo.edu

5th Krittika D'Silva
Dept. of Computer Science
University of Cambridge
Cambridge, UK
krittika.dsilva@cl.cam.ac.uk

6th Frank Soboczenski
SPHES
King's College London
London, UK
frank.soboczenksi@kcl.ac.uk

7th Graham Mackintosh
Advanced Super Computing Division
NASA
USA
graham.mackintosh@nasa.gov

8th Nicholas Chia
Center for Individualized Medicine
Department of Surgery, Mayo Clinic
Rochester, USA
chia.nicholas@mayo.edu

9th John Kalantari
Center for Individualized Medicine
Department of Surgery, Mayo Clinic
Rochester, USA
john.kalantari@mayo.edu

***Abstract*—We introduce CRISP, a Causal Research and Inference Search Platform. It is designed to assist biological and medical research by applying a variety of causal discovery methods to heterogeneous and high-dimensional observational data. CRISP aims to identify a small set of input variables which are most likely to have a causal effect on a target variable. The output of CRISP, thus, highlights the most promising candidates for further targeted research. We illustrate the utility of CRISP with a case study in oncology, using a multi-omic colorectal cancer data set to identify causal drivers differentiating two subtypes of colorectal cancer.**

***Index Terms*—Causal Discovery, causal inference, colorectal cancer**

I. Introduction

Cancer is a heterogeneous disease with many factors contributing to its development and progression. A lack of knowledge about cancer aetiology in addition to the inter-tumor and intra-tumor heterogeneity observed among tumors have posed significant challenges in the discovery of new therapeutics and preventative countermeasures. As a consequence, cancer still takes more than 600,000 lives every year in the US alone. [1]. While there has been a continued effort to apply machine learning to cancer research, ranging from melanoma detection [2] to survival prediction [3], the application of such methods has been mostly restricted to identifying associations in the data. However, in order to fully understand, treat, and prevent cancer, a causal understanding of the mechanisms driving the disease is necessary.

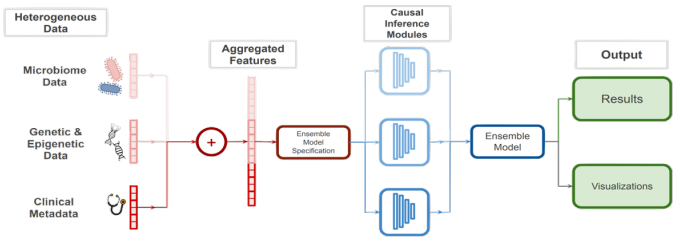

Fig. 1. Schematic diagram of CRISP applied to multi-omic colorectal cancer data

In the causal inference literature, a number of methods have been developed and adopted for observational data (cf. [4] and references therein). However, their application to biological data poses significant challenges. First, each of these methods requires specific data assumptions, many of which are hard to verify in practice (e.g., non-confoundedness). Second, only some of them can cope computationally with high-dimensional data.

We aim to overcome these challenges in order to enable the use of causal discovery methods in cancer research, and more

Frontier Development Lab, Mayo Clinic, NASA, SETI

§Equal contribution

broadly facilitate the adoption and application of machine learning assisted causal research in biomedicine. As a first step towards this goal, we introduce CRISP (Causal Research and Inference Search Platform). A summary of our contributions is listed below:

- We combine six different causal discovery methods into a single causal discovery platform that together provide a more holistic picture of causal relationships in biological data.
- We illustrate the benefits of the resulting platform by applying it to a heterogeneous, multi-omic colorectal cancer dataset.
- We enable the application of causal discovery methods to high-dimensional data by combining expert guidance and automated dimensionality reduction.

II. Related Work

Applications of machine learning to cancer research are numerous. Prominent recent examples include [5] who classify brain tumors better than human experts using random forest based methods, [6] who beat human experts in breast cancer diagnosis using deep learning models, as well as [7] who predict lung cancer risk using deep learning models. A good review of earlier work is provided in [8]. Unlike all these impressive works, our work goes beyond prediction and classification by using causal inference to generate an understanding of the causal drivers of cancer.

In causal machine learning methods, we are only aware of one work applied to cancer research ([9]), which uses inverse reinforcement learning to understand the development and progression of colon cancer. Our approach is orthogonal to this work in that we use a different set of causal inference methods to achieve a similar goal.

The causal inference methods we use are described in Methods. They are mainly based on [10], [11], [12], and [13], as well as earlier works contributing to them. Compared to these works, we apply their methods to cancer data that is high-dimensional and low in sample size.

III. Methods

CRISP combines six different methods for causal discovery, described below. We provide further details on how these are integrated as well as describe the automated dimensionality reduction routine that enables CRISP to run on high dimensional data.

A. The components of CRISP

The causal and/or invariant prediction methods used in CRISP are: linear and non-linear Invariant Causal Prediction (ICP) [13], [14], linear and non-linear Invariant Risk Minimisation (IRM) [10], Average Treatment Effect (ATE) [12], and the Deconfounder (DCF) [11].

Both ICP and IRM are methods that search for input variables that invariantly predict a target variable across different environments. Environments are defined as subsets of the data that do not share the same underlying data generating distribution, but are expected to share the same causal relationships. While the non-linear versions can naturally cope better with potential non-linear relationships in the data, the linear versions are less affected by the combination of low sample size and high dimensionality.

Methods for estimating the ATE seek to understand the strength of the effect of an input variable ('treatment') on a target variable. To use such methods for causal discovery, we sequentially estimate the ATE for all input variables and include those with significant ATE in CRISP.

ICP, IRM, and ATE all assume that there are no unobserved confounding variables to the data. The DCF is an approach designed to cope with such variables. DCF first fits a factor model to the observed data, and then augments this data by the mean latent values of the fitted model as surrogate confounders. Based on this augmented data set, an unconfounded model (e.g., a linear model) is trained to predict the outcome.

1) Invariant Causal Prediction: Invariant Causal Prediction (ICP) is a causal inference approach that searches for a combination of features in an input feature set that invariantly predicts a target variable across environments. That is, a model is fit to predict a target variable from the combination of features being tested; the data is split into environment groups according to an 'environment' feature supplied by the user; and a statistical test is applied to evaluate whether the residuals of the predictions are invariant across environments, returning a p-value. It can be applied to both linear ([14]) and nonlinear ([13]) model settings, with the difference being in the statistical test that is applied to the residuals.

In a linear setting, a two-sided t-test is applied to test whether the mean of the residuals between the two environment groups are equal; and a cumulative distribution function of the ratio of the variances between the two environments is used to test whether the variance of the residuals across environments are equal. Finally, the minimum of these two p-values is accepted, and a bonferroni corrected final p-value is returned.

In the nonlinear setting, nonparametric tests are used to test the invariance of residuals across environments. A wilcoxon rank sums test is applied to evaluate that residuals from different environments are drawn from the same distribution. A Levene test is applied to evaluate whether residuals across environments have equal variances. As in linear ICP, the bonferroni-corrected minimum of the two p-values returned by these tests is accepted.

A p-value threshold is supplied by the user, and any combination of features that is above the p-value threshold is considered to have equal residuals across environments and is accepted. From the accepted combinations of features that invariantly predict the target variable, the intersection of these sets is accepted as the final feature set. If no intersection exists, a 'defining set' is determined (as in [13]) as the subset of features such that each accepted set has at least one feature that is contained in every defining set. This

defining set approach is more flexible to highly correlated features, which can be difficult to distinguish in practice. Environment variables are chosen by the user, and the only requirement is that the environmental variable should not directly affect the target variable. Additionally, while every combination set of every possible size from $1..N$, where N is the number of features in a dataset, is tested in the original ICP implementations, this can result in a very high number of combinations which takes considerable time to compute. CRISP enables users to specify the maximum number of features in a set to test.

2) Invariant Risk Minimization: Invariant Risk Minimization (IRM) is a method that is designed to discover invariant relationships from empirical data. In particular, IRM is a learning method that seeks to identify classifiers that are optimal across different environments. More precisely, IRM is expressed as a constrained optimization of the following form

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \quad \sum_{e\in\mathcal{E}_{\text{tr}}} R^e\left(w\circ\Phi\right) \tag{1}$$

$$\text{subject to} \quad w\in\underset{\tilde{w}:\mathcal{H}\to\mathcal{Y}}{\operatorname{argmin}} R^e\left(\tilde{w}\circ\Phi\right),\ \forall e\in\mathcal{E}_{\text{tr}}, \tag{2}$$

where $\mathcal{E}_{\text{tr}}$ is the set of training environments, and R^e denotes the risk under environment e. Notice that removing the constraint in optimization problem (2) recovers the classical empirical risk minimization problem. Incorporating this constraint results in a bi-leveled optimization problem, which is computationally challenging. Arjovsky et al. [10] proposes a variant of IRM that is more practical.

$$\min_{\Phi:\mathcal{X}\to\mathcal{H}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e\left(w_0\circ\Phi\right)+\lambda\left\|\nabla_{w|w=w_0}R^e\left(w\circ\Phi\right)\right\|^2 \tag{3}$$

where w_0 is a user specified vector with a user specified dimension (e.g., a scalar).

3) Average Treatment Effect for causal discovery: Methods for estimating the average treatment effect often seek to understand the strength of a causal effect on an outcome. While this only aims to understand the relationship between one specific variable and the outcome, the CRISP framework extends this approach so that it can be applied in causal discovery. Specifically, the average treatment effect is iteratively estimated for all potential causes, including any significant causal effect found into our framework. The implementation follows Gelman ([12], Chapter 9 and 10) and uses a linear regression model to estimate the effect and significance of a cause on the outcome after first binarizing the cause under inspection. Binarizing is usually done by setting all values which are greater than 0 to 1 and leaving the remaining ones at 0.

4) Deconfounder: All of the methods presented so far share the implicit assumption that all confounding variables are part of the observed data. One recent approach to cope with situations when this is not the case is the Deconfounder (DCF). This method was developed by Wang & Blei ([11]) as a causal approach that can cope with unobserved confounding variables that have an effect on at least two observed variables. While its potential to accurately identify causal effects has been questioned in a theoretical discussion ([15], [16]), it has exhibited empirical potential in applications where unobserved confounders are present. Since the presence of such unobserved confounders is very much likely in complex biological data, we have decided to include DCF into CRISP. The main elements of the DCF are as follows.

1) *An assignment model:* in order to account for unobserved multi-cause confounding variables, DCF fits a probabilistic factor model (e.g. probabilistic principal components analysis (PPCA), mixture models, variational autoencoders) to the observed causes in order to infer substitute confounders from the data as given by the latent space distribution of the model conditioned on the observed causes. In the current implementation of the framework, the probabilistic PCA is used.
2) *An outcome model:* Augmenting the observed data by the inferred conditional mean of the substitute confounders given the observed data points, DCF fits an outcome model that links the augmented data to the observed outcomes. In theory, different models can be used here, including models that establish environment invariance. However, in practice, it is often found that a simple linear model works well ([11]) and hence a linear outcome model is currently implemented.

B. Combining the Components

Each of the methods described above relies on different assumptions (e.g. regarding confoundedness) and their outputs can be sensitive to the training routine. For these two reasons, combining these approaches may improve their individual performances. We first compute the causal impact attributed to each input variable $k \in \{1,\ldots,d\}$ for each individual method $m \in \{1,\ldots,M\}$ via sensitivity analysis. For linear models such as linear ICP, linear IRM, ATE, and our implementation of DCF, this is straightforward as it is given by the corresponding linear coefficient. For the non-linear models, we evaluate this impact as the difference between the outputs of the method derived from setting the corresponding input variable to its minimum and maximum values. We then combine the obtained sensitivity coefficients $s_m(k)$ of each individual method into a single value that indicates the causal potential $\mathrm{CP}(k)$. This metric reflects the likelihood assigned by CRISP that input variable k has a causal effect on the target variable (as well as the direction of that effect), as follows.

We sort the d input variables for each method m by largest absolute values $|s_m(k)|$, and calculate the fraction of methods for which an input variable k is in the top T as p_k[1]. For each method, the predictive accuracy on unseen test data is

[1]We suggest selecting T in relation to total number of input variables d

calculated as a_m. Based on these, we define the heuristic for the combined input variable causal potential $\mathrm{CP}(k)$ as

$$\mathrm{CP}(k) = p_k \cdot \frac{1}{M} \sum_{m}^{M} a_m \cdot \frac{s_m(k)}{\max_k |s_m(k)|}. \quad (4)$$

Thus, for an input variable k, $\mathrm{CP}(k) = 1$, if and only if every method in CRISP achieves predictive accuracy of 1, and every method in CRISP selects that input variable with its highest absolute sensitivity coefficient. In this way we up-weight the causal potential of an input variable if it is selected by multiple methods, but down-weight the contribution of a method if its predictive accuracy is low.

C. Automated dimensionality reduction

A common problem in oncological data is that it is typically very high-dimensional (order of millions to billions), but low in sample size (order of tens to hundreds). This makes expert guided dimensionality reduction necessary, but its extent may not be sufficient to apply some causal inference methods. For example, ICP has a computational complexity that is exponential in the number of input variables. Therefore, it is vital that CRISP has an additional automated dimensionality reduction subroutine which does not omit crucial causal input variables. This is an open challenge, but CRISP currently uses non-linear IRM to select the most important 100 input variables before running the full causal analysis.

IV. Data

To illustrate the efficacy of CRISP for causal inference in medicine, we use a multi-omic colorectal cancer (CRC) data set provided by *Anonymous*. It consists of tumor samples collected via surgical resection from 100 CRC patients. It consists of heterogeneous data types: data describing genomic and methylation variants of cancer cells relative to surrounding noncancerous cells, gut microbiome community composition from the area surrounding the tumor, and clinical metadata describing the patient. The anonymized clinical metadata contains the age group, BMI group, sex, location of the tumor in the colon, and cancer stage of the patient at the time the tumor sample was taken. Age and BMI were grouped into categories from 1-5 to further preserve patient privacy. Genomic and methylation variant data included the location of the variant, the original and altered sequences (in the case of genomic variants), and the type of region affected (e.g. 'intergenic region', 'disruptive inframe deletion', 'intron variant'). For privacy reasons, the data was provided in standard variant call format (vcf) describing the nature and location of the difference between the tumor genome to the healthy cell genome, without including the full original genome sequence of the patient. Microbiome data consisted of abundance counts of the 8471 most abundant operational taxonomic units (OTUs), based on 16S ribosomal DNA profiling of the microbial community surrounding the tumor. These OTUs were further classified into predicted taxonomies which ranged in specificity from family to species level depending on the OTU. The classification of the cancer type into Mismatch Repair Deficient (DMMR) or Mismatch Repair Proficient (PMMR) was also provided, based on an immunohistochemistry (IHC) test. IHC distinguishes the two types based on tumor sample loss of the protein product of the affected mismatch repair (MMR) genes [17].

V. Experiment - CRC subtype prediction

In this experiment, we employ CRISP to extract causal variables from the CRC data set and compare this result to analyses based on correlation or other non-causal measures of association. The aim is to determine causal drivers that explain the difference in the sub-type of CRC - PMMR vs. DMMR - from a combination of clinical metadata and somatic mutation data. To this end, we formed environments based on age group of patients, as we suspect the true causal drivers of CRC subtype should not vary significantly by age, whereas other factors that are only correlated with it might differ. Figure 2

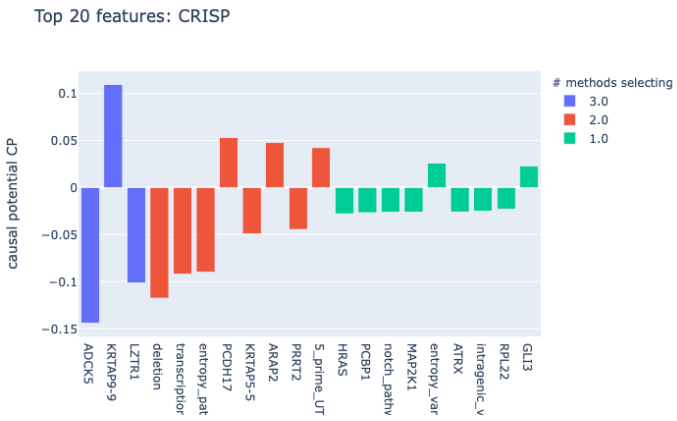

Fig. 2. Top 20 most explaining variables selected by CRISP as likely to be causal.

shows the features that CRISP selects as most likely to be causal. Notably, three explaining variables have been selected by half the methods we include in CRISP – ADCK5 and LZTR1 gene modifications by non-linear IRM, non-linear ICP and DCF, and KRTAP9-9 gene modifications by non-linear IRM, non-linear ICP and linear IRM. All three of these are modifications of protein coding genes, which is interesting in light of the nature of the IHC test (measuring loss of protein product). Furthermore, LZTR1 is a known tumor suppressor gene [18]. Yet, none of these three genes have been previously linked to a specific CRC subtype, thus this result calls for further investigation in clinical research. Importantly, none of these three input variables were selected by all six methods, highlighting the utility of CRISP in identifying potential causal variables that analyses based on only one method may miss. Lastly, in Figure 3 it can be seen that these three variables are neither found by a correlation analysis nor by inspecting the feature importances of a strongly predictive, but non-causal model (random forest with test accuracy 1). This highlights the necessity of causal inference methods when looking for causal explanations.

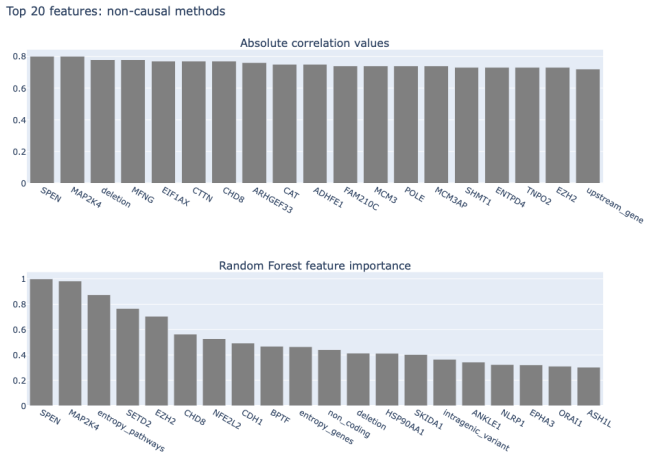

Fig. 3. Top 20 features for non-causal methods. Values are normed by largest feature importance for random forest.

VI. Future Work

To assess the efficacy of CRISP and the aggregation heuristic in 4 more thoroughly, we aim to develop a set of three comprehensive validation protocols. The first procedure will use synthetic data with a biologically appropriate causal structure to validate and calibrate the default settings of CRISP. The second procedure will use real-world data with known causal structure to verify CRISP's performance in a real-world setting. The final procedure will provide theoretical statistical guarantees for the results provided by CRISP, which will be vital for building trust in safety critical domains such as healthcare.

VII. Conclusion

We have presented CRISP, a platform of causal discovery methods applicable to high-dimensional biological data with low sample size. CRISP is an ideal tool to test biological hypotheses of causal drivers and to identify causal mechanisms with high potential for further investigation. Future validation work seeks to further improve the performance and utility of the platform to guide clinical prevention and treatment in practice.

Acknowledgment

The authors would like to thank the Frontier Development Lab, NASA, and the SETI Institute for their continuous support during this work. The authors are also grateful for the computational resources provided by Google Cloud. The authors would like to thank Yarin Gal and Miguel Marthino for their guidance and thank Marina Walther-Antonio, Nenad Tomasev, Martin Arjovsky, David Lopez-Paz, Ishaan Gulrajani, Juan Gamella Martin, Christina Heinze-Deml and Brock Sishc for their guiding feedback throughout the project.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian, "Melanoma detection by analysis of clinical images using convolutional neural network," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 1373–1376.

[3] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, "Risk classification of cancer survival using ann with gene expression data from multiple laboratories," *Computers in biology and medicine*, vol. 48, pp. 1–7, 2014.

[4] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference*. The MIT Press, 2017.

[5] D. Capper, D. T. Jones, M. Sill, V. Hovestadt, D. Schrimpf, D. Sturm, C. Koelsche, F. Sahm, L. Chavez, D. E. Reuss *et al.*, "Dna methylation-based classification of central nervous system tumours," *Nature*, vol. 555, no. 7697, pp. 469–474, 2018.

[6] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[7] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[8] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[9] J. Kalantari, H. Nelson, and N. Chia, "The unreasonable effectiveness of inverse reinforcement learning in advancing cancer research," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 437–445.

[10] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[11] Y. Wang and D. M. Blei, "The blessings of multiple causes," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1574–1596, 2019.

[12] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[13] C. Heinze-Deml, J. Peters, and N. Meinshausen, "Invariant causal prediction for nonlinear models," *Journal of Causal Inference*, vol. 6, no. 2, 2018.

[14] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference using invariant prediction: identification and confidence intervals," 2015.

[15] A. D'Amour, "Comment: Reflections on the deconfounder," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1597–1601, 2019.

[16] E. L. Ogburn, I. Shpitser, and E. J. T. Tchetgen, "Counterexamples to" the blessings of multiple causes" by wang and blei," *arXiv preprint arXiv:2001.06555*, 2020.

[17] F. A. Sinicrope and Z. J. Yang, "Prognostic and predictive impact of dna mismatch repair in the management of colorectal cancer," *Future oncology*, vol. 7, no. 3, pp. 467–474, 2011.

[18] T. Abe, I. Umeki, S.-i. Kanno, S.-i. Inoue, T. Niihori, and Y. Aoki, "Lztr1 facilitates polyubiquitination and degradation of ras-gtpases," *Cell Death & Differentiation*, pp. 1–13, 2019.