



Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium

Sjoerd Viktor Beentjes ^{1,2,*} and Ava Khamseh ^{3,4,5,†}

¹*Hausdorff Center for Mathematics, Universität Bonn, Endenicher Allee 60, D-53115 Bonn, Germany*

²*School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom*

³*MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom*

⁴*School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom*

⁵*Higgs Centre for Theoretical Physics, School of Physics & Astronomy, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom*



(Received 20 June 2020; accepted 21 October 2020; published 24 November 2020)

The problem of inferring pairwise and higher-order interactions in complex systems involving large numbers of interacting variables, from observational data, is fundamental to many fields. Known to the statistical physics community as the inverse problem, it has become accessible in recent years due to real and simulated big data being generated. Current approaches to the inverse problem rely on parametric assumptions, physical approximations, e.g., mean-field theory, and ignoring higher-order interactions which may lead to biased or incorrect estimates. We bypass these shortcomings using a cross-disciplinary approach and demonstrate that none of these assumptions and approximations are necessary: We introduce a universal, model-independent, and fundamentally unbiased estimator of all-order symmetric interactions, via the nonparametric framework of targeted learning, a subfield of mathematical statistics. Due to its universality, our definition is readily applicable to *any* system at equilibrium with binary and categorical variables, be it magnetic spins, nodes in a neural network, or protein networks in biology. Our approach is targeted, not requiring fitting unnecessary parameters. Instead, it expends all data on estimating interactions, hence substantially increasing accuracy. We demonstrate the generality of our technique both analytically and numerically on (i) the two-dimensional Ising model, (ii) an Ising-type model with four-point interactions, (iii) the restricted Boltzmann machine, and (iv) simulated individual-level human DNA variants and representative traits. The latter demonstrates the applicability of this approach to discover epistatic interactions causal of disease in population biomedicine.

DOI: [10.1103/PhysRevE.102.053314](https://doi.org/10.1103/PhysRevE.102.053314)

I. INTRODUCTION

Starting from microscopic laws of Nature, the aim of statistical physics is to provide a macroscopic description of Nature by deriving observable quantities from the underlying laws. In the *inverse problem*, the starting point is observations for which the underlying microscopic properties, such as interactions within the constituents of the system of interest, are unknown and to be inferred. Taking the Ising model of binary magnetic spins as an example, the goal of the forward problem is to obtain observables such as magnetization, energy, and correlation, given the Hamiltonian with its parameters. Conversely, the goal of the inverse problem is to derive unknown interactions within spins directly from data.

In recent years, the inverse problems are often motivated by challenges in “big data” biology due to modern high-throughput sequencing experiments and large-scale patient databases. There is a rich literature for inverse problems with the aim of inferring model parameters describing a system, e.g., via a Hamiltonian, from observational data (see, e.g., [1] and the references therein). Most of these methods rely

on making assumptions about the parametric form of the Hamiltonian, which may not accurately reflect the true distribution of the data. For instance, a misspecified parametric form often results in biases in the estimation of the quantities of interest when sample sizes grow without the variance in the estimation decreasing sufficiently fast. Furthermore, in most real-world settings such as interactions in biomedical data, there is no heuristic, let alone a theory, suggesting that the effects of higher-order interactions are negligible and can be ignored without consequence. Most methods in the literature simply truncate the problem by allowing for at most pairwise interactions [1–5]. This in turn results in biased estimates, even for two-point interactions.

The aim of this work is to introduce a universal, unbiased, and targeted framework in which symmetric two-point and higher-order interactions can be estimated from *any* discrete data set. We propose a *model-independent* definition of n -point interaction among binary and categorical random variables. In contrast to earlier approaches to the inverse problems in the literature, our definition is fully *nonparametric*: we make no assumptions on the parametric form of the joint or marginal probability distributions of the random variables. Moreover, in contrast to other approaches, which consider pairwise interactions only, ours can access higher-order interactions [1–5]. We note that the nonparametric approach

*sjoerd.beentjes@ed.ac.uk

†ava.khamseh@ed.ac.uk

in Ref. [6], although pairwise, does incorporate dynamical interactions. From a theoretical perspective, our definition benefits from the following three properties: (i) it is unbiased by construction and hence converges to the ground truth in the infinite data limit, (ii) it provides a natural, model-independent interpretation of higher-order interactions, and (iii) it reduces to well-known intuitive notions of interaction in parametric statistical physics models described by a Hamiltonian. From a computational point of view, our definition of n -point interaction may be directly estimated from observational data by simply taking suitable combinations of expectation values. The variance on the resulting estimate solely depends on how deeply relevant states are sampled, and it can be substantially improved when (conditional) independence between variables is known or derived. In most practical situations where the Markovian condition is assumed, e.g., for causal identifiability [7], (conditional) independence may be derived using causal structure learning algorithms such as [8–10].

Our nonparametric definition of n -point interactions among binary random variables fits in the targeted learning framework of [11], a subfield of mathematical statistics. Targeted learning is a probabilistic framework to estimate (causal) quantities of interest directly from a data set \mathcal{O} , without the need to successfully estimate the true (but unknown, and often unknowable) joint probability distribution p_0 that generated \mathcal{O} , or to expend data on estimating parameters θ of a potentially misspecified parametric model p_θ . Crucially, the framework requires a *model-independent* definition of the (causal) quantity of interest α , known as the *target parameter*, as a functional of any candidate probability distribution p , not in terms of a parameter of a parametric ansatz. This eliminates bias due to the choice of model while safeguarding the interpretation of α as a meaningful statistical quantity revealing true knowledge about the ground truth p_0 . Once the target parameter is established, all statistical power is used for its estimation. The targeted learning framework has already been successfully applied in biomedicine and epidemiological studies [11].

This paper is structured as follows. We discuss the nonparametric formulation of interactions using the targeted learning framework in Sec. II, for the case of binary and categorical variables. We propose two definitions of interaction, namely, *additive* and *multiplicative*, and illustrate their relation. For a given data set and application, one choice may be more intuitive than the other, but the information they hold is equivalent. The additive formulation in Sec. II B applies to scenarios where the subject expert takes one of the variables in the system as the “outcome” variable and is interested in estimating the effect of the interaction among other variables on this outcome. The multiplicative formulation in Sec. II C treats the variables on the same footing, and instead considers their effect (via interactions) on the energy function, and hence the joint probability distribution. The former is more used in biomedical applications when a treatment-outcome relationship is set out at the beginning, whereas the latter is more relevant for statistical physics and, e.g., molecular networks in biology.

Next, we provide a general formula for extracting n -point interactions and their interpretation directly from data. We conclude Sec. II by discussing how establishing conditional

independence among variables, e.g., via the nonparametric χ -squared test or more sophisticated state-of-the-art algorithms such as [8,9], leads to improved estimates of the n -point interaction.

As a first result, we provide a concrete biological example in Sec. III, based on interactions among DNA variants (epistasis) contributing to trait or disease, with data generated using a linear model. We demonstrate analytically and numerically that the targeted learning estimator obtains the correct ground truth interaction, even though it is *entirely agnostic* to both the data generating process and its linearity. This simplified example is used to guide the reader through the theoretical concepts introduced in Sec. II.

To demonstrate universal applicability of our estimator, in Sec. IV B, we consider a more complex Hamiltonian, namely, that of the restricted Boltzmann machine (RBM), and analytically obtain its all-order couplings without the need for an asymptotic expansion and resummation as originally employed in [12]. In Sec. IV C, we consider the two-dimensional (2D) Ising model and show how the *same* estimator is able to predict two-point interactions among nearest and non-nearest neighbor spin pairs, at various temperatures and lattice sizes. Moreover, it correctly predicts that three- and four-point interactions vanish. We compare our estimations to predictions from an RBM, on data generated from the 2D Ising model. We limit our comparisons to the RBM as, unlike other parametric methods, it does not truncate higher-order interactions and hence does not bias lower-order interactions.

Finally, in Sec. V, we generate data from a Hamiltonian with self-, two-, three-, and four-point interactions and show that our targeted learning estimator accurately predicts higher-order interactions. We present numerical results at various temperatures. This indicates that the targeted learning (TL) estimator can be applied to obtain higher-order interactions in the case of biological networks, such as biomarker and gene expression networks. For instance, this method is applicable to modern biomedical data sets, such as large-scale patient databases, e.g., UKBiobank, containing half a million patient samples [13], or high-throughput sequencing experiments, e.g., the 1.3×10^6 single-cell experiment by 10X genomics [14] and the Human Cell Atlas project, so far containing 4.5×10^6 cells [15].

II. NONPARAMETRIC FORMULATION OF INTERACTION

A. Targeted learning

Let \mathcal{O} be a data set of n observations \mathcal{O}_i generated by an experiment with random variable O , and let p_0 denote its probability distribution $O \sim p_0$. The fundamental goal in probabilistic modeling is to obtain an estimate \bar{p} of p_0 given the data \mathcal{O} . With \bar{p} in hand, a relevant quantity α concerning the data set \mathcal{O} can then be estimated, such as a moment, an interaction coefficient, or a (causal) effect.

In typical situations, however, given the data \mathcal{O} the ground truth p_0 is completely out of reach due to, e.g., a small sample size n as compared to the dimensionality of the data. To remedy this, a parametric form p_θ of \bar{p} may be proposed, and the data may be used to fit unknown parameters θ , but this often leads to an incorrect ansatz for the parametric model due

to bias. Alternatively, one may use model selection based on the data \mathcal{O} , but will subsequently suffer from overconfidence in reporting the estimate $\bar{\alpha}$ of the quantity of interest α .

Targeted learning [11] is a probabilistic framework to estimate (causal) quantities of interest directly, without the need to successfully estimate p_0 or to expend data on estimating parameters θ of a (misspecified) parametric model p_θ . As such, it avoids the above pitfalls of the estimation problem. Targeted learning consists of the following steps:

(1) Define the *statistical model* \mathcal{M} : this is the, in general infinite dimensional, space of candidate probability distributions,

$$\mathcal{M} = \{p \mid p \text{ a probability compatible with } \mathcal{O}\},$$

based on the data \mathcal{O} . By compatibility, we mean that the statistical model accommodates for *a priori* knowledge regarding the data and how it is generated. For example, if \mathcal{O} is generated by n binary random variables, then \mathcal{M} only contains $p = p(T_1, \dots, T_n)$ with T_i binary variables. Similarly, if the expectation value $\mathbb{E}(T_i)$ of a variable is known to be positive, or if one or more variables are known to be (conditionally) independent, this true knowledge can be incorporated. Finally, the statistical model contains the true probability distribution $p_0 \in \mathcal{M}$ by definition.

(2) Define the *target mapping* $\Phi : \mathcal{M} \rightarrow \mathbb{R}^d$ that expresses the quantity of interest α as a function of the distribution p . In particular, $\alpha_0 = \Phi(p_0)$ is the ground truth for α . For example, Φ could be a (conditional) expectation value over some or all of the variables. As another example, suppose that \mathcal{O} is generated by a random variable $O = (Y, T, W)$ where Y is a continuous outcome, T is a binary random variable which we will call treatment, and W is a covariate. The treatment effect

$$\Phi(p) = \mathbb{E}_W[\mathbb{E}(Y \mid T = 1, W) - \mathbb{E}(Y \mid T = 0, W)]$$

is another example of a target parameter, often used in epidemiological studies to estimate the causal effect of a drug or treatment T on health outcome Y while correcting for confounding effects due to the covariate W .

(3) Apply statistical methods to obtain an estimate $\bar{\alpha}$ of the target parameter. We indicate a method for obtaining improved estimates of n -point interaction in Sec. II F, but otherwise refer the reader to [11].

There are a number of important remarks to be made regarding the targeted learning paradigm. First of all, the *definition* of the quantity of interest α and its subsequent *estimation* are two separate steps. On the one hand, the quantity of interest is no longer a parameter in a potentially misspecified parametric model p_θ , but is associated to a candidate probability distribution p via the map Φ as $\Phi(p)$; thus, the quantity of interest needs to be expressed *nonparametrically* as a function of p forcing one to reevaluate the interest of said quantity. On the other hand, the method of estimation may be chosen independently from either model or target parameter. Second, by expressing the quantity of interest α as a target parameter $\Phi(p)$ one avoids introducing bias by making an incorrect parametric ansatz p_θ while safeguarding the interpretation of α as a meaningful statistical quantity revealing true knowledge about the ground truth p_0 . And, third, due to bias every misspecified parametric model will not converge to

the ground truth as sample size increases and variance shrinks. Thus, a nonparametric definition of a quantity of interest is essential to make full use of big data.

In this paper, we apply the framework of targeted learning to our quantity of interest, n -point interaction, and illustrate its application on data generated from various models.

B. Additive interaction

Consider a random variable $O = (Y, T_1, \dots, T_r, W)$ where Y is a discrete or continuous outcome, the T_i are binary random variables causally leading to the outcome Y , and W is a covariate. In this section, we wish to causally infer the effect of the interaction of the treatment variables T_i on the outcome Y , for simplicity having already corrected for confounding effects W . In other words, we implicitly take expectation values over strata of the covariate W . For example, we abbreviate

$$\mathbb{E}(Y \mid T_1 = 1) = \mathbb{E}_W[\mathbb{E}(Y \mid T_1 = 1, W)], \quad (1)$$

where \mathbb{E} denotes the expectation value over $Y \mid T_1 = 1$, and \mathbb{E}_W denotes the expectation value over W . Note, however, that all definitions and results hold in the more general case of a fixed value $W = w$ of the covariate.

First of all, we define the statistical model, incorporating all *a priori* knowledge, as in Sec. II A:

$$\mathcal{M} = \{p(Y, T_1, T_2, \dots, T_r, W) \mid Y \text{ continuous, } T_i \text{ binary, } W \text{ a covariate}\}.$$

Before defining the target parameter, we introduce some notation that will be used throughout the paper. If a subset T_{i_1}, \dots, T_{i_n} of the variables T_1, \dots, T_r is specified, then we write \underline{T} for all of the remaining variables. For example, $\mathbb{E}(T_1 \mid T_3 = 1, \underline{T} = 0)$ denotes the conditional expectation value of T_1 , given $T_3 = 1$ and $\underline{T} = 0$, meaning $T_2 = T_4 = T_5 = \dots = T_r = 0$. We abbreviate $(T_i, T_j) = (a, b)$ to $T_{ij} = (a, b)$.

In biomedicine and epidemiological studies, a particular quantity of interest to be estimated is the causal effect of a treatment on an outcome, the *average treatment effect*, e.g., the effect of a drug on health. We express our additive notion of interaction with notation compatible with the existing literature [7,11,16]. The *average treatment effect* (ATE) of T_i on Y is given by

$$\text{ATE}_{T_i}(Y) = \mathbb{E}(Y \mid T_i = 1) - \mathbb{E}(Y \mid T_i = 0). \quad (2)$$

This expression is the first order derivative with respect to T_i evaluated at $T_i = 0$ of the function $T_i \mapsto \mathbb{E}(Y \mid T_i)$. Indeed, for a function f of a binary variable T we have $\partial_T f = f(1) - f(0)$.

Next, given two binary variables T_i, T_j encoding two different treatments, we obtain the ATE of treatment T_i on Y and the ATE of treatment T_j on Y . A natural question is as follows: *How do these treatments interact?* In words, how does applying treatment T_i affect the effect of treatment T_j on Y , and vice versa? In order to isolate the effects of T_i and T_j on Y , the other treatments are not applied, i.e., we condition on $\underline{T} = 0$. We now define the first target mapping $\Phi_{i,j}^a$, which is our nonparametric additive formulation of two-point interaction between binary random variables. The *additive interaction* $I_{i,j}^a$ between the binary variables T_i and T_j is given by the difference of the effect of changing $T_i : 0 \rightarrow 1$ on Y

given $T_j = 1$, and the effect of changing $T_i : 0 \rightarrow 1$ on Y given $T_j = 0$, i.e.,

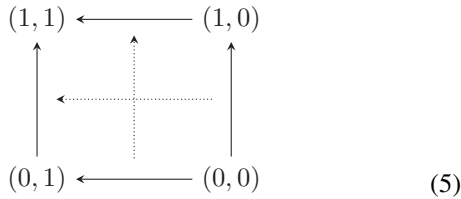
$$\begin{aligned} \mathcal{M} \ni p \mapsto \Phi_{i,j}^a(p) := I_{i,j}^a &= [\mathbb{E}(Y | T_{ij} = (1, 1), \underline{T} = 0) \\ &\quad - \mathbb{E}(Y | T_{ij} = (0, 1), \underline{T} = 0)] \\ &\quad - [\mathbb{E}(Y | T_{ij} = (1, 0), \underline{T} = 0) \\ &\quad - \mathbb{E}(Y | T_{ij} = (0, 0), \underline{T} = 0)]. \end{aligned} \quad (3)$$

Note that interaction is a difference of ATEs, i.e., $I_{i,j}^a = \text{ATE}_{T_i}(Y | T_j = 1, \underline{T} = 0) - \text{ATE}_{T_i}(Y | T_j = 0, \underline{T} = 0)$. Thus, the interaction $I_{i,j}^a$ is the change of effect of T_i on Y when changing T_j , conditioned on $\underline{T} = 0$. This change of effect may be expressed as the (symmetric) double derivative with respect to T_i and T_j , and so $I_{1,2}^a$ is also the change of effect of T_j on Y when changing T_i . Formally, this reads as

$$I_{i,j}^a = I_{j,i}^a, \quad (4)$$

as one readily deduces from Eq. (3). Indeed, given a function $f : \{0, 1\}^2 \rightarrow \mathbb{R}$ of two binary variables x and y , $\partial_x \partial_y f = \partial_y \partial_x f$.

Although numerically, the effect of T_i on the effect of T_j on Y is the same as the effect of T_j on the effect of T_i on Y , only one direction might admit a sensible interpretation. This is similar to the causal interpretation of the set of equations $Y = mX + b$ or $X = m'Y + b'$ that is provided by a directed acyclic graph (DAG) [7] and is not captured by the equation alone. In contrast, note that the *sign* of the interaction is uniquely determined since a *direction* is specified: it is the effect on Y of changing T_i from 0 to 1, not from 1 to 0, that we compare to the effect on Y of changing T_j from 0 to 1. Both the symmetry and the sign of $I_{i,j}^a$ are illustrated in the following diagram:



We introduce the shorthand $A(t_i, t_j) = \mathbb{E}(Y | T_{ij} = (t_i, t_j), \underline{T} = 0)$ where $t_i, t_j \in \{0, 1\}$. In the diagram, vertex (t_i, t_j) represents the expected outcome $A(t_i, t_j)$. An arrow represents the average treatment effect of the variable of which the value changes, where the sign is dictated by “target minus source.” For example, the left vertical arrow encodes the average treatment effect of $T_i : 0 \rightarrow 1$ on Y given $T_j = 1$, i.e.,

$$A(1, 1) - A(0, 1) = \text{ATE}_{T_i}(Y | T_j = 1, \underline{T} = 0). \quad (6)$$

Finally, either dotted arrow encodes the interaction between the effects of T_i and T_j on the outcome Y , together with its inherent symmetry. Indeed, via the sign convention “target

minus source,” the diagram yields relations

$$\begin{aligned} I_{i,j}^a &= \text{ATE}_{T_i}(Y | T_j = 1, \underline{T} = 0) - \text{ATE}_{T_i}(Y | T_j = 0, \underline{T} = 0), \\ I_{j,i}^a &= \text{ATE}_{T_j}(Y | T_i = 1, \underline{T} = 0) - \text{ATE}_{T_j}(Y | T_i = 0, \underline{T} = 0), \end{aligned}$$

where the first line is encoded by the horizontal arrow and the second line by the vertical arrow.

Next, we define the additive n -point interaction on the outcome Y . Whereas the two-point interaction is a difference of two ATEs, hence a sum of $2^2 = 4$ expectation values, the three-point interaction involves $2^3 = 8$ such terms and, more generally, the n -point interaction involves 2^n terms. We introduce notation in order to state the formula of a general n -point interaction.

Consider a subset $K = \{i_1, \dots, i_{\ell(K)}\} \subset \{1, \dots, r\}$ of the indices for the treatment variables T_1, \dots, T_r in the random variable O . Here, in general, given a further subset $J \subset K$ we denote its number of elements by $\ell(J)$. We write $e_J^{(\ell(K))}$ for the $\ell(K)$ -tuple of elements,

$$e_J^{(\ell(K))} = (e_{i_1}, \dots, e_{i_{\ell(K)}}), \quad (7)$$

where e_{i_j} equals 1 if $i_j \in J$ and 0 if $i_j \notin J$. For example, if $J = \{2, 7\} \subset \{1, 2, 4, 5, 7\} = K$, then

$$e_J^{(\ell(K))} = e_J^{(5)} = (0, 1, 0, 0, 1). \quad (8)$$

Finally, we write $T_K = (T_{i_1}, \dots, T_{i_{\ell(K)}})$ where $i_j \in K$ for all $1 \leq j \leq \ell(K)$. Continuing the previous example, we have $\ell(K) = 5$ and $\ell(J) = 2$. The five-point interaction between the variables $T_K = (T_1, T_2, T_4, T_5, T_7)$ is a sum of $2^5 = 32$ terms, and it will involve the expectation value

$$\begin{aligned} \mathbb{E}(Y | T_K = e_J^{(5)}, \underline{T} = 0), \\ \mathbb{E}(Y | T_1, T_2, T_4, T_5, T_7) = (0, 1, 0, 0, 1), \underline{T} = 0). \end{aligned} \quad (9)$$

The next target mapping, Φ_{i_1, \dots, i_n}^a , is our nonparametric additive formulation of n -point interaction.

Definition 1. Let $K = \{i_1, \dots, i_n\} \subset \{1, \dots, r\}$ be a subset of indices. The additive n -point interaction among the effects of the binary treatments $T_K = (T_{i_1}, \dots, T_{i_n})$ on the outcome Y is

$$\begin{aligned} \mathcal{M} \ni p \mapsto \Phi_{i_1, \dots, i_n}^a(p) &:= I_{i_1, \dots, i_n}^a \\ &= \sum_{j=0}^n (-1)^{n-j} \left(\sum_{J \subset K: \ell(J)=j} \mathbb{E}(Y | T_K = e_J^{(n)}, \underline{T} = 0) \right), \end{aligned} \quad (10)$$

where the internal sum runs over all subsets $J \subset K$ of length $\ell(J) = j$.

This is the n th order Boolean derivative of the function $(T_1, \dots, T_n) \mapsto \mathbb{E}(Y | T_1, \dots, T_n)$. As an example, consider the three-point interaction $I_{1,2,3}^a$ among the effects of the binary random variables T_1, T_2, T_3 on the outcome Y . Then, $T_K = (T_1, T_2, T_3)$ with $K = \{1, 2, 3\}$, and $I_{1,2,3}^a$ consists of $2^3 = 8$ terms. Explicitly, the interaction reads as

$$\begin{aligned} I_{1,2,3}^a &= \mathbb{E}(Y | T_K = (1, 1, 1), \underline{T} = 0) - \mathbb{E}(Y | T_K = (1, 1, 0), \underline{T} = 0) \\ &\quad - \mathbb{E}(Y | T_K = (1, 0, 1), \underline{T} = 0) - \mathbb{E}(Y | T_K = (0, 1, 1), \underline{T} = 0) \\ &\quad + \mathbb{E}(Y | T_K = (1, 0, 0), \underline{T} = 0) + \mathbb{E}(Y | T_K = (0, 1, 0), \underline{T} = 0) \\ &\quad + \mathbb{E}(Y | T_K = (0, 0, 1), \underline{T} = 0) - \mathbb{E}(Y | T_K = (0, 0, 0), \underline{T} = 0). \end{aligned}$$

Note that the four terms with a “+” are those for which an *odd* number of variables satisfies $T_i = 1$, whereas the four terms with a “−” are those for which an *even* number of variables satisfies $T_i = 1$. This is the other way around for two-point interactions [see Eq. (3)] and depends on the parity of the number n in general as follows from Eq. (10).

For a diagrammatic relation between the three-point interaction and the two-point interactions from which it is built, as in Eq. (5), together with an interpretation of n -point interaction in general, we refer the reader to Sec. II E. Finally, we show in Corollary 2 that I_{i_1, \dots, i_n}^a is symmetric under any permutation of its indices i_1, \dots, i_n .

Our additive notion of n -point interaction among binary random variables readily generalizes to the setting of categorical variables. Recall that a *categorical random variable* X distinguishes $k + 1$ categories, typically labeled by integers $0, 1, \dots, k$, where the probability of being in category i equals $p(X = i) = p_i$ and the $p_i \in [0, 1]$ sum to 1. If $k = 1$, then X is a binary random variable. The categorical case leads to new phenomena, most importantly the dependence of the interaction I_{i_1, \dots, i_n}^a on the particular categories of T_{i_1}, \dots, T_{i_n} one considers. Indeed, e.g., $I_{i,j}^a$ in the binary case has a unique double derivative whereas in general a derivative is a *function* that needs to be evaluated at a point (i.e., a category) in order to obtain a *value*.

Before we define interaction as a target parameter, we again specify the statistical model:

$$\mathcal{M} = \{p(Y, T_1, T_2, \dots, T_r, W) \mid Y \text{ continuous, } T_i \text{ categorical with } k_i \in \mathbb{N} \text{ categories, } W \text{ a covariate}\}$$

Let t_i, t'_i and t_j, t'_j be categories of T_i and T_j , respectively. First, we define the interaction between the effects of T_i on Y as T_i changes from t_i to t'_i and the effect of T_j on Y as T_j changes from t_j to t'_j . We write $T_i : t_i \rightarrow t'_i$ to mean that T_i changes from t_i to t'_i . For example, the average treatment effect of $T_i : t_i \rightarrow t'_i$ on Y , given $T_j = t_j$, reads as

$$\text{ATE}_{T_i:t_i \rightarrow t'_i}(Y \mid T_j = t_j) = \mathbb{E}(Y \mid T_i = t'_i, T_j = t_j) - \mathbb{E}(Y \mid T_i = t_i, T_j = t_j). \quad (11)$$

The target mapping for the additive interaction between the effects of T_i and T_j on the outcome Y is the following. The *additive interaction* $I_{i,j}^a(t_i t'_i; t_j t'_j)$ between the effect of the categorical variables $T_i : t_i \rightarrow t'_i$ on Y and the effect of $T_j : t_j \rightarrow t'_j$ on Y is given by the difference of their respective treatment effects, i.e.,

$$I_{i,j}^a(t_i t'_i; t_j t'_j) = \text{ATE}_{T_i:t_i \rightarrow t'_i}(Y \mid T_j = t'_j, \underline{T} = 0) - \text{ATE}_{T_i:t_i \rightarrow t'_i}(Y \mid T_j = t_j, \underline{T} = 0). \quad (12)$$

This definition reduces to that of Eq. (3) in the case where both T_i and T_j are binary with labels $\{0, 1\}$, i.e.,

$$I_{i,j}^a(01; 01) = I_{i,j}^a. \quad (13)$$

For properties of n -point interaction in this more general setting, such as transitivity, see Appendix A.

C. Multiplicative interaction

In this section, we define the multiplicative interaction among n binary random variables X_i forming part of a

random variable $O = (X_0, \dots, X_r)$ with joint probability density function p_0 . First of all, we specify the statistical model as in Sec. II A:

$$\mathcal{M} = \{p(X_0, X_1, \dots, X_r) \mid X_i \text{ binary random variables}\}.$$

The target map $\Phi_{i,j}^m$ is our nonparametric multiplicative formulation of two-point interaction between the binary random variables X_i and X_j :

$$\mathcal{M} \ni p \mapsto \Phi_{i,j}^m(p) := I_{i,j}^m = \frac{p(X_{ij} = (1, 1) \mid \underline{X} = 0) p(X_{ij} = (0, 0) \mid \underline{X} = 0)}{p(X_{ij} = (1, 0) \mid \underline{X} = 0) p(X_{ij} = (0, 1) \mid \underline{X} = 0)}. \quad (14)$$

The above ratios of conditional probability distributions may be expressed in terms of the joint probability distribution p since all are conditioned on $\underline{X} = 0$. As a result, the two-point interaction between, e.g., X_1 and X_2 , can be directly estimated from the data, as it reduces to

$$I_{1,2}^m = \frac{p(1, 1, 0, \dots, 0) p(0, 0, 0, \dots, 0)}{p(1, 0, 0, \dots, 0) p(0, 1, 0, \dots, 0)}. \quad (15)$$

Moreover, if a variable X_k appearing in the \underline{X} is independent of both X_i and X_j , then one need not condition on X_k . In this case, statistics may be improved as X_k drops out of the conditional joint distribution $p(X_i, X_j \mid \underline{X})$ for (X_i, X_j) . See Sec. III F where this argument is explained in detail.

The multiplicative two-point interaction $I_{i,j}^m$ of Eq. (14) between the binary random variables X_i, X_j can also be expressed in terms of their (conditional) expectation values. Numerically, this reformulation allows one to obtain uncertainties on the estimates of $I_{i,j}^m$ using, e.g., the empirical bootstrap procedure (see Sec. IV C). The expression of $I_{i,j}^m$ in terms of expectation values is derived via the product rule for probabilities, which yields

$$\frac{p(X_{ij} = (0, 0) \mid \underline{X} = 0)}{p(X_{ij} = (1, 0) \mid \underline{X} = 0)} = \frac{1 - \mathbb{E}(X_i \mid X_j = 0, \underline{X} = 0)}{\mathbb{E}(X_i \mid X_j = 0, \underline{X} = 0)},$$

and similarly for the remaining two probabilities. Therefore, the multiplicative two-point interaction (14) can be written as a combination of expectation values:

$$I_{i,j}^m = \frac{\mathbb{E}(X_i \mid X_j = 1, \underline{X} = 0) (1 - \mathbb{E}(X_i \mid X_j = 0, \underline{X} = 0))}{\mathbb{E}(X_i \mid X_j = 0, \underline{X} = 0) (1 - \mathbb{E}(X_i \mid X_j = 1, \underline{X} = 0))}. \quad (16)$$

It is not hard to see that this expression is symmetric under $X_i \leftrightarrow X_j$. For a general statement, see Proposition 2.

The following is the target map for our nonparametric multiplicative formulation of n -point interaction.

Definition 2. Let $K = \{i_1, \dots, i_n\} \subset \{0, 1, \dots, r\}$ be a subset of indices. The multiplicative n -point interaction among the binary random variables $X_K = (X_{i_1}, \dots, X_{i_n})$ is defined as

$$\mathcal{M} \ni p \mapsto \Phi_{i_1, \dots, i_n}^m(p) := I_{i_1, \dots, i_n}^m = \prod_{j=0}^n \left(\prod_{\substack{K \subset K: \ell(J)=j}} p(X_K = e_j^{(n)} \mid \underline{X} = 0)^{(-1)^{n-j}} \right), \quad (17)$$

where the internal product runs over all subsets $J \subset K$ of length $\ell(J) = j$.

As an example, consider the three-point interaction $I_{1,2,3}^m$ among the binary random variables X_1, X_2, X_3 . It consists of

$$I_{1,2,3}^m = \frac{p(X_K = (1, 1, 1) | \underline{X} = 0) p(X_K = (1, 0, 0) | \underline{X} = 0) p(X_K = (0, 1, 0) | \underline{X} = 0) p(X_K = (0, 0, 1) | \underline{X} = 0)}{p(X_K = (1, 1, 0) | \underline{X} = 0) p(X_K = (1, 0, 1) | \underline{X} = 0) p(X_K = (0, 1, 1) | \underline{X} = 0) p(X_K = (0, 0, 0) | \underline{X} = 0)}. \quad (18)$$

Note that the four terms in the numerator are those for which an *odd* number of variables satisfies $X_i = 1$, whereas the four terms in the denominator are those for which an *even* number of variables satisfies $X_i = 1$. This is the other way around for two-point interactions [see Eq. (14)] and depends on the parity of the number n in general as follows from Eq. (17). There is a large amount of symmetry in this expression:

$$I_{1,2,3}^m = \frac{I_{1,2}^m(X_3 = 1)}{I_{1,2}^m(X_3 = 0)} = \frac{I_{1,3}^m(X_2 = 1)}{I_{1,3}^m(X_2 = 0)} = \frac{I_{2,3}^m(X_1 = 1)}{I_{2,3}^m(X_1 = 0)}, \quad (19)$$

where $I_{1,2}^m(X_3 = 1)$ means that all instances of X_3 are conditioned as $X_3 = 1$, as opposed to $X_3 = 0$. The fact that all three expressions (and the remaining three) are equal follows from the $3! = 6$ symmetries of $I_{1,2,3}^m$ of Proposition 2 below. We also remark that $I_{1,2,3}^m$ can be readily computed from data since the ratios of conditional probability distributions appearing in this equation may be expressed in terms of the joint probability distribution p of O . As for the two-point interaction, a general three-point interaction $I_{i,j,k}^m$ can be expressed in terms of expectation values:

$$I_{i,j,k}^m = \frac{R_{i,jk}(1, 1) R_{i,jk}(0, 0)}{R_{i,jk}(1, 0) R_{i,jk}(0, 1)}, \quad (20)$$

where we have defined, for any variable X_i conditioned on $X_{jk} = (X_j, X_k) = (a, b)$, the following expression:

$$R_{i,jk}(a, b) = \frac{\mathbb{E}(X_i | X_{jk} = (a, b), \underline{X} = 0)}{1 - \mathbb{E}(X_i | X_{jk} = (a, b), \underline{X} = 0)}. \quad (21)$$

For any binary variable T with $p(T = 1) = p$, this fraction encodes the ratio $p/(1 - p)$. The expression of the three-point interaction $I_{i,j,k}^m$ in terms of expectation values over binary random variables is used in Sec. IV C for the purposes of numerical estimation via statistical bootstrap. It is straightforward to write an expression similar to that of Eq. (20) for any n -point interaction, making statistical bootstrap applicable in general.

Finally, we make explicit a basic and natural symmetry that is inherent in our nonparametric formulation of n -point interaction I_{i_1, \dots, i_n}^m among the binary random variables X_{i_1}, \dots, X_{i_n} : n -point interaction is invariant under any permutation σ of the n variables, namely,

$$I_{i_1, \dots, i_n}^m = I_{\sigma(i_1, \dots, i_n)}^m. \quad (22)$$

We refer the interested reader to Proposition 2 for a proof.

D. Relating additive and multiplicative formulations

Consider binary random variables X_i forming part of a random variable $O = (X_0, \dots, X_r)$ with joint probability density function p . In this section, we show that the nonparametric formulation of multiplicative n -point interaction among the

$2^3 = 8$ terms. Writing $X_K = X_{1,2,3}$ for the triple (X_1, X_2, X_3) , the interaction reads as

variables X_{i_1}, \dots, X_{i_n} is equivalent to the additive n -point interaction among the effects of the variables X_{i_1}, \dots, X_{i_n} on a particular outcome canonically related to p ; in fact, when both interactions are defined, they are related by a logarithm. This outcome is the negative of the energy function $E(\underline{X})$, obtained from the joint distribution p via

$$p(\underline{X}) = \exp(-[-\ln p(\underline{X})]) \quad \text{and} \quad E(\underline{X}) = -\ln p(\underline{X}). \quad (23)$$

Note that the expectation value of $E(\underline{X})$ is the Shannon entropy of the probability distribution p . More precisely, the additive and multiplicative n -point interactions among the X_{i_1}, \dots, X_{i_n} are related via

$$\ln(I_{i_1, \dots, i_n}^m) = I_{i_1, \dots, i_n}^a, \quad (24)$$

where the additive n -point interaction is computed with respect to the outcome $Y = -E(\underline{X})$. Indeed, this follows directly as taking the logarithm of Eq. (17) yields Eq. (10). Here we have used that

$$\frac{p(X_{i_1, \dots, i_n} = e_J^{(n)} | \underline{X} = 0)}{p(X_{i_1, \dots, i_n} = e_{J'}^{(n)} | \underline{X} = 0)} = \frac{p(X_{i_1, \dots, i_n} = e_J^{(n)}, \underline{X} = 0)}{p(X_{i_1, \dots, i_n} = e_{J'}^{(n)}, \underline{X} = 0)}, \quad (25)$$

i.e., a ratio of conditional probabilities is equal to the corresponding ratio of joint probabilities, together with the fact that an expectation value of the number

$$\alpha = \ln p(X_{i_1, \dots, i_n} = e_J^{(n)}, \underline{X} = 0)$$

equals the number itself: $\mathbb{E}(\alpha) = \alpha$. Take, as an example, the two-point interaction $I_{1,2}^m$ between X_1 and X_2 of Eq. (14):

$$\begin{aligned} I_{1,2}^m &= \frac{p(X_{12} = (1, 1) | \underline{X} = 0) p(X_{12} = (0, 0) | \underline{X} = 0)}{p(X_{12} = (1, 0) | \underline{X} = 0) p(X_{12} = (0, 1) | \underline{X} = 0)} \\ &= \frac{p(X_{12} = (1, 1), \underline{X} = 0) p(X_{12} = (0, 0), \underline{X} = 0)}{p(X_{12} = (1, 0), \underline{X} = 0) p(X_{12} = (0, 1), \underline{X} = 0)}. \end{aligned}$$

Taking the logarithm and simplifying notation to $p_{12}(X_1, X_2) = p(X_1, X_2, \underline{X} = 0)$ yields

$$\begin{aligned} \ln I_{1,2}^m &= \ln p_{12}(1, 1) - \ln p_{12}(1, 0) \\ &\quad - \ln p_{12}(0, 1) + \ln p_{12}(0, 0) = I_{1,2}^a, \end{aligned}$$

as claimed. Note that we recognize the canonical outcome $Y = -E(\underline{X}) = \ln p(\underline{X})$.

As a corollary, we deduce the general permutation symmetry of the additive n -point interaction, namely,

$$I_{i_1, \dots, i_n}^a = I_{\sigma(i_1, \dots, i_n)}^a \quad (26)$$

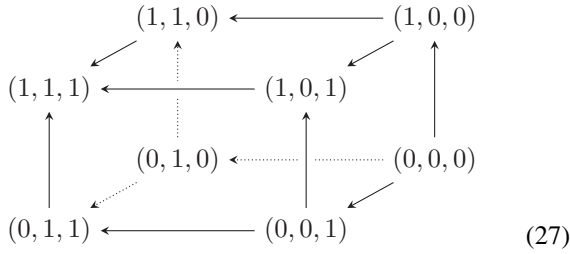
for any permutation σ ; see Corollary 2 for a proof.

E. Interpreting higher-order interactions

The nonparametric n -point interaction consists of 2^n terms, as it involves n binary variables turning on or off. Consequently, the interpretation of such higher-order interactions is

somewhat delicate. To fix ideas, we focus on the case of additive three-point interactions, the discussion readily generalizes to n -point interactions.

Let T_1, T_2, T_3 be three binary random variables and let Y denote the outcome. The interpretation of the three-point interaction $I_{1,2,3}^a$ of Sec. II B is similar to that of the two-point interaction in Eq. (5). Consider the following diagram:



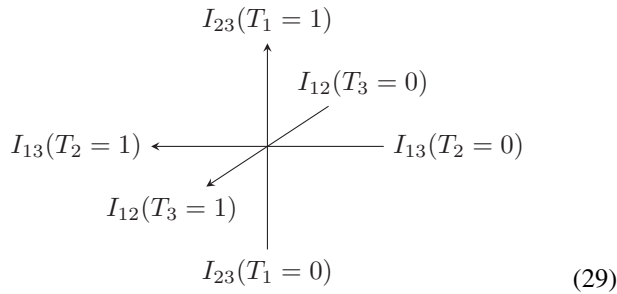
We have introduced the shorthand

$$A(t_1, t_2, t_3) = \mathbb{E}(Y \mid T_{123} = (t_1, t_2, t_3), \underline{T} = 0), \quad (28)$$

where $t_1, t_2, t_3 \in \{0, 1\}$. Vertex (t_1, t_2, t_3) represents the expected outcome $A(t_1, t_2, t_3)$. An arrow represents the ATE of the variable of which the value changes, where the sign is again dictated by target minus source. For example, the front left vertical arrow encodes the ATE:

$$A(1, 1, 1) - A(0, 1, 1) = \text{ATE}_{T_1}(Y \mid T_{23} = (1, 1), \underline{T} = 0).$$

The 12 arrows along the 6 faces of the cube (one horizontal and one vertical each) encode the 6 additive two-point interactions between the effects of two out of the three variable T_1, T_2, T_3 on the outcome Y , with the third variables fixed to 0 or 1, together with their inherent symmetry as discussed in Sec. II B. Either of the three arrows through the sides of the cube, depicted in the figure below, encodes the additive three-point interaction between the effects of T_1, T_2, T_3 on the outcome Y :



We have the relations target minus source:

$$\begin{aligned} I_{1,2,3}^a &= I_{1,2}^a(T_3 = 1) - I_{1,2}^a(T_3 = 0) \\ &= I_{1,3}^a(T_2 = 1) - I_{1,3}^a(T_2 = 0) \\ &= I_{2,3}^a(T_1 = 1) - I_{2,3}^a(T_1 = 0). \end{aligned} \quad (30)$$

This is our threefold interpretation of three-point interaction: it is the change in the two-point interaction between T_1 and T_2 , i.e., $I_{1,2}^a = I_{1,2}^a(T_3 = 0)$, as T_3 is turned on $T_3 : 0 \rightarrow 1$, yielding $I_{1,2}^a(T_3 = 1)$. In other words, $I_{1,2,3}^a$ captures the dependence of the two-point interaction between T_1 and T_2 as a function of T_3 . We conclude that the sign and magnitude of a three-point interaction can be interpreted relative to any of the two-point interactions between two out of the three variables.

As an illustration, we present the natural interpretation of symmetric higher-order interactions in the following real-world examples:

(1) Genomic variant interaction leading to disease: The additive two-point interaction answers the following question: Does variant i influence disease differently depending on the status of variant j , and by how much? The three-point interaction answers the following question: Does the interaction between variant i and variant j influence disease differently depending on the status of variant k , and by how much? The same interpretation applies to combination therapy where the effects of multiple drug interactions on health are examined.

(2) Molecular networks: The multiplicative two-point interaction answers the following question: Does the likelihood of gene i being on increase or decrease depending on whether gene j is on or off, and by how much? Similarly, the three-point interaction answers the following question: Does the interaction between gene i and gene j influence outcome differently, depending on the status of gene k , and by how much?

The cause-effect directionalities are either provided by subject experts, discovered by perturbation experiments, or derived by causal discovery algorithms.

F. Improving statistics via (conditional) independence

The nonparametric formulations of n -point interaction among the random variables X_{i_1}, \dots, X_{i_n} [Eqs. (10) and (17)] require conditioning on all remaining variables in the system. In order to improve statistical power when estimating interactions directly from data, this requirement can be relaxed under the assumption that the system is *Markovian*. Then, one need only condition on the *parents* of the variables X_{i_1}, \dots, X_{i_n} involved in the interaction. A finite collection of categorical random variables $\{X_i\}_{i=1}^r$ is a *Markov random field* if

(1) the joint distribution is strictly positive, i.e., $p(X_i = x_i \text{ for } 1 \leq i \leq r) > 0$, and

(2) for each X_i there exists a set of *parents* $\mathcal{P}_i \subset \{1, 2, \dots, r\}$, not including i , which is the minimal set such that the following condition holds:

$$p(X_i = x_i \mid \underline{X} = \underline{x}) = p(X_i = x_i \mid X_j = x_j \text{ for } j \in \mathcal{P}_i).$$

In words, the conditional probability of $X_i = x_i$ only depends on its parents $X_j = x_j, j \in \mathcal{P}_i$.

It is not hard to see that the set of parents \mathcal{P}_i of the variable i is unique. To any Markov random field one can associate a finite undirected graph with a vertex for each variable X_i and an edge connecting X_i and X_j if $j \in \mathcal{P}_i$, i.e., X_j is a parent of X_i . The Hammersley-Clifford theorem [17] (see also [18]) states that $\{X_i\}_{i=1}^r$ is a Markov random field if and only the joint probability distribution $p(X_1, \dots, X_r)$ is a *Gibbs ensemble*, i.e., there exists a Hamiltonian $E(X_1, \dots, X_r)$ such that

$$p(X_1, \dots, X_r) = \frac{1}{\mathcal{Z}} \exp[-E(X_1, \dots, X_r)], \quad (31)$$

where \mathcal{Z} denotes the partition function normalizing the distribution. As a result, *all* energy-based models of binary and categorical random variables are Markov random fields, and may thus benefit from the aforementioned improvement in statistical power when computing n -point interactions directly

from data. These facts are leveraged in the numerical Secs. IV C and V B below. We also remark that we regard the assumption that $\{X_i\}_{i=1}^r$ be a Markov random field as minimal in the context of inverse problems since it is a basic axiom in any treatment of causality, e.g., in the works of Pearl [7] or Rubin [16]. In practice, it may be the case that the parent structure of a Markov random field $\{X_i\}_{i=1}^r$ is not *a priori* known and is to be inferred from data. This can be achieved by applying algorithms designed to estimate conditional independence among variables in a given system, from data. These algorithms use parametric or nonparametric statistical methods, such as Pearson's χ -squared test, to establish conditional independence among categorical random variables [8–10].

As an example of a structure discovery algorithm, the Peter-Clark (PC) algorithm only scales exponentially in the *worst* case scenario. The sparser the ground truth network structure is, the faster the algorithm will converge. In Ref. [8], parallelized PC is benchmarked for constructing gene network neighboring structures for yeast (5361 variables), a bacterium (2810 variables), and DREAM5-Insilico data set (1643 variables). The algorithm was shown to converge in less than 12 h in all cases, on a personal computer with 8-cores. Once the graph structure is known or learned, estimating interactions scales as efficiently as computing averages over the data. The algorithm is therefore approximately as fast as estimating the bootstrap error on the interaction estimates.

As a simple illustration, in Sec. IV D we demonstrate the results of conditional independence tests on data generated by the two-dimensional Ising model, using the χ -squared test, and discuss the improved statistics of the interaction estimates.

III. RESULTS I: ANALYTICAL MAP TO REGRESSION AND NUMERICAL RESULTS FOR THE UK BIOBANK SIMULATION

As an elementary and concrete example, in this section we show that the nonparametric additive definition of interactions (Definition 1) reduces to an interaction coefficient in a linear regression model. We illustrate this example in the context of a biomedical application.

A. Application: Interactions in biomedicine

Genome-wide association studies (GWAS) are methods to identify genetic variants in the genome of individuals in a population, that could be associated with a disease or trait. In case-control GWAS, one searches for variants, a collection of single nucleotide changes in the DNA, that occur more frequently in people with a particular disease (cases) as compared to those without the disease (controls). The goal of GWAS is to find candidate genes that could potentially increase the risk of a certain disease, with the medical aim of identifying potential drug targets. Currently, one of the main aims of this field of study is to move away from associational to causal variant-trait relations. For the magnitude of causal effects of genomic variants on traits to be inferred accurately, one is required to (i) relax parametric assumptions such as the linear dependencies of the traits on the variants, and (ii) take into account interactions among the variants affecting traits, known as *epistasis*. In contrast to the methods used in

some of the key literature in the field [19,20], our definition of interaction via the targeted learning framework satisfies requirement (i) by removing the need for parametric assumptions altogether, and incorporates (ii) by taking into account epistatic interactions.

B. Epistatic interactions

Consider (i) a transcription factor protein which modifies gene expression by binding the DNA. The degree of binding, however, depends on the underlying DNA variants to which the transcription factor is binding. Now, suppose that (ii) there are multiple other variants across the genome that regulate the effect of another transcription factor protein, hence changing levels of gene expressions. Then, (i) and (ii) have downstream interactions that affect particular traits or diseases in humans. As the considerations of genetics and causality are beyond the scope of this work, we limit ourselves here to a sample application of our techniques in extracting such epistatic interactions, using simulated data of trait and disease representative of the summary-level UK BioBank population [13]. We consider the case of a complex continuous trait, height, as an example.

There are many variants across the genome contributing a small fraction to a complex trait such as height; this is known as the omnigenic model [21]. Suppose that we have an *a priori* understanding of which genomic variants are relevant to consider, e.g., those in the vicinity of bone developmental genes. Consider the following linear ground truth, involving six variants V_j , for $j = 1, 2, \dots, 6$, across the genome each contributing via a positive or negative coefficient to the value of height. Without loss of generality, suppose that only two of them also have a nonzero interaction (the generalization to more interactions is trivial):

$$\text{Height}^{(i)} \sim \alpha_0 + \sum_{j=1}^6 \alpha_j V_j^{(i)} + \gamma V_1 V_2 + \epsilon, \quad (32)$$

where i represents an individual, ϵ is the noise in height, and α_0 corresponds to unobserved, but independent, variants contributing to height.

We use our model-agnostic nonparametric additive two-point interaction estimator $I_{1,2}^a$ [Eq. (3)] to show we recover the coefficient γ representing the ground truth interaction between V_1 and V_2 . To see this, we simply compute the four expected outcomes in Eq. (3):

$$\mathbb{E}(H \mid V_1 = 1, V_2 = 1, V_{3,4,5,6} = 0) = \alpha_0 + \alpha_1 + \alpha_2 + \gamma,$$

$$\mathbb{E}(H \mid V_1 = 1, V_2 = 0, V_{3,4,5,6} = 0) = \alpha_0 + \alpha_1,$$

$$\mathbb{E}(H \mid V_1 = 0, V_2 = 1, V_{3,4,5,6} = 0) = \alpha_0 + \alpha_2,$$

$$\mathbb{E}(H \mid V_1 = 0, V_2 = 0, V_{3,4,5,6} = 0) = \alpha_0.$$

We obtain the following expressions for the four average treatment effects:

$$\text{ATE}_{V_1}(H \mid T_V = 1) = \alpha_1 + \gamma,$$

$$\text{ATE}_{V_1}(H \mid T_V = 0) = \alpha_1$$

$$\text{ATE}_{V_2}(H \mid T_V = 1) = \alpha_2 + \gamma,$$

$$\text{ATE}_{V_2}(H \mid T_V = 0) = \alpha_2. \quad (33)$$

The interactions both ways around are $I_{1,2}^a = \gamma = I_{2,1}^a$, as expected since interaction is symmetric by Corollary 2. In

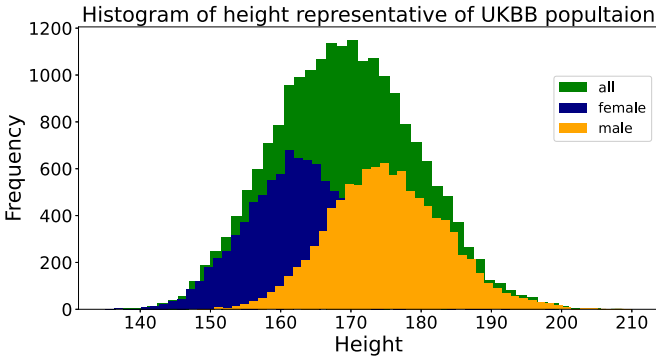


FIG. 1. Histogram of female, male, and combined heights on simulated data, such that it is representative of the UK BioBank population (UK BioBank, standing height).

conclusion, we have $I_{1,2}^a = \gamma$ as claimed. Generalizations to higher-point interactions are trivial. For a numerical example with three-point interactions, see Appendix D.

C. Numerical simulations based on the UK BioBank traits

We generate data from the above ground truth, Eq. (32). The coefficients are chosen without loss of generality to reproduce a realistic distribution of heights which is representative of the UK BioBank population [13], with approximately the same mean (168.5 cm) and standard deviation (9.3 cm) (UK BioBank, standing height).

The male and female populations are generated separately and merged to form the full distribution of height, consisting of 20 000 individuals, as presented in Fig. 1. More explicitly, without loss of generality, $\alpha_0 = 154$ for females and $\alpha_0 = 166$ for males, together with $\{\alpha_1, \dots, \alpha_6\} = \{2, 6, -3, 6, -1.5, 6\}$ with $\gamma = \epsilon = 5$. Notice that the two-point interaction γ between the two aforementioned variants is chosen to approximately equal the level of noise in height across the population. The variant allele frequencies for $V_1, V_2 \sim \text{Binom}(0.8), \text{Binom}(0.7)$, respectively, and for $V_3, \dots, V_6 \sim \text{Binom}(0.5)$.

We apply the additive targeted learning estimator of interaction (10) to the data. We obtain the targeted learning prediction $\gamma = 4.77(1.36)$ which agrees with the ground truth value $\gamma = 5$, within statistics.

NB. Since the targeted learning (TL) estimator is nonparametric, it is completely agnostic to form, e.g., linearity or nonlinearity, of the data generating process. In particular, in the case of categorical variants, there is no biological basis for the linearity assumption often used in modeling variant-trait relations. The above example merely serves to illustrate that if the underlying truth were to be linear, then the TL estimator correctly recovers this linearity. In fact, TL can be used to *test* if the effect of variants on trait is linear.

The targeted learning estimator of epistatic interactions applies to all scenarios, be they linear, nonlinear, or nonmonotonic, without requiring any parametric ansatz regarding the form of the fit function. This generality is of crucial importance since transcription factors often consist of large protein complexes that can introduce highly nontrivial behavior as well as other higher-order interactions. Such scenarios will be

missed by standard linear parametric fits. Using individual-level DNA variant and trait population data, our estimator's agnosticism and flexibility allows for new discoveries of novel and more complex interaction networks.

IV. RESULTS II: ANALYTICAL MAP AND NUMERICAL RESULTS OF THE 2D ISING MODEL AND RESTRICTED BOLTZMANN MACHINES (RBM)

In this section, we discuss Boltzmann probability distributions. In Sec. IV A, we recover the two-point couplings in an Ising Hamiltonian from the multiplicative formulation (14). In Sec. IV B, we consider a more complex Hamiltonian: The restricted Boltzmann machine (RBM). We analytically obtain its all-order couplings *without* any need for an asymptotic expansion and resummation as originally employed in [12], using the *same* universal multiplicative estimator (14). In Sec. IV C, we compare numerical results and finally, in Sec. IV D, we evaluate the improvement in the numerical results when applying Markovian conditional independence criteria.

A. Two-dimensional Ising model

We briefly recall the two-dimensional Ising model. Consider a two-dimensional square lattice of size L^2 with periodic boundary conditions, with a spin \tilde{v}_i on each lattice point i taking on values $\tilde{v}_i = \pm 1$. A *state* of the Ising model is the assignment $\tilde{\mathbf{v}}$ of a value $+1$ or -1 to each of the L^2 spins. Given a temperature T , the Boltzmann distribution describes the probability $p(\tilde{\mathbf{v}}|T)$ that the system takes on a particular state $\tilde{\mathbf{v}}$ at temperature T . Explicitly,

$$p(\tilde{\mathbf{v}}|T) = \frac{1}{\mathcal{Z}(T)} e^{-E(\tilde{\mathbf{v}})} \quad \text{where } E(\tilde{\mathbf{v}}) = - \sum_{i,j} J_{i,j} \tilde{v}_i \tilde{v}_j, \quad (34)$$

where the sum runs over all pairs of lattice sites (i, j) , where $J_{i,j}$ is the coupling between spins \tilde{v}_i and \tilde{v}_j , the external magnetic field is zero, and $\mathcal{Z}(T)$ is the partition function that normalizes this probability distribution.

In the basic version of the Ising model, the interaction between non-nearest neighbor spins is put to zero, and $J_{i,j} = \frac{1}{2T}$ for all nearest neighbor spins \tilde{v}_i, \tilde{v}_j ; this is not required in general. However, $J_{i,j} = J_{j,i}$ is symmetric.

The *inverse Ising problem* is concerned with estimating the coupling $J_{i,j}$ from data. Our nonparametric definition (14) of multiplicative two-point interaction between the binary random variables v_i and v_j recovers the coupling coefficient $J_{i,j}$ directly from the probability distribution, after applying $\ln(-)/8$; the factor of 8 is due to double counting as explained below. To see this, we first apply the bijective transformation $\tilde{v}_i = 2v_i - 1$ expressing the values of a spin v_i in terms of $\{0, 1\}$ as opposed to $\{-1, 1\}$ in order to use our definition of multiplicative two-point interaction (14). Thus, $\tilde{v}_i = -1$ corresponds to $v_i = 0$, whereas $\tilde{v}_i = 1$ corresponds to $v_i = 1$. The energy function corresponds to

$$E(\mathbf{v}) = -4 \sum_{i,j} J_{i,j} v_i v_j + 4 \sum_i \left(\sum_j J_{i,j} \right) v_i - \left(\sum_{i,j} J_{i,j} \right),$$

where we have used the symmetry $J_{i,j} = J_{j,i}$.

Next, we compute the multiplicative two-point interaction $I_{i,j}^m$ between two spins. Without loss of generality, we do this for spins v_1 and v_2 . We compute the probabilities that (v_1, v_2) takes on the values $\{(1, 1), (1, 0), (0, 1), (0, 0)\}$ with all other spins being zero, i.e., $\underline{v} = 0$. We find

$$\frac{p(1, 1, \underline{v} = 0)}{p(1, 0, \underline{v} = 0)} = \exp\left(4J_{1,2} + 4J_{2,1} - 4 \sum_{j \neq 1} J_{1,j}\right), \quad (35)$$

$$\frac{p(0, 0, \underline{v} = 0)}{p(0, 1, \underline{v} = 0)} = \exp\left(4 \sum_{j \neq 1} J_{1,j}\right), \quad (36)$$

and multiplying both yields $I_{1,2}^m = \exp(8J_{1,2})$. Hence, $\ln(I_{1,2}^m)/8 = J_{1,2}$ as claimed.

Whether or not $I_{1,2}^m$ is smaller or larger than 1 is due to the interpretation of the interaction. In this case, it is the two-point interaction between *turning on* both spins, i.e., $v_1 : 0 \rightarrow 1$ and $v_2 : 0 \rightarrow 1$, not turning them off. Alternatively, computing the additive interaction between $v_1 : 0 \rightarrow 1$ and $v_2 : 0 \rightarrow 1$ on the outcome $-E(\mathbf{v})$ is easily seen to be $I_{1,2}^a = 8J_{1,2}$. The factor of 8 is due to the change of variables $\tilde{v}_i \mapsto v_i$ and a double counting in Eq. (34). Finally, the coupling $J_{i,j}$ can be obtained directly by taking the double derivative of the outcome $-E(\mathbf{v})$ with respect to v_1 and v_2 .

In Sec. IV C, we extract $J_{i,j}$ directly from data. In order to improve the estimate of the two-point interaction $I_{i,j}^m$ from data, one may appeal to the Hammersley-Clifford theorem of Sec. II F to increase statistics by only conditioning on the relevant *parent* variables, i.e., in this case the nearest neighbors of v_i and v_j . In fact, the Monte Carlo algorithm, e.g., Metropolis, generating Ising configurations uses this feature in its update step by computing the change in energy only using nearest neighbor spins. For completeness, we analytically demonstrate that the Hammersley-Clifford theorem applies to the Ising model in Appendix C.

B. Restricted Boltzmann machine

A restricted Boltzmann machine (RBM) is a type of undirected Markov random field (MRF) with a two layer architecture. An RBM consists of m visible nodes v_j , $j \in \{1, \dots, m\}$, collectively denoted by \mathbf{v} and representing the observed input data, and n hidden nodes h_i , $i \in \{1, \dots, n\}$, collectively denoted by \mathbf{h} . We consider binary variables, i.e., $v_j, h_i \in \{0, 1\}$. The energy of the joint state $\{\mathbf{v}, \mathbf{h}\}$ of the machine is as follows:

$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_{i=1}^n \sum_{j=1}^m h_i w_{ij} v_j - \sum_{j=1}^m b_j v_j - \sum_{i=1}^n c_i h_i, \quad (37)$$

and we collectively call $\theta = \{\mathbf{w}, \mathbf{b}, \mathbf{c}\}$ the model parameters. The RBM is used to encode the joint conditional probability distribution of a state $\{\mathbf{v}, \mathbf{h}\}$ given a set of parameters θ :

$$p(\mathbf{v}, \mathbf{h}|\theta) = \frac{1}{\mathcal{Z}(\theta)} e^{-E(\mathbf{v}, \mathbf{h}; \theta)}, \quad (38)$$

where the partition function $\mathcal{Z}(\theta)$ normalizes the probability distribution. Marginalizing over the binary hidden variables h_i yields the probability distribution of the variables in the

visible layer [22]:

$$p(\mathbf{v}|\theta) = \frac{1}{\mathcal{Z}(\theta)} \prod_{j=1}^m (e^{b_j v_j}) \prod_{i=1}^n (1 + e^{c_i + \sum_{j=1}^m w_{ij} v_j}). \quad (39)$$

By equating the RBM energy function to the two-dimensional Ising energy function, the expression

$$J_{j_1, j_2} = \frac{1}{8} \ln \prod_{i=1}^n \frac{(1 + e^{c_i + w_{ij_1} + w_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + w_{ij_1}})(1 + e^{c_i + w_{ij_2}})} \quad (40)$$

is obtained in [12]. This expresses the Ising coupling J_{j_1, j_2} in terms of the model parameters of the RBM. The proof uses an asymptotic expansion and a resummation. Computing the nonparametric two-point interaction, as in Eq. (14), of the RBM readily yields the above formula:

$$\frac{1}{8} \ln (I_{j_1, j_2}^m) = J_{j_1, j_2}, \quad (41)$$

where I_{j_1, j_2}^m is computed from Eq. (38). Indeed, this follows from Eq. (14) by a direct computation since

$$\begin{aligned} I_{j_1, j_2}^m &= \frac{p(v_{j_1 j_2} = (1, 1), \underline{v} = 0) p(v_{j_1 j_2} = (0, 0), \underline{v} = 0)}{p(v_{j_1 j_2} = (1, 0), \underline{v} = 0) p(v_{j_1 j_2} = (0, 1), \underline{v} = 0)} \\ &= \prod_{i=1}^n \frac{(1 + e^{c_i + w_{ij_1} + w_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + w_{ij_1}})(1 + e^{c_i + w_{ij_2}})}. \end{aligned}$$

Indeed, both the partition functions and the b_j coefficients cancel out. By the same argument, one immediately recovers the closed form expression for the three-point interaction between $v_{j_1}, v_{j_2}, v_{j_3}$ as derived in [12, Eq. (66)], and the closed form expressions for all n -point interactions, without having to resolve to an asymptotic expansion and resummation as in [12].

C. Numerical results for the Ising model and comparisons with the RBM

In this section, we generate two-dimensional Ising configurations at various values of temperature using MAGNETO [23], a fast parallel C++ Monte Carlo code available online. We set $J_{ij} = 1/2T$ in Eq. (34). We then use the nonparametric multiplicative definition of interactions, Sec. II C, to extract the couplings J_{ij} directly from the data, i.e., we solve the inverse problem. We demonstrate agreement with the ground truth and compare the performance of the estimation of interactions directly from the data with the estimates obtained via the RBM [12]. Ising states generated by MAGNETO consist of spins ± 1 . Note that these are converted to 0, 1 as input to both the multiplicative interaction formulation and the RBM, as already discussed in Sec. IV A. Before delving into the numerical analysis, our main results are summarized in the paragraph below.

In general, the nonparametric interaction converges to the true value in the infinite data limit as it is unbiased, whereas the RBM need not do so as the original data are almost surely not generated from an RBM distribution. However, for finite sample sizes, the direct computation may become noisy and unstable without additional information, such as conditional independence among the variables. Take, for example, the case of the Ising configuration in different temperature

regimes. At low temperatures the system is highly coupled and symmetric with respect to configurations mostly containing spin zeros and those mostly containing spin ones. In this regime, there are enough samples to estimate conditional probabilities appearing in Eq. (14). On the other hand, it is harder to train an RBM in highly coupled systems, e.g., in [12] more precise hyperparameter tuning and longer training was required. This behavior of the RBMs has been reported previously in the literature [22] and is due to the machine remaining in local minima of the activation function. To avoid this problem, the RBM needs to be trained using more advanced algorithms such as parallel tempering [22] which allows the machine to exit potential local minima. Of course, this in turn requires tuning of extra hyperparameters and results in longer training times. For temperatures above the critical temperature, the system becomes weakly coupled and moves toward more randomly distributed zero and one spin configurations. In this scenario, conditioning on all but two variables in the system results in very low sample sizes and unstable estimates of the interactions unless the total sample size is very large. The RBM, on the other hand, captures the interactions well given a comparable sample size. If however, information about conditional independence among the variables in the system is used, the nonparametric estimates perform better than the RBM in terms of bias, variance, and compute time. In what follows, we quantify the above statements explicitly.

Before we present numerical results, we note that excluding higher-order interaction terms from the outset necessarily results in biased or incorrect estimates of even the two-point and self-couplings. To give a simple example, consider the following formula:

$$\begin{aligned}
 E &= E_0 + h_1 v_1 + h_2 v_2 + J_{12} v_1 v_2 + J_{123} v_1 v_2 v_3 \\
 &= E_0 + h_1 v_1 + h_2 v_2 + (J_{12} + J_{123} v_3) v_1 v_2. \quad (42)
 \end{aligned}$$

Thus, any parametric fit ignoring third order (and higher) interactions will incorrectly report $J_{12} + J_{123} \mathbb{E}(v_3)$ as the two-point interaction. More disturbingly, in a situation where the ground truth satisfies $J_{12} = 0$ but $J_{123} \neq 0$, a truncated parametric fit will incorrectly produce the nonexistent two-point interaction $J_{123} \mathbb{E}(v_3)$. Our method avoids this problem entirely.

Using the TL universal estimator (14) directly, it is possible to obtain an accurate estimate of the couplings at cold temperatures, *without* conditioning on the Markovian parents or using translational invariance. Unlike Refs. [1–5] no parametric assumptions, regularization, truncation of higher-order interactions, or other approximations are required. The results are shown in Fig. 2.

Above the critical temperature, however, TL estimation requires larger samples sizes. More explicitly, beyond $T = 2.4$, the states become more random, and conditioning on all v_i 's to be zero, apart from the two spins whose interaction is to be estimated, results in low sample sizes and unstable predictions of the conditional probabilities appearing in Eq. (14). This is demonstrated by plotting the bin sizes used to estimate the probabilities at various values of temperature in Fig. 3.

Note that, as mentioned earlier, the nonparametric approach of estimating coupling from the data is an unbiased estimator and only limited by the amount of data. Therefore,

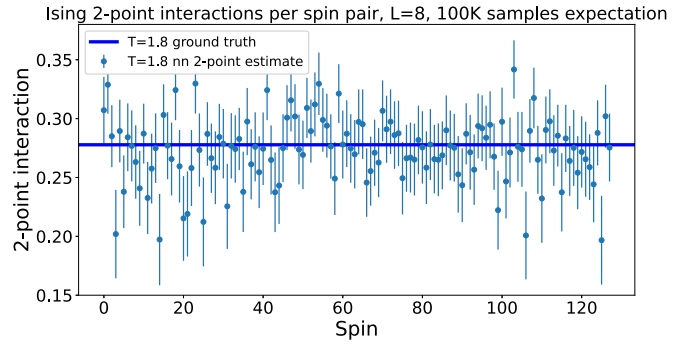


FIG. 2. All nonzero two-point interaction estimates using Eq. (14) directly, at temperature $T = 1.8$, in an Ising system of size $L^2 = 8^2$ with periodic boundary conditions. 100 000 samples are used for this estimation. No conditioning on the Markovian parents is performed, no translational invariance assumptions are made.

larger samples sizes are required, if one wishes to make no physical approximation or further assumptions about, e.g., conditional independence among the variables. Figure 4

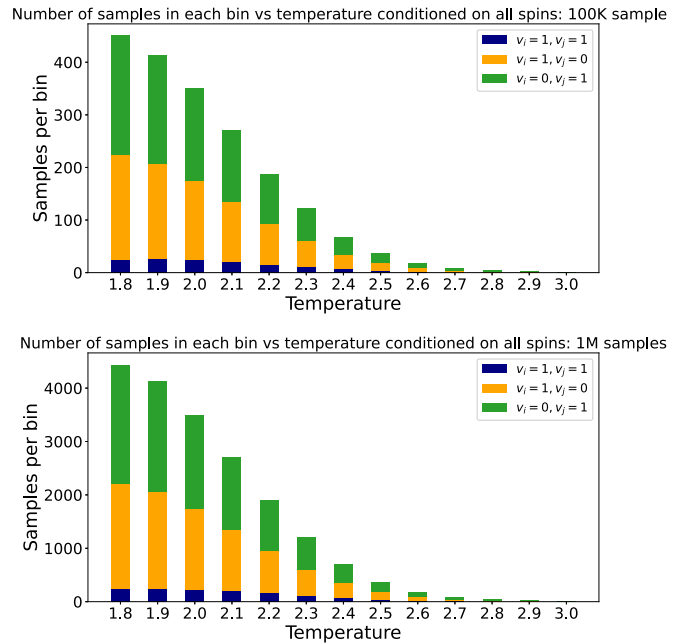


FIG. 3. Average sample sizes for conditional probabilities entering the computation of the two-point interaction for the nearest neighbor pairs in an $L^2 = 8^2$ lattice. These values are obtained by conditioning on all other spins. The bin $v_i = v_j = 0$ is left out as it has the largest size as compared to the other three. The top plot is from 100 000 samples, and the bottom is from 1×10^6 samples. Notice that each of the bin sizes increases 10-fold as we go from 100 000 to 1×10^6 samples, as expected. Observing the 100 000 plot, it is clear that above $T = 2.6$, there are not enough samples in the $v_i = v_j = 1$ bin to yield reliable estimates of the interactions, with $T = 2.6$ containing approximately nine samples on average. With 1×10^6 total samples, one can obtain estimates for $T = 2.7$, which on average contain 10 samples in the $v_i = v_j = 1$ bin, respectively. Beyond this temperature, one has to again increase the sample size to 2×10^6 or more.

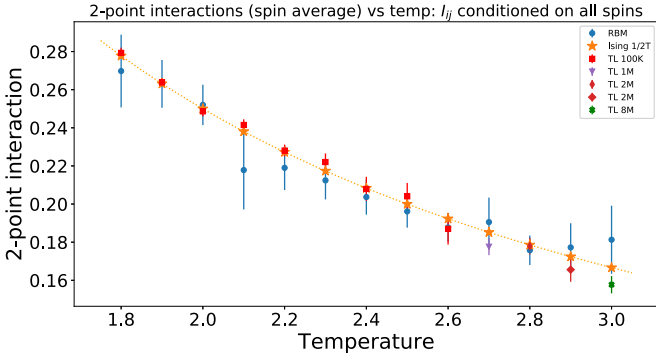


FIG. 4. A comparison of estimates of the two-point interaction among nearest neighbor spins as the temperature varies, in an Ising system of size $L^2 = 8^2$ with periodic boundary conditions, averaged over all 128 pairs of nearest neighbors for summary illustration. Each point represents a bootstrap average with error bar given by the bootstrap error. For $T \leq 2.6$, 100 000 samples are enough to estimate the nearest neighbor interactions. For $T > 2.6$ substantially more samples are required for stable estimates of the interactions. At $T = 3.0$, 8M samples are required for a stable estimate.

indicates this requirement: Above the critical temperatures, the sample sizes need to be increased from 100 000 to 1×10^6 and 10×10^6 , at very hot temperatures, in order to estimate the couplings. As expected, in Fig. 4 the estimates converge to the theoretical ground truth when the samples sizes are sufficiently increased. Note that translational invariance is not a requirement and is merely used as a summary to illustrate convergence of the nonzero couplings to the correct ground truth value.

We now demonstrate improvements in the estimates of interactions at all values of temperatures, by using information on conditional independence among the spins. This allows for a substantial reduction in the sample sizes required, especially at high temperature. As discussed earlier in Sec. II F and will be further explained in Sec. IV D, to obtain correct estimates of interaction among spins of interest, it is sufficient to condition on their parents, i.e., nearest neighbor spins, as opposed to all other spins in the rest of the lattice. For interactions between pairs of nearest neighbor spins, we condition on their six nearest neighbors, while for interactions between pairs

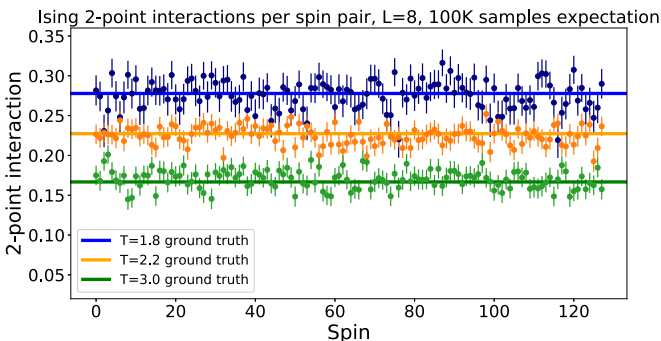


FIG. 5. Conditioning on the nearest neighbors (as prior information) to estimate I_{ij}^m substantially improves the estimates. 100 000 samples for estimations at $T = 1.8, 2.2, 3.0$, $L^2 = 8^2$.

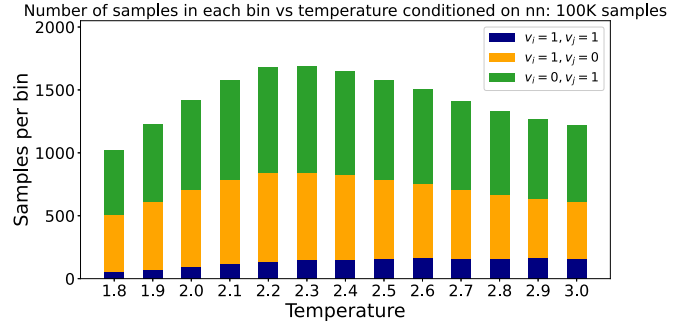


FIG. 6. Average sample sizes for conditional probabilities entering the computation of the two-point interaction for the nearest neighbor pairs in an $L^2 = 8^2$ lattice. These values are obtained by conditioning on the nearest neighbor spins only. The bin $v_i = v_j = 0$ is left out as it has the largest size as compared to the other three. There are enough samples in each bin to yield stable estimates of each conditional probability/expectation value.

of non-nearest neighbor spins we condition on their 4 + 4 nearest neighbor spins.

The *individual* per spin pair results, *without* using translational averaging, for $T = 1.8, 2.2, 3.0$ are shown in Fig. 5. Individual vanishing per spin triplet and quadruplet three- and four-point interactions are presented in Appendix F (Fig. 24) with $T = 1.8$ as an example. Figure 6 indicates an increase in the smallest bin size, i.e., $v_i = v_j = 1$, at all temperatures. This results in more precise estimates for the couplings, presented in Fig. 7,¹ by using translational invariance. Again, note that translational invariance used in Fig. 7 is not a requirement and is merely used as a summary for comparison with the RBM results in [12].

Figure 8 (upper) indicates individual spin pair couplings I_{ij}^m , estimated using Eq. (16) over 100 000 samples as compared to 20 000 (lower) for both nearest and non-nearest

¹All run times are measured on a MacBook Pro (2018) machine, 6-Core Intel i9 with 16GB memory.

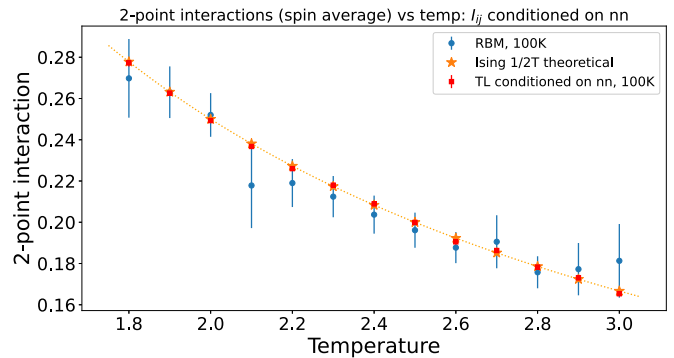


FIG. 7. Conditioning on the nearest neighbors (as prior information) to estimate I_{ij}^m substantially improves the estimates as compared to Fig. 4. 100 000 samples are used for both training the RBM and estimating the interactions directly using TL. See Fig. 23 for the successful estimation of interactions and their uncertainty using TL, with 10 000 samples. The run time for each estimation using TL is at the order of a few seconds.

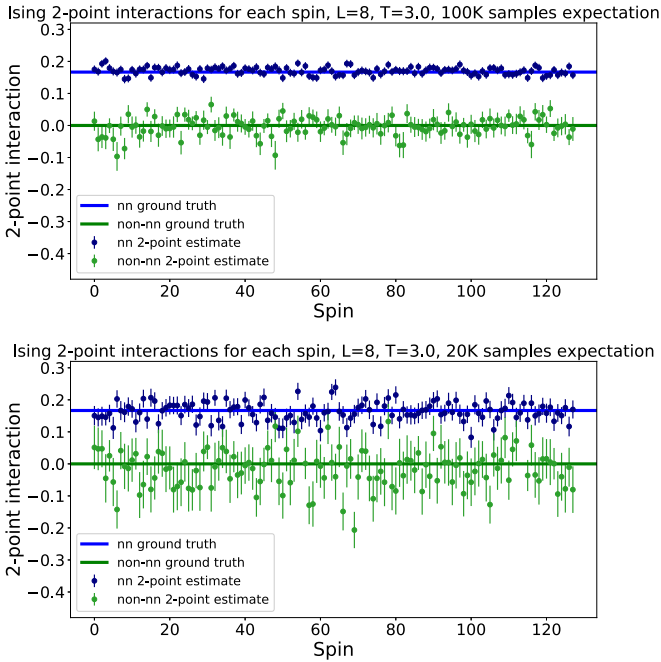


FIG. 8. $L^2 = 8^2$, $T = 3.0$, with conditioning on the nearest neighbors to estimate I_{ij}^m for both nearest and non-nearest neighbor spin pairs. In order to reduce clutter, the same number of non-nearest as nearest neighbor couplings are shown (128). No translational invariance is used. Top: the results are computed over a total of 100 000 samples, using Eq. (16) and statistical bootstrap, as compared to bottom: The results are computed over a total of 20 000 samples. For the latter, approximately 2% of spins had no samples in the p_{11} bin. This is because it is unlikely that two spins have value one, while their eight nearest neighbors all have spin value zero, as the total sample size reduces.

neighbor spins. The latter results are more noisy as expected. As compared to the 100 000, 20 000 total samples approximately had 2% of spin pairs with no samples in the p_{11} bin. This is due to the fact that it is unlikely that two spins having value one, while their eight nearest neighbors all have spin zero. This scenario is observed more often at colder temperatures (see Figs. 21 and 22 in Appendix F). Note that the nonparametric method of estimation, combined with information on conditional independence among the variables, has nevertheless enabled us to obtain accurate estimates of the interactions relying on a smaller number of samples in total. For example, using this method, there is enough power to estimate all the nearest neighbor spin pair interactions and approximately 83% of the non-nearest neighbor spin pair interactions for temperature $T = 2.2$ using 10 000 sample only, as demonstrated in Fig. 8. In contrast, e.g., the RBM does not train well on Ising data with 10 000 samples (see [12, Fig. 31]), and therefore is not able to provide accurate estimates of the interactions at low sample sizes.

Finally, we present the results of estimating the two-point interactions per individual spin pair, for a $L^2 = 32^2$ lattice at temperature $T = 3.0$, in Fig. 9. As expected, the results for the case of 20 000 total samples is more noisy, however, the signal is clearly distinguishable from background with most of the nearest-neighbor interactions being more than

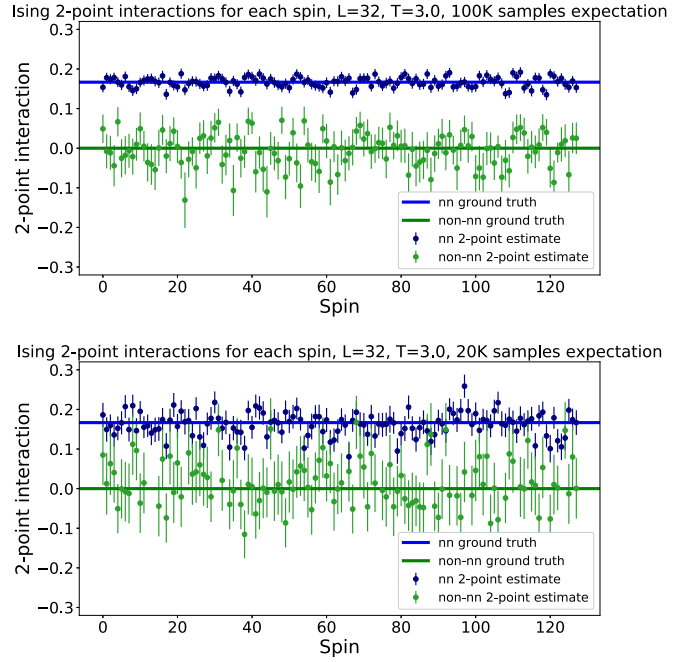


FIG. 9. $L^2 = 32^2$, $T = 3.0$, with conditioning on the nearest neighbors to estimate I_{ij}^m for both nearest and non-nearest neighbor spin pairs. In order to reduce clutter, 2×128 interactions are shown. No translational invariance is used. Top: The results are computed over a total of 100 000 samples, using Eq. (16) and statistical bootstrap, as compared to bottom: The results are computed over a total of 20 000 samples. For the latter, there is sufficient power to accurately estimate all the nearest neighbor interactions, as well as approximately 98% of non-nearest neighbor interactions.

3σ away from the zero line. We note that training an RBM on a lattice of this size, if possible, is expected to be computationally expensive and not possible for low numbers of sample sizes. This is due to the fact that a $L^2 = 32^2$ lattice contains 1024 spins which would correspond to an RBM with 1024×1024 weights + 2×1024 bias terms, i.e., 1 050 624 parameters to be determined, when the number of hidden nodes (1024) is set equal to the visible nodes (1024). The run time of the nonparametric approach is of the order of minutes on a local computer.

D. Numerical evidence for conditional independence

In the first step of the targeted learning road map stated in Sec. II A, we select the set of probability distributions p that are compatible with *a priori* knowledge regarding the data and how it is generated. For example, in the case of the Ising model, this knowledge could include information regarding the nearest neighbor structure, namely, that by conditioning on the parental spins of two spins, the two spins become independent of each other and the rest of the spins if they are non-nearest neighbors. If they are nearest neighbors, then they only become independent of the rest of the spins but not of each other. Then, using the Markovian property and the Hammersley-Clifford theorem of Sec. II F, to obtain the interactions between pairs of spins, it suffices to condition on their nearest neighbors to be zero, rather than all the rest of the spins (see Appendix C for a proof). This results in improved

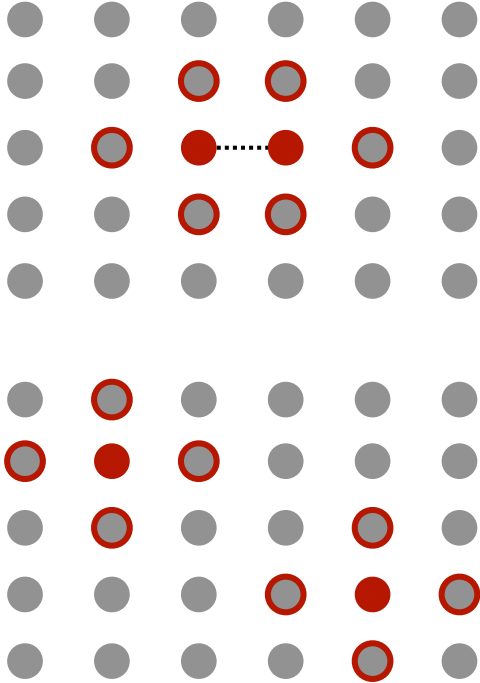


FIG. 10. Nearest neighbor structure in the two-dimensional Ising model. Parents of the pairs of interest required for conditional independence: the six parents of a nearest neighbor pair (top), and the eight parents of a non-nearest neighbor pair (bottom).

statistical estimates, as the number of samples that satisfy the latter condition will be significantly larger than the former. The Markovian parent structure of nearest and non-nearest neighbors in the two-dimensional Ising model is presented in Fig. 10.

If *a priori* information on conditional independence is not known, one can use nonparametric statistical testing to determine such independence criteria, in order to improve the estimates of interactions. The χ -squared test of independence can be used for the case of binary or categorical variables and, e.g., an information-theoretic independence criterion for continuous variables [24]. Algorithms such as Peter-Clark can then be employed to automatically detect (conditional) independence using a given test in an efficient way [10]. Discussion on the latter is beyond the scope of this work, and we only briefly present results on applying a χ -squared test directly on Ising data as an example.

We perform the χ -squared test of independence on Ising configurations generated at the critical temperature which is approximately $T = 2.3$. The null hypothesis H_0 of χ -squared is that the variables are *independent* of each other. Given a particular threshold, if the computed p values become less than the threshold, we reject the null hypothesis in favor of the alternative hypothesis H_1 , i.e., that the variables in question are indeed dependent. For the two-dimensional Ising model at the critical temperature we expect the correlation length to diverge, and therefore to observe a large degree of dependence among all spins. Therefore, taking pairs of spins, while conditioning on no other spins in the system, we expect the χ -squared test to result in small- p values, indicating dependence among the spins. Indeed, we observed $p \approx 0$ for all pairs of

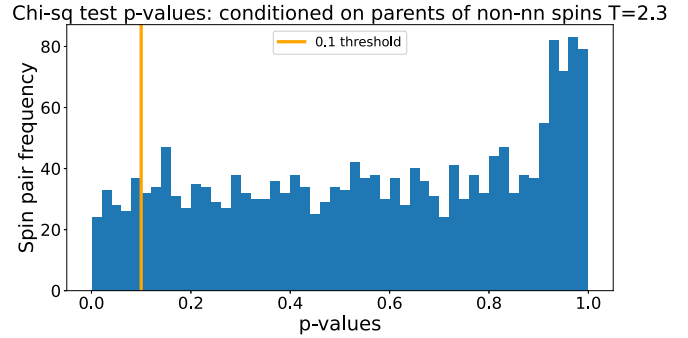


FIG. 11. Histogram of χ -squared test p values for non-nearest neighbor spins pairs, conditioned on all of the eight parents, for the $T = 2.3$ Ising model. We expect the null hypothesis of independence not to be rejected, i.e., high- p values. This is indeed observed with less than 10% of the p values being less than the chosen threshold 0.1. The χ -squared test has incorrectly taken these as dependent, however, taking more spins into account when conditioning does not introduce any bias in the estimation of the interactions.

spins in this case. If, on the other hand, we condition on all eight nearest neighbor spins of any non-nearest neighbor spin pair, we observed that most of the p values are large, indicating independence as expected. However, the test does result in less than 10% of the non-nearest neighbor spin pairs having small- p values, namely, less than the chosen threshold of 0.1 (see Fig. 11). These are the result of a type I error, or false claim of dependence, which do not bias the estimation of the interactions but merely render the procedure more conservative than necessary, at the cost of larger variance.

Next, we observe what happens if we, wrongly, do not condition on all the parents of variables that χ -squared otherwise declares as dependent. As an example, conditioning on only two of the total of eight nearest neighbors, the χ -squared test declares all $p \approx 0$. Estimating the interaction between non-nearest neighbor spin pairs, while conditioning on two parents only, results in highly biased estimates of the interactions, as expected, as indicated on the right-hand side of Fig. 12.

Finally, we condition on four out of the eight nearest neighbors, for all the non-nearest neighbor spin pairs, with all four blocking one of the spins from the rest of the system. In this case the χ -squared test seems to declare independence in most cases. This is a type II error: failure to reject a false null hypothesis of independence. We examine the resulting bias on the estimates for the associated two-point interactions in Fig. 13: The level of statistical variation in the data is large enough to compensate for the bias introduced by not conditioning on all the Markovian parents. In the tests that we have performed, we have observed these features both at cold and hot temperatures as well.

In summary, when *a priori* knowledge regarding independence among variables is not available and has to be derived from the data, one can perform the nonparametric χ -squared test for binary and categorical data. If χ squared declares dependence among variables, we must ensure to condition on these when estimating the interactions. If χ squared declares false independence, potentially due to the level of variance or noise in the data, it is likely to be the case that this missed degree of dependence is not so large as to bias the estimates of

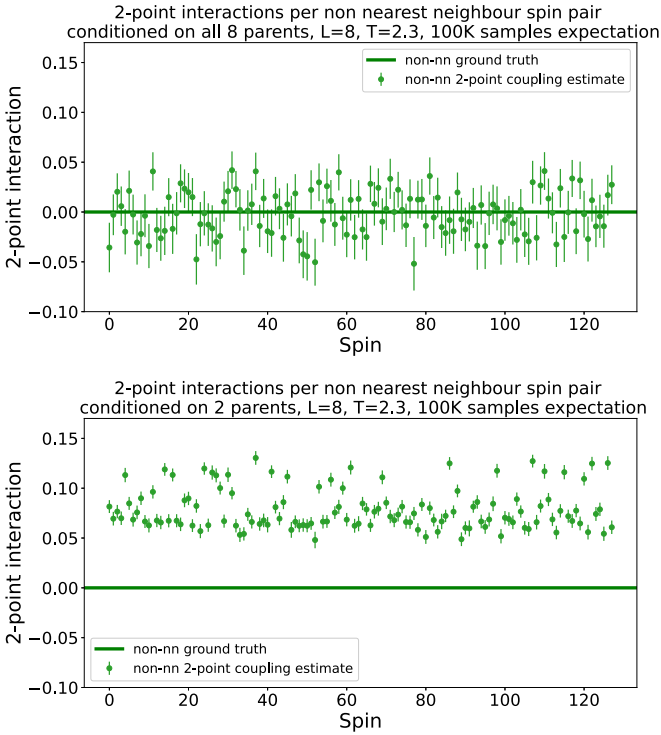


FIG. 12. Non-nearest neighbor two-point interactions for Ising configurations near the critical temperature $T = 2.3$, 100 000 samples. 128 spin pairs are taken as representatives of all 1888 non-nearest neighbor spin pairs. Top: conditioning on all eight parents, estimation accurately recovers the ground truth. Bottom: Conditioning on only two parents, even though χ -squared has accurately detected dependence, results in biased estimates of the interactions.

n -point interaction, again given the level of variance or noise in the data.

V. RESULTS III: A HAMILTONIAN WITH ONE-, TWO-, THREE-, AND FOUR-POINT INTERACTIONS

A. Analytical formulation

In this section, we consider an Ising-type Hamiltonian in the $\{-1, 1\}$ basis with four-point couplings. After transforming to the $\{0, 1\}$ basis, this results in a Hamiltonian with nonzero self-, two-, three-, and four-point couplings. The setup is as follows. Consider a two-dimensional square lattice of size L^2 with periodic boundary conditions, with a spin \tilde{v}_i on each lattice point i taking on values $\tilde{v}_i = \pm 1$. A *state* is the assignment $\tilde{\mathbf{v}}$ of a value $+1$ or -1 to each of the L^2 spins. The Boltzmann distribution describes the probability $p(\tilde{\mathbf{v}}|T)$ that the system takes on a particular state $\tilde{\mathbf{v}}$ at temperature T , i.e.,

$$p(\tilde{\mathbf{v}}|T) = \frac{1}{\mathcal{Z}(T)} e^{-E(\tilde{\mathbf{v}})}, \quad (43)$$

where

$$E(\tilde{\mathbf{v}}) = -\frac{1}{T} \sum_{(i,j)} J_{i,j} \tilde{v}_{(i,j)} \tilde{v}_{(i+1,j)} \tilde{v}_{(i,j+1)} \tilde{v}_{(i+1,j+1)}. \quad (44)$$

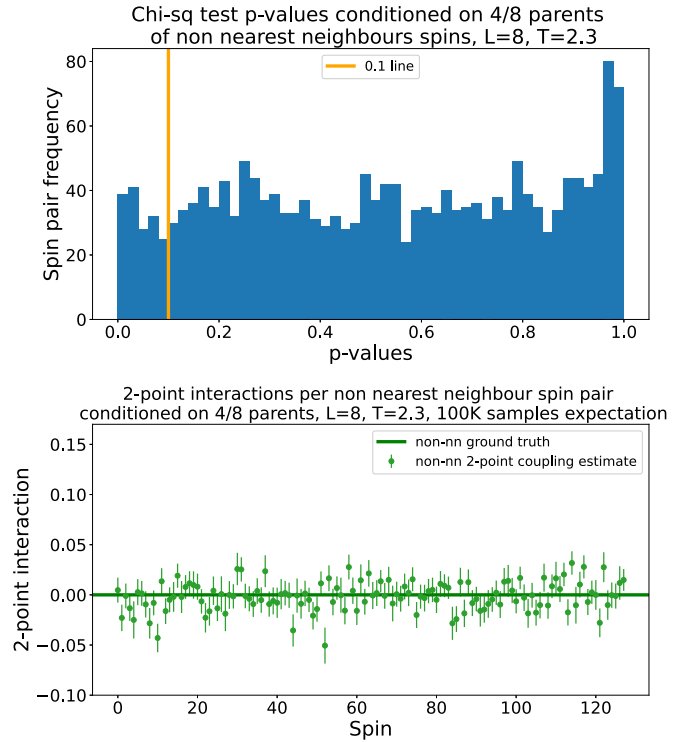


FIG. 13. Non-nearest neighbor two-point interactions for Ising configurations near the critical temperature $T = 2.3$, 100 000 samples. 128 spin pairs are taken as representatives of all 1888 non-nearest neighbor spin pairs. Top: Conditioning on four out of the total of eight parents, the χ -squared test is unable to detect dependence. Bottom: numerical results indicate that when χ squared does not detect dependence in the data, conditioning on four out of the total of eight parents does not introduce strong bias in estimating the interactions accurately.

The sum runs over all L^2 lattice sites $(i, j) \in \{1, 2, \dots, L\}^2$ and $J_{i,j}$ is the coupling among the square of spins $\{\tilde{v}_{(i,j)}, \tilde{v}_{(i+1,j)}, \tilde{v}_{(i,j+1)}, \tilde{v}_{(i+1,j+1)}\}$, see Fig. 14.

We first solve the inverse problem defined by the Hamiltonian of Eq. (44) analytically. Our nonparametric Definition 2 of multiplicative self-, two-, three-, and four-point interaction

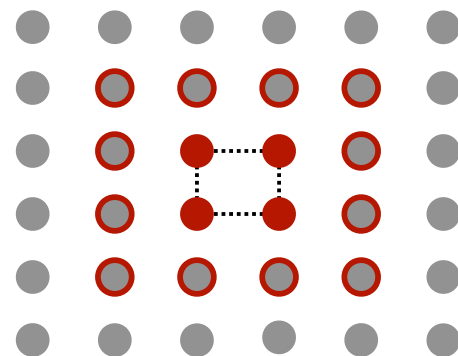


FIG. 14. Nearest neighbor structure in the Ising-type Hamiltonian with four-point interactions. There are 12 parents to be conditioned on for estimating the four-point interaction among the quadruple of spins of interest.

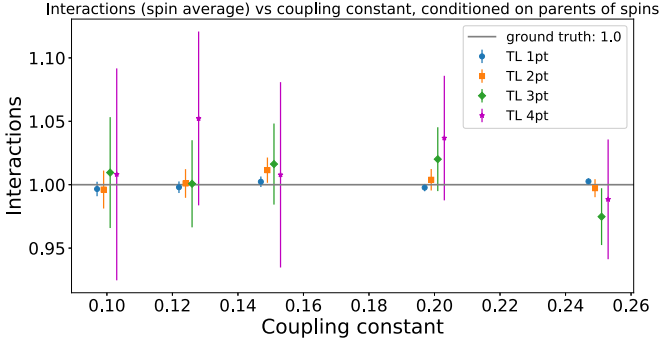


FIG. 15. Estimates of the self- to four-point interactions I_{ijkl}^m averaged across spins and normalized by various values of coupling constants in the Hamiltonian 0.1, 0.125, 0.15, 0.2, 0.25. Estimations are performed using 1×10^6 samples. As the total number of samples used for estimation is lowered, the power to detect higher-order interactions is reduced.

among binary variables immediately recovers the couplings $-8J_{i,j}$, $8J_{i,j}$, $-8J_{i,j}$, and $16J_{i,j}$ respectively, from the probability distribution of Eq. (44), after applying $\ln(-)$ and correcting for double counting due to the change of basis $\{-1, 1\} \mapsto \{0, 1\}$. To see this, we first apply the transformation $\tilde{v}_{(i,j)} = 2v_{(i,j)} - 1$ expressing the values of a spin in terms of $\{0, 1\}$ as opposed to $\{-1, 1\}$ in order to apply the definition of multiplicative n -point interaction of Eq. (17). Thus, $\tilde{v}_{(i,j)} = -1$ corresponds to $v_{(i,j)} = 0$, whereas $\tilde{v}_{(i,j)} = 1$ corresponds to $v_{(i,j)} = 1$. This yields

$$J_{i,j} \tilde{v}_{(i,j)} \tilde{v}_{(i+1,j)} \tilde{v}_{(i,j+1)} \tilde{v}_{(i+1,j+1)} = J_{i,j} (2v_{(i,j)} - 1) (2v_{(i+1,j)} - 1) (2v_{(i,j+1)} - 1) (2v_{(i+1,j+1)} - 1),$$

for the contribution to $E(\mathbf{v})$ of a single square of spins with the top left spin at lattice site (i, j) . The interactions may now be computed by taking suitable derivatives of the energy function $E(\mathbf{v})$ in the $\{0, 1\}$ basis, while putting the remaining spins to zero, and taking care of double counting due to the change of basis.

B. A Hamiltonian with four-point interactions

In this section, we evaluate the performance of our nonparametric formulation of multiplicative interaction on data generated by an Ising-type Hamiltonian with four-point couplings in the $\{-1, 1\}$ basis. This corresponds to having nonzero self-, two-, three-, and four-point interactions in the $\{0, 1\}$ basis.

One million samples were generated using the Metropolis algorithm at $T = 1$ and different coupling constants 0.1, 0.125, 0.15, 0.2, 0.25. The results for self- to four-point interactions, normalized by the corresponding coupling constant and corrected for change of basis factors, are presented in Fig. 15. As expected, the uncertainty on the estimations increases as we consider higher-order interactions. Nevertheless, at one million samples, the uncertainty on the average four-point interaction is approximately less than 10% in this system. Reducing the sample sizes from 1×10^6 to 500 000, then to 200 000, results in not having sufficient power to estimate the four-point and the three-point interactions,

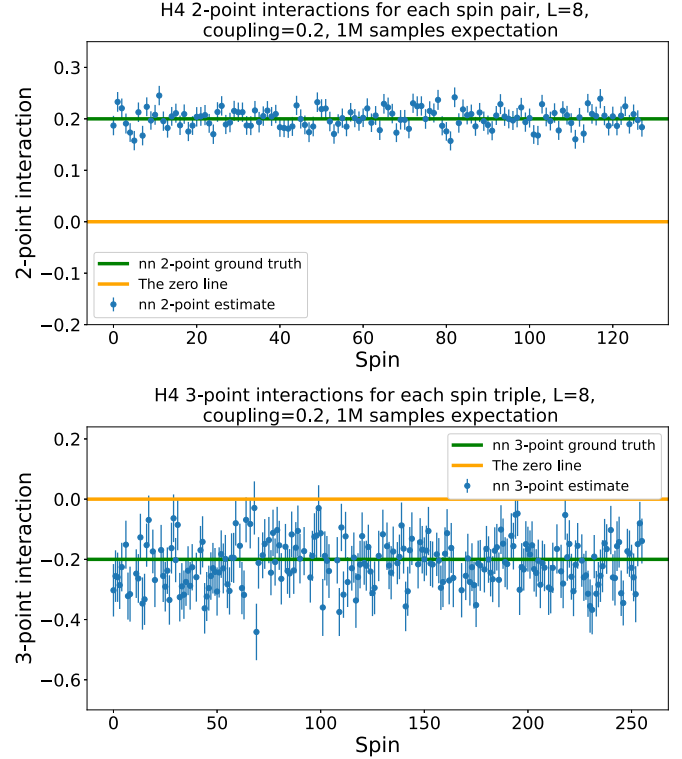


FIG. 16. Two-point (top) and three-point (bottom) per spin estimates of interactions for the ground truth coupling constant 0.2. Estimations are performed on 1×10^6 samples.

respectively. The results for the interactions per pair, triple, and quadruple of spins are presented in Figs. 16 and 17.

C. Interaction in energy-based models

Our nonparametric definition of n -point interaction applies to any set of n binary and categorical random variables in any probability distribution p . For example, if the probability distribution is believed to be a Boltzmann distribution, our formulation can be used to estimate all the n -point

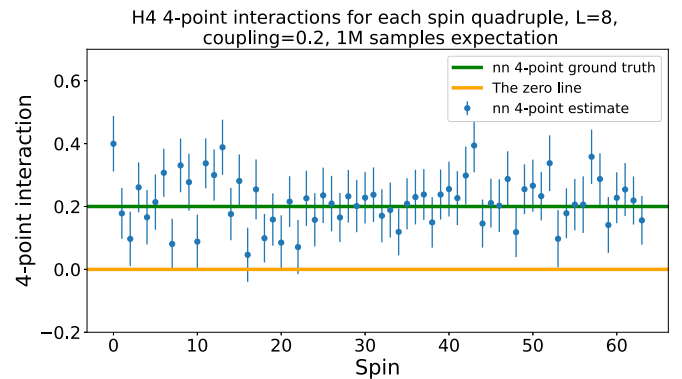


FIG. 17. Four-point per spin estimates of interactions for the ground truth coupling constant 0.2. Estimations are performed using 1×10^6 samples. We observe that the variance is large, in the sense that if the ground truth were to be unknown, some of the couplings would be considered as insignificant.

interactions, i.e., the coefficients in the Hamiltonian up to statistics, e.g., as shown in Sec. VB numerically. In particular, given any parametric form p_θ , our formulation yields an analytical, closed form expression for all n -point interactions in terms of the parameters θ of the given model. For example, the restricted Boltzmann machine was dealt with in Sec. IV B. Note, however, that in such energy-based neural networks determining the n -point interaction is a two-step procedure: (i) marginalizing of the hidden (latent) variables to obtain the probability distribution in terms of the visible variables only, and (ii) replacing the probabilities p in Eq. (17) with the parametric form p_θ . Thanks to the targeted learning framework, the last step can be performed directly without the need for asymptotic expansions and resummations.

VI. CONCLUSIONS AND FUTURE WORK

In this work, we have provided a nonparametric solution to the inverse problem of estimating n -point interactions among binary and categorical random variables directly from data, using the framework of targeted learning. In doing so, no parametric assumptions have to be made, yielding a fully model-independent and unbiased estimator of interaction at all orders. We have shown that interaction can naturally be interpreted as a derivative and, more specifically, that n -point interactions are inductively interpretable as a *change* in $(n-1)$ -point interaction when fixing any one of the n variables. Under a Markovian assumption, which is satisfied by all energy-based models in statistical physics and machine learning, we have demonstrated that interaction can be efficiently estimated from data by only conditioning on parent variables. If the parent structure is known, or has been inferred from a nonparametric independence test, one can substantially reduce the sample size required to obtain an accurate estimate. Furthermore, as the estimator only consists of expectation values over the data, the run time on a local machine is of the order of a few minutes. We have illustrated the above both analytically and numerically on a two-dimensional Ising Hamiltonian, a four-point Ising-type Hamiltonian, and the distribution of a restricted Boltzmann machine. Moreover, we have argued that our formulation can be used to extract closed form expressions of n -point interaction in any system of binary and categorical random variables, such as energy-based neural networks, where this coupling cannot directly be read off from a Hamiltonian, e.g., due to multiple hidden nodes. Finally, we have indicated how our definition of interaction via targeted learning has applications in population biomedicine, in particular genome-wide association studies (GWAS), since it both removes the need for parametric assumptions altogether and correctly accounts for molecular interaction effects (epistasis), in contrast to current approaches in the literature.

In future work, we plan to examine the bias-variance trade-off in extracting n -point interactions from other generative networks, such as variational autoencoders (VAE) and generative adversarial networks (GAN).

ACKNOWLEDGMENTS

We are most grateful to M. van der Laan for his suggestions regarding the formulation of 2-variable interactions

using the targeted learning framework, in a private conversation at the *causal machine learning master class*, the Alan Turing Institute, London. We are thankful to L. Del Debbio for his comments on the numerical results, as well as A. Papanastasiou and A. Jansma for reading and commenting on the manuscript. We are also thankful to C. Ponting and N. Clark, for their insights into the biological applicability of our work. S.V.B. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy–EXC-2047/1–390685813. A.K. is a cross-disciplinary post-doctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (MC_UU_00009/2).

APPENDIX A: ADDITIVE INTERACTION FOR CATEGORICAL VARIABLES

We make the following remarks regarding Eq. (12) of additive two-point interaction for categorical variables.

(1) Similar to the notion of interaction in the binary case, the notion of interaction for categorical variables is inherently symmetric under the exchange of the variables ($T_1 : t_1 \rightarrow t'_1$) and ($T_2 : t_2 \rightarrow t'_2$), i.e.,

$$I_{1,2}^a(t_1 t'_1; t_2 t'_2) = I_{2,1}^a(t_2 t'_2; t_1 t'_1). \quad (\text{A1})$$

(2) The interaction between the effect of $T_1 : t_1 \rightarrow t'_1$ on Y and $T_2 : t_2 \rightarrow t'_2$ on Y is opposite in sign to the effect of $T_1 : t'_1 \rightarrow t_1$ on Y (we swap t_1 and t'_1) and $T_2 : t_2 \rightarrow t'_2$ on Y , i.e.,

$$I_{1,2}^a(t_1 t'_1; t_2 t'_2) = -I_{1,2}^a(t'_1 t_1; t_2 t'_2). \quad (\text{A2})$$

For example, the interaction between the effect of *turning on* variable $T_1 : 0 \rightarrow 1$ on Y and the effect of $T_2 : t_2 \rightarrow t'_2$ on Y , is opposite in sign to the interaction between the effect of *turning off* variable $T_1 : 1 \rightarrow 0$ on Y and the effect of $T_2 : t_2 \rightarrow t'_2$ on Y .

(3) As a result of the above remark, swapping both categories yields the same interaction

$$I_{1,2}^a(t_1 t'_1; t_2 t'_2) = (-1)^2 I_{1,2}^a(t'_1 t_1; t'_2 t_2). \quad (\text{A3})$$

Finally, the additive two-point interaction between categorical variables satisfies the following *transitivity*:

Proposition 1. Let T_1, T_2 be two categorical variables, let $\{0, 1, 2\}$ denote the labels of three categories of T_1 , and let $\{0, 1\}$ denote the labels of two categories of T_2 . Then, the interactions satisfy transitivity, i.e.,

$$I_{1,2}^a(01; 01) + I_{1,2}^a(12; 01) = I_{1,2}^a(02; 01). \quad (\text{A4})$$

Heuristically, the result states that the sum of the effect on Y of changing T_1 from 0 to 1 and then changing T_1 from 1 to 2 equals the effect on Y of changing T_1 from 0 to 2 directly. The same heuristic holds for the interaction with the effect of $T_2 : 0 \rightarrow 1$ on Y as this effect is the same during all three steps of the procedure.

Proof. We define the function $f : \{0, 1, 2\} \times \{0, 1\} \rightarrow \mathbb{R}$ as

$$f(t_1, t_2) := \mathbb{E}(Y | T_1 = t_1, T_2 = t_2, \underline{T} = 0). \quad (\text{A5})$$

We may express the average treatment effect in terms of f as $\text{ATE}_{T_1:t_1 \rightarrow t'_1}(Y | T_2 = t_2, \underline{T} = 0) = f(t'_1, t_2) - f(t_1, t_2)$. This leads to the following expression for the interaction in terms of f :

$$I_{1,2}^a(t_1 t'_1; t_2 t'_2) = [f(t'_1, t'_2) - f(t_1, t'_2)] - [f(t'_1, t_2) - f(t_1, t_2)]. \quad (\text{A6})$$

Equation (A4) now follows by writing out both sides:

$$\begin{aligned} I_{1,2}^a(01; 01) + I_{1,2}^a(12; 01) &= [f(1, 1) - f(0, 1)] - [f(1, 0) - f(0, 0)] \\ &\quad + [f(2, 1) - f(1, 1)] - [f(2, 0) - f(1, 0)] \\ &= [f(2, 1) - f(0, 1)] - [f(2, 0) - f(0, 0)] \\ &= I_{1,2}^a(02; 01). \end{aligned}$$

This completes the proof. \blacksquare

As an important corollary, we obtain a *criterion* for linear dependence of the interaction $I_{1,2}^a$ on particular labels of the categorical variables. The precise statement is the following.

Corollary 1. Let T_1, T_2 be two categorical variables, let $\{0, 1, 2\}$ denote the labels of three categories of T_1 , and let $\{0, 1\}$ denote the labels of two categories of T_2 . If

$$I_{1,2}^a(01; 01) = I_{1,2}^a(12; 01), \quad (\text{A7})$$

then the interaction $I_{1,2}^a(_; 01)$ between the effect of T_1 on Y and the effect of $T_2 : 0 \rightarrow 1$ on Y depends linearly on the label of the categorical variable T_1 , in the sense that

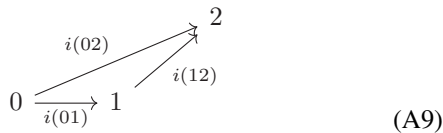
$$I_{1,2}^a(02; 01) = 2I_{1,2}^a(01; 01). \quad (\text{A8})$$

Thus, the 2 of the label 02 can be taken outside to multiply the interaction leaving the label 01, hence the term *linear*.

Proof. This follows directly from Proposition 1. \blacksquare

A similar statement holds for the interaction conditioned on a particular covariate $W = w$, and when interchanging the roles of T_1 and T_2 by considering two categories for T_1 and three for T_2 .

This result has a graphical interpretation in terms of the following triangle:



$$(A9)$$

where we denote the corresponding interaction by $i(t_1 t'_1) = I_{1,2}^a(t_1 t'_1; 01)$ which is represented by the length of the *vertical* component of the arrow. For example, in the above picture $i(01) = 0$ since the arrow is horizontal, and $i(12) = i(02)$ as the vertical components of both arrows have the same length. Thus, the transitive relation

$$i(01) + i(12) = i(02) \quad (\text{A10})$$

allows us to draw this triangle. Under the condition of Corollary 1, the vertical components of the arrow $0 \rightarrow 1$ and $1 \rightarrow 2$ are equal, i.e., $i(01) = i(12)$, in which case the above triangle is *degenerate*, i.e., a line segment. In conclusion, the linearity of the dependence on the categorical variable T_1 of the interaction $I_{1,2}^a(_; 01)$ between the effect of $T_1 : 0 \rightarrow 1$ on Y and

the effect of $T_2 : 0 \rightarrow 1$ on Y , in the sense that

$$i(02) = 2i(01) \quad (\text{A11})$$

corresponds to degeneracy of the above triangle. This is a geometrical criterion for linearity.

The notion of interaction as in Eq. (12) is *independent* of the chosen labels for the categorical random variables T_1, T_2 whether they be numbers, farm animals, or names of cabinet ministers. The *interpretation* of Eq. (A8) in terms of linearity *depends* on the chosen labels since it forces them to appear in the mathematical formula (A8). Naturally, the above discussion admits a direct generalization to the case of categorical variables describing more than three categories. In fact, all results are formulated in this general setting already, apart from assigning the particular labels $\{0, 1, 2\}$ or $\{0, 1\}$.

APPENDIX B: SYMMETRY OF n -POINT INTERACTION

In this Appendix, we prove the symmetry under any permutation of the variables X_{i_1}, \dots, X_{i_n} of the multiplicative formulation of n -point interaction.

Proposition 2. Let $K = \{i_1, \dots, i_n\} \subset \{0, 1, \dots, r\}$ be a subset of indices, and let σ be any of the $n!$ permutations of $\{1, 2, \dots, n\}$ that acts on the n -tuple K as $\sigma(K) = \sigma(i_1, \dots, i_n) = \{i_{\sigma(1)}, \dots, i_{\sigma(n)}\}$. Then, we have

$$I_{i_1, \dots, i_n}^m = I_{\sigma(i_1), \dots, \sigma(i_n)}^m. \quad (\text{B1})$$

Proof. Let $J \subset K$ be a subset of j indices and recall that $e_J^{(n)} = (e_{i_1}, \dots, e_{i_n})$ is the unique n -tuple such that $e_{i_j} = 1$ if $i_j \in J$ and $e_{i_j} = 0$ otherwise; in particular, this n -tuple contains j ones and $n - j$ zeros. The same property holds for the n -tuple $e_{\sigma(J)}^{(n)}$, where σ is any permutation of K . As a result, it suffices to show that σ satisfies

$$I_{\sigma(i_1, \dots, i_n)}^m(j) = I_{i_1, \dots, i_n}^m(j), \quad (\text{B2})$$

where

$$I_{i_1, \dots, i_n}^m = \prod_{j=0}^n I_{i_1, \dots, i_n}^m(j), \quad (\text{B3})$$

i.e., that it fixes the $n + 1$ factors $I_{i_1, \dots, i_n}^m(j)$ of I_{i_1, \dots, i_n}^m separately. But any permutation of $K = \{i_1, \dots, i_n\}$ simply permutes all subsets $J \subset K$ of fixed length $\ell(J) = j$ among each other. This completes the proof. \blacksquare

As a corollary, we deduce the general permutation symmetry of the additive n -point interaction.

Corollary 2. Let $K = \{i_1, \dots, i_n\} \subset \{1, 2, \dots, r\}$ be a subset, and let σ be any of the $n!$ permutations of $\{1, 2, \dots, n\}$ acting on K as $\sigma(K) = \sigma(i_1, \dots, i_n) = \{i_{\sigma(1)}, \dots, i_{\sigma(n)}\}$. The additive n -point interaction satisfies

$$I_{i_1, \dots, i_n}^a = I_{\sigma(i_1), \dots, \sigma(i_n)}^a. \quad (\text{B4})$$

Proof. For the outcome $Y = -E(\underline{X})$, this follows directly by combining Eqs. (22) and (24). For a general outcome Y , it follows by the argument of Proposition 2. \blacksquare

APPENDIX C: HAMMERSLEY-CLIFFORD THEOREM FOR THE ISING MODEL

Recall the two-dimensional Ising model of spins $\{v_i\}$ taking on the value ± 1 . As an example, we explicitly establish the

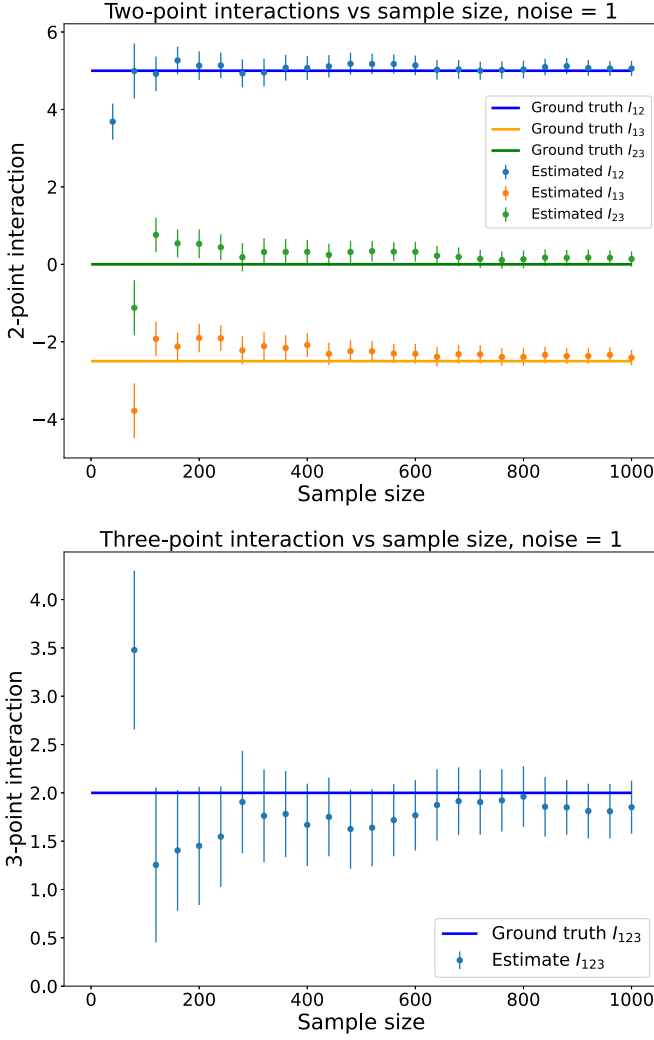


FIG. 18. Estimates of two-point (top) and three-point (bottom) interaction as a function of sample size, with noise $\sigma^2 = 1$. The uncertainties on the estimates are derived using statistical bootstrap. See Fig. 19 in Appendix E for a comparison of bin sizes for each of the expectation values as the total sample size increases.

Hammersley-Clifford theorem of Sec. II F in this case by verifying that its Hamiltonian

$$p(\mathbf{v}) = \frac{1}{Z(T)} e^{-E(\mathbf{v})} \quad \text{where } E(\mathbf{v}) = - \sum_{i,j} J_{i,j} v_i v_j \quad (\text{C1})$$

from Eq. (34) is locally, and hence globally, Markovian. To do so, we denote by \mathcal{N} the set of all spins in the system, by \mathcal{N}_i the set of (four) spins neighboring spin i , and we denote by \mathcal{N}_{-i} the set of all spins in the system apart from spin i . The probability p is locally Markovian if we have the equality

$$p(v_i = \pm 1 \mid v_j \text{ for } j \neq i) = p(v_i = \pm 1 \mid v_j \text{ for } j \in \mathcal{N}_i), \quad (\text{C2})$$

for each $i \in \mathcal{N}$. Fix a spin v_0 and denote its neighbors by $\mathcal{N}_0 = \{v_1, v_2, v_3, v_4\}$. We will check that in the conditional probability on the left-hand side of Eq. (C2), one only needs

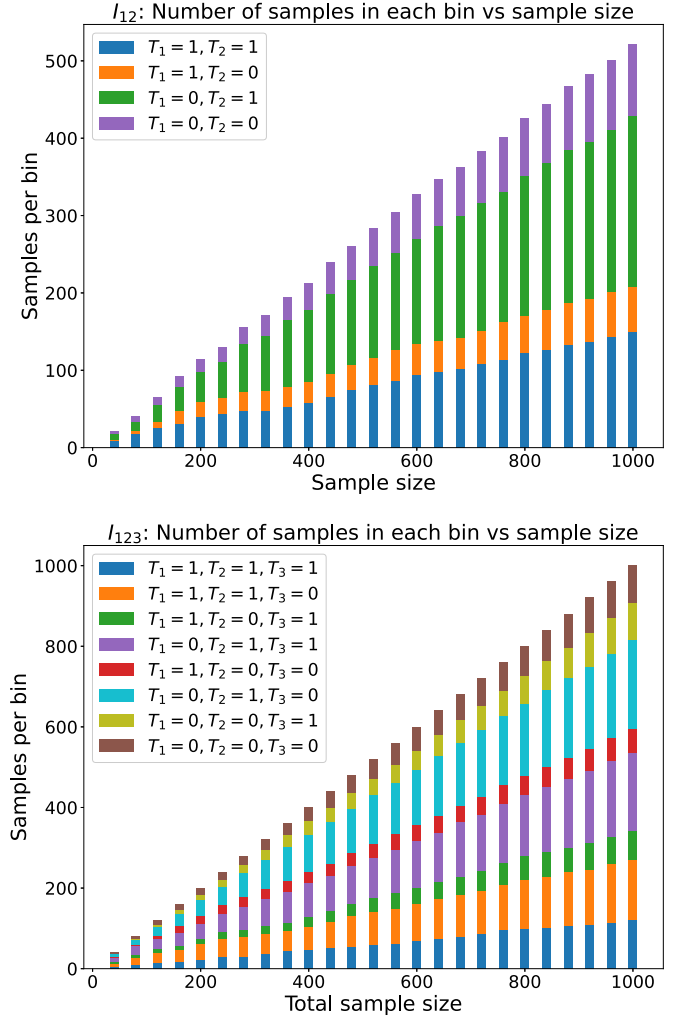


FIG. 19. Number of samples for each of the expectation values vs total sample size. Top: for the two-point interaction I_{12} . The variables are distributed as $T_1 \sim \text{Binom}(0.4)$ and $T_2 \sim \text{Binom}(0.7)$ so that, e.g., the bin size of $(T_1, T_2) = (1, 0)$ is the smallest, whereas the one of $(T_1, T_2) = (0, 1)$ is the largest. Bottom: for the three-point interaction I_{123} , where $T_3 \sim \text{Binom}(0.5)$. The legend $T_1 = T_2 = T_3 = 1$ and $T_1 = T_2 = T_3 = 0$ are placed lowest and highest in the bar plot, respectively.

to condition on the spins v_1, v_2, v_3, v_4 . Here,

$$p(v_0, v_j \mid j \in \mathcal{N}_{-0}) = \frac{1}{Z(T)} e^{-\sum_{i,j \neq 0} J_{i,j} v_i v_j} e^{-v_0 \sum_{i=1}^4 (J_{0,i} v_i + J_{i,0} v_i)},$$

$$p(v_j \mid j \in \mathcal{N}_{-0}) = \frac{1}{Z(T)} e^{-\sum_{i,j \neq 0} J_{i,j} v_i v_j} \times [e^{-\sum_{i=1}^4 (J_{0,i} v_i + J_{i,0} v_i)} + e^{\sum_{i=1}^4 (J_{0,i} v_i + J_{i,0} v_i)}].$$

It follows that their ratio, which is by definition the binary probability distribution $p(v_0 \mid v_j \text{ for } j \neq i)$, is fully determined once one conditions on the four nearest neighbor spins v_1, v_2, v_3, v_4 of v_0 . This proves the claim.

APPENDIX D: LINEAR REGRESSION

Let us consider the regression model with quadratic and cubic terms, representing additive two- and three-point

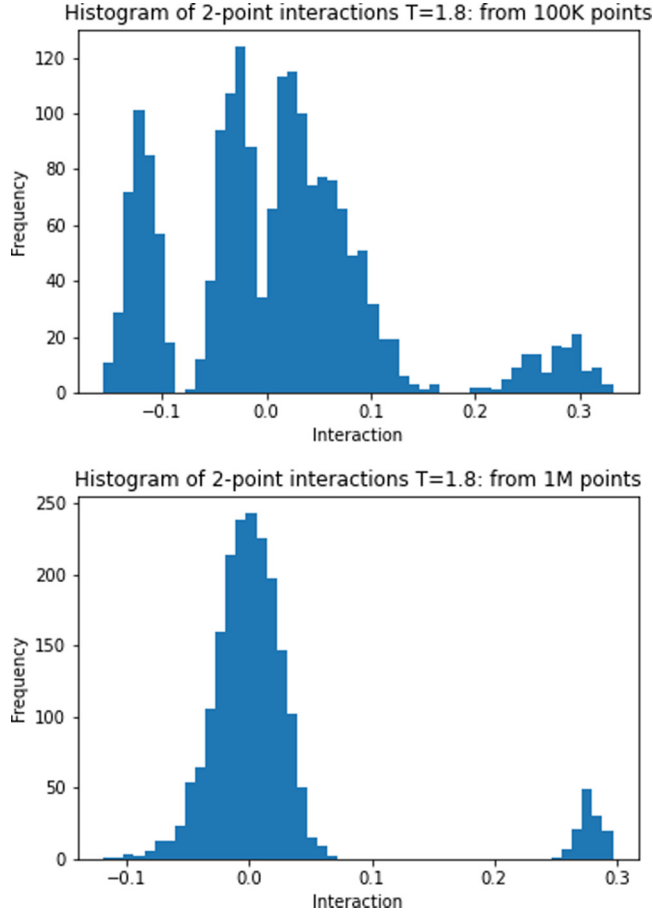


FIG. 20. Histograms of 100 000 (top) and 1×10^6 (bottom) estimates of the two-point interaction at $T = 1.8$, in an Ising system of size $L^2 = 8^2$. The interactions are computed directly from the data using the nonparametric multiplicative formulation in Eq. (14). As expected, with larger sample sizes, the peaks corresponding to non-nearest neighbor interactions, around zero, and nearest neighbor interactions, around $\frac{1}{27} \approx 0.28$, become more distinct with less noise.

interactions among the effects of the binary random variables T_1 , T_2 , and T_3 on Y :

$$Y = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_3 + \alpha_{12} T_1 T_2 + \alpha_{13} T_1 T_3 + \alpha_{23} T_2 T_3 + \gamma T_1 T_2 T_3 + \epsilon. \quad (\text{D1})$$

The noise term ϵ is normally distributed $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$. Without loss of generality, the ground truth three-point interaction γ is set to twice the value of the noise, i.e., $\gamma = 2$, while the two-point interactions are set to $\alpha_{12}, \alpha_{13}, \alpha_{23} = 5.0, -2.5, 0$, respectively. The zeroth order coefficient $\alpha_0 = -1.5$ and the linear coefficient are set to $\alpha_1, \alpha_2, \alpha_3 = -2, 10, 0$. We generate $N_s = 40, 80, \dots, 1000$ samples with $T_1 \sim \text{Binom}(0.4)$, $T_2 \sim \text{Binom}(0.7)$, $T_3 \sim \text{Binom}(0.5)$, where we have fixed regression coefficients to be as above. We then take as input (Y, T_1, T_2, T_3) , and compute the expectation values in Eq. (10) to estimate the two- and three-point interactions, for varying sample sizes N_s , and compare with the ground truth values used to generate the data.

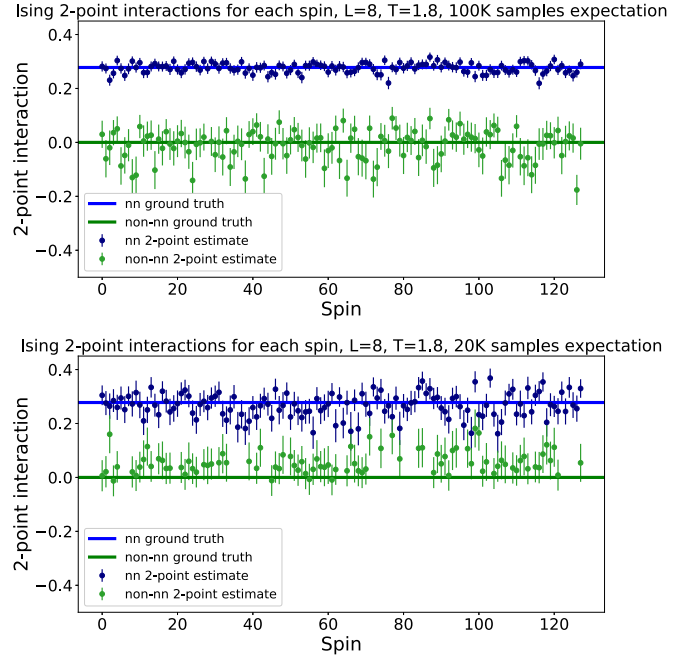


FIG. 21. $L^2 = 8^2$, $T = 1.8$, with conditioning on the nearest neighbors to estimate I_{ij}^m for both nearest and non-nearest neighbors. In order to reduce clutter, the same number of non-nearest couplings as nearest neighbors are shown (128). No translational invariance is used. Top: The results over 100 000 samples, using Eq. (16) and statistical bootstrap, as compared to bottom: the results over 20 000 samples. For the latter, approximately 30% of spins had no samples in the p_{11} bin. This is due to the fact that it is very rare to find two spins having value one, while their eight nearest neighbors all have spin value 0, particularly at cold temperatures, as the total sample size becomes smaller.

In order to ensure the estimates are robust, sufficiently many subsamples have to be available for estimating each of the four conditional expectation values appearing in Eq. (10). As with any statistical estimator, having very few samples for one of the conditional expectation values may result in unstable estimates of the expectation value and its variance. This will in turn introduce instabilities in the estimates of the interactions. See Appendix E for a comparison of bin sizes for each of the expectation values as the total sample size increases.

The three two-point interactions and the three-point interaction among variables T_1, T_2, T_3 are presented in Fig. 18. The uncertainties on the estimates are derived using statistical bootstrap [25]. One can readily observe that as the sample size increases, the estimates converge to the correct value with smaller variance as expected.

APPENDIX E: LINEAR REGRESSION: BIN SIZES AS A FUNCTION OF SAMPLE SIZE

In Fig. 19 we plot the bin sizes for each of the four expectation values appearing in Eq. (3) as the sample size grows. When the total sample size is, e.g., $N_s = 40$, some of the conditional expectation values are estimated using one or two samples only and thus are unreliable.

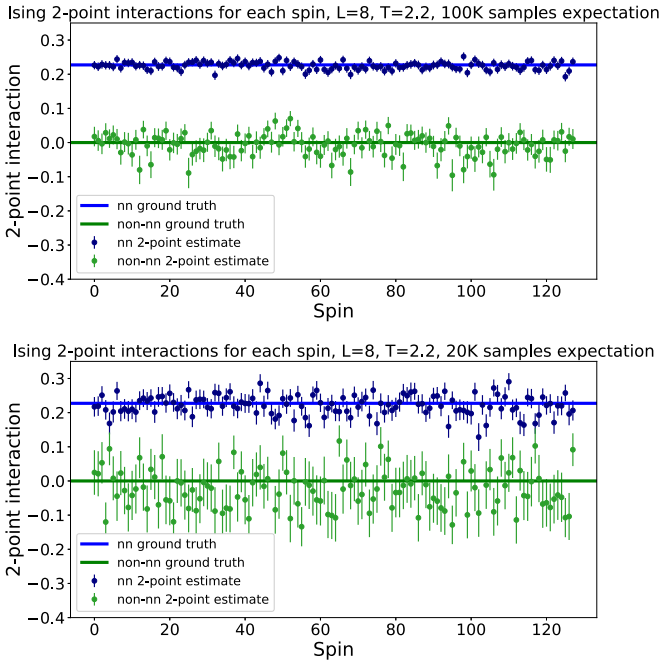


FIG. 22. $L^2 = 8^2$, $T = 2.2$, with conditioning on the nearest neighbors to estimate I_{ij}^m for both nearest and non-nearest neighbors. In order to reduce clutter, the same number of non-nearest couplings as nearest neighbors are shown (128). No translational invariance is used. Top: the results over 100 000 samples, using Eq. (16) and statistical bootstrap, as compared to bottom: the results over 20 000 samples. At 20 000 samples we have power to accurately estimate approximately 98% of non-nearest neighbor spin pairs.

APPENDIX F: INTERACTION ESTIMATES PER SPIN PAIR FOR THE ISING MODEL

We present the histogram of two-point interactions among all pairs of (non)-nearest neighbors, using Eq. (14) for Ising states simulated at temperature $T = 1.8$ and $L^2 = 8^2$. As follows from Fig. 20, as the total sample size increases the two peaks corresponding to zero couplings between non-nearest

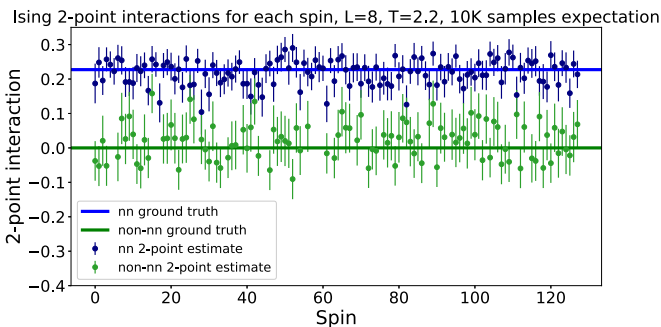


FIG. 23. $L^2 = 8^2$, $T = 2.2$, with conditioning on the nearest neighbors to estimate I_{ij}^m for (non)-nearest neighbors. In order to reduce clutter, the same number of non-nearest couplings as nearest neighbors are shown (128). Similar to the results in Fig. 22, except the total sample size is now 10 000 only. There is enough power to accurately estimate I_{ij}^m for all nearest neighbor pairs, and approximately 83% of the non-nearest neighbor pairs. In contrast, e.g., the RBM does not train on 10 000 samples (see [12, Fig. 31]).

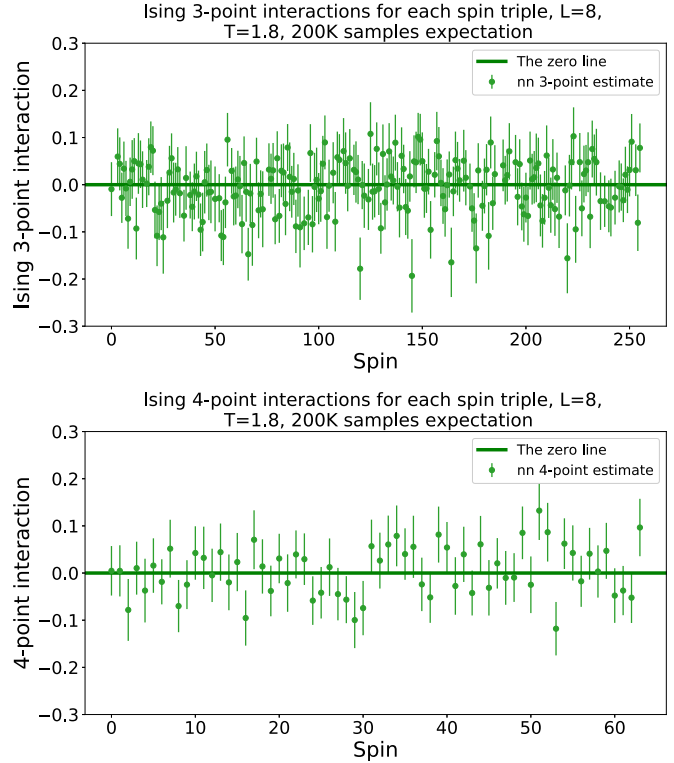


FIG. 24. $L^2 = 8^2$, $T = 1.8$, with conditioning on the nearest neighbors to estimate three-point (top) and four-point (bottom) interaction for the nearest neighbors. Due to the cold temperature, 85% of triples can be estimated, all four points are estimated. If 100 000 samples are used 40% of the three points can be estimated, but they are all accurately zero within statistics, similar to the top plot.

neighbor pairs and positive couplings at $\frac{1}{2T} \approx 0.28$ corresponding to the nearest neighbor pairs become more distinct.

The estimates of two-point couplings for both the nearest neighbor and non-nearest neighbor spin pairs, using 100 000 (top) and 20 000 (bottom) sample sizes, are presented in Fig. 21. As mentioned in Sec. IV C, one can use smaller sample sizes to estimate the couplings at the cost of reduced power. For colder temperatures and small sample sizes,

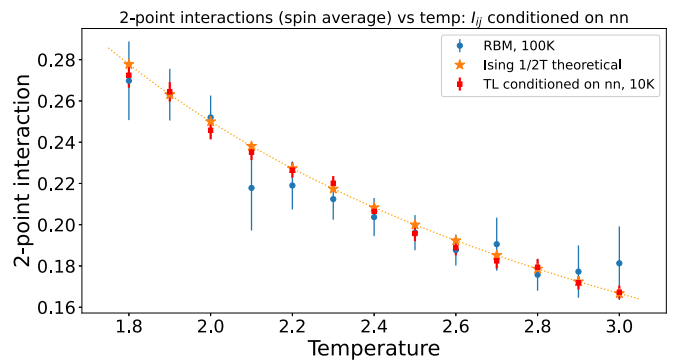


FIG. 25. Conditioning on the nearest neighbors to estimate I_{ij}^m substantially improves the estimates as compared to Fig. 4. The square points are estimations of interactions and their uncertainty using TL with 10 000 samples. The run time for each estimation using TL is at the order of a few seconds.

there may be no states in the p_{11} bin, for the case of non-nearest neighbor spin pairs. For $T = 1.8$ over 20K samples, we have power to accurately estimate all the nearest neighbor couplings, but only have power to accurately estimate approximately 70% of couplings between non-nearest neighbor pairs. As expected, increasing the sample size to 100 K improves the latter to 99%. Note that with real data sets, one may have limitations on the sample size. For example, as shown in Figs. 22 and 23, the nonparametric estimator,

combined with conditional independence among the variables has nevertheless enabled us to obtain accurate estimates using 10 000 samples only. In contrast, e.g., the RBM does not train well on Ising data with 10 000 samples (see [12, Fig. 31]). Individual vanishing per spin triplet and quadruplet three- and four-point interactions are presented in Fig. 24, for $T = 1.8$.

Figure 25 illustrates the estimates for nearest neighbor interactions vs temperature with 10 000 total samples using the TL framework.

-
- [1] H. Chau Nguyen, R. Zecchina, and J. Berg, Inverse statistical problems: From the inverse Ising problem to data science, *Adv. Phys.* **66**, 197 (2017).
- [2] A. Decelle and F. Ricci-Tersenghi, Pseudolikelihood Decimation Algorithm Improving the Inference of the Interaction Network in a General Class of Ising Models, *Phys. Rev. Lett.* **112**, 070603 (2014).
- [3] E. Aurell and M. Ekeberg, Inverse Ising Inference Using All the Data, *Phys. Rev. Lett.* **108**, 090201 (2012).
- [4] P. Ravikumar, Martin J. Wainwright, and John D. Lafferty, High-dimensional Ising model selection using ℓ_1 -regularized logistic regression, *Ann. Statist.* **38**, 1287 (2010).
- [5] H. Kiwata, Simple method for inference in inverse Ising problem using full data, *Phys. A (Amsterdam)* **436**, 321 (2015).
- [6] F. Lu, M. Zhong, S. Tang, and M. Maggioni, Nonparametric inference of interaction laws in systems of agents from trajectory data, *Proc. Natl. Acad. Sci. USA* **116**, 14424 (2019).
- [7] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. (Cambridge University Press, New York, 2009).
- [8] Thuc D. Le, T. Hoang, J. Li, L. Liu, H. Liu, and S. Hu, A fast PC algorithm for high dimensional causal discovery with multi-core PCs, *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **16**, 1483 (2019).
- [9] J. Kuipers and G. Moffa, Partition MCMC for inference on acyclic digraphs, *J. Am. Stat. Assoc.* **112**, 282 (2017).
- [10] C. Glymour, K. Zhang, and P. Spirtes, Review of causal discovery methods based on graphical models, *Front. Genetics* **10**, 524 (2019).
- [11] M. J. van der Laan and S. Rose, *Targeted Learning*, Springer Series in Statistics (Springer, New York, 2011).
- [12] G. Cossu, Luigi Del Debbio, T. Giani, A. Khamseh, and M. Wilson, Machine learning determination of dynamical parameters: The Ising model case, *Phys. Rev. B* **100**, 064304 (2019).
- [13] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLoS Medicine* **12**, (2015).
- [14] 10X Genomics, Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium Single Cell 3' Solution (<https://support.10xgenomics.com/single-cell-gene-expression/datasets>).
- [15] HCA DCP Data Portal, <https://data.humancellatlas.org/>.
- [16] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, Cambridge, 2015).
- [17] J. M. Hammersley and P. Clifford, Markov fields on finite graphs and lattices, Oxford University, 1971 (unpublished).
- [18] G. R. Grimmett, A theorem about random fields, *Bull. London Math. Soc.* **5**, 81 (1973).
- [19] X. Liu, Y. I. Li, and Jonathan K. Pritchard, Trans effects on gene expression can drive omnigenic inheritance, *Cell* **177**, 1022 (2019).
- [20] M. Claussnitzer, J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. E. Kenny, C. M. Lindgren, D. G. MacArthur, K. N. North, S. E. Plon, H. L. Rehm, N. Risch, C. N. Rotimi, J. Shendure, N. Soranzo, and M. I. McCarthy, A brief history of human disease genetics, *Nature (London)* **577**, 179 (2020).
- [21] E. A. Boyle, Y. I. Li, and J. K. Pritchard, An expanded view of complex traits: From polygenic to omnigenic, *Cell* **169**, 1177 (2017).
- [22] A. Fischer and C. Igel, Training restricted Boltzmann machines: An introduction, *Pattern Recognit.* **47**, 25 (2014).
- [23] Magneto: 2D Ising model in C++, <https://github.com/s9w/magneto>
- [24] T. B. Berrett and R. J. Samworth, Nonparametric independence testing via mutual information, *Biometrika* **106**, 547 (2019).
- [25] B. Efron, Computers and the theory of statistics: Thinking the unthinkable, *SIAM Rev.* **21**, 460 (1979).