Conditional Independence Testing using Generative Adversarial Networks

Alexis Bellot, Mihaela van der Schaar

University of Cambridge and The Alan Turing Institute

Cambridge, United Kingdom

ab2583@cam.ac.uk, mschaar@turing.ac.uk

Abstract

We consider the hypothesis testing problem of detecting conditional dependence, with a focus on high-dimensional feature spaces. Our contribution is a new test statistic based on samples from a generative adversarial network designed to approximate directly a conditional distribution that encodes the null hypothesis, in a manner that maximizes power (the rate of true negatives). We show that such an approach requires only that density approximation be viable in order to ensure that we control type I error (the rate of false positives); in particular, no assumptions need to be made on the form of the distributions or feature dependencies. Using synthetic simulations with high-dimensional data we demonstrate significant gains in power over competing methods. In addition, we illustrate the use of our test to discover causal markers of disease in genetic data.

1   Introduction

Conditional independence tests are concerned with the question of whether two variables X and Y behave independently of each other, after accounting for the effect of confounders Z. Such questions can be written as a hypothesis testing problem:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y | Z \quad \text{versus} \quad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y | Z$$

Tests for this problem have recently become increasingly popular in the Machine Learning literature [23, 29, 22, 21, 9] and find natural applications in causal discovery studies in all areas of science [15, 18]. An area of research where such tests are important is genetics, where one problem is to find genomic mutations directly linked to disease for the design of personalized therapies [31, 14]. In this case, researchers have a limited number of data samples to test relationships even though they expect complex dependencies between variables and often high-dimensional confounding variables Z. In settings like this, existing tests may be ineffective because the accumulation of spurious correlations from a large number of variables makes it difficult to discriminate between the hypotheses. As an example the work in [20] shows empirically that kernel-based tests have rapidly decreasing power with increasing data dimensionality.

In this paper, we present a test for conditional independence that relies on a different set of assumptions that we show to be more robust for testing in high-dimensional samples (X, Y, Z). In particular, we show that given only a viable approximation to a conditional distribution one can derive conditional independence tests that are approximately valid in finite samples and that have non-trivial power. Our test is based on a modification of Generative Adversarial Networks (GANs) [11] that simulates from a distribution under the assumption of conditional independence, while maintaining good power in high dimensional data. In our procedure, after training, the first step involves simulating from our network to generate data sets consistent with $\mathcal{H}_0$. We then define a test statistic to capture the $X - Y$ dependency in each sample and compute an empirical distribution which approximates the behaviour

of the statistic under $\mathcal{H}_0$ and can be directly compared to the statistic observed on the real data to make a decision.

The paper is outlined as follows. In section 2, we provide an overview of conditional hypothesis testing and related work. In section 3, we provide details of our test and give our main theoretical results. Sections 4 and 5 provide experiments on synthetic and real data respectively, before concluding in Section 6.

2 Background

We start by introducing our notation and define central notions of hypothesis testing. Throughout, we will assume the observed data consists of n $i.i.d$ tuples (X_i, Y_i, Z_i), defined in a potentially high-dimensional space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, typically $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_z}$. Conditional independence tests statistics $T : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ summarize the evidence in the observational data against the hypothesis $\mathcal{H}_0 : X \perp\!\!\!\perp Y | Z$ in a real-valued scalar. Its value from observed data, compared to a defined threshold then determines a decision of whether to reject the null hypothesis $\mathcal{H}_0$ or not reject $\mathcal{H}_0$. Hypothesis tests can fail in two ways:

- Type I error: rejecting $\mathcal{H}_0$ when it is true.
- Type II error: not rejecting $\mathcal{H}_0$ when it is false.

We define the p-value of a test as the probability of making a type I error, and its power as the probability of correctly rejecting H_0 (that is 1 - Type II error). A good test requires the p-value to be upper-bounded by a user defined significance level α (typically $\alpha = 0.05$) and seeks maximum power. Testing for conditional independence is a challenging problem. Shah et al. [24] showed that no conditional independence test maintains non-trivial power while controlling type I error over any null distribution. In high dimensional samples (relative to sample size), the problem of maintaining good power is exacerbated by spurious correlations which tend to make X and Y appear independent (conditional on Z) when they are not.

2.1 Related work

A recent favoured line of research has characterized conditional independence in a **reproducing kernel Hilbert space** (RKHS) [29, 9]. The dependence between variables is assessed considering all moments of the joint distributions which potentially captures finer differences between them. [29] uses a measure of partial association in a RKHS to define the KCIT test with provable control on type I error asymptotically in the number of samples. Numerous extensions have also been proposed to remedy high computational costs, such as [25] that approximates the KCIT with random Fourier features making it significantly faster. Computing the limiting distribution of the test becomes harder to accurately estimate in practice [29], and different bandwidth parameters give widely divergent results with dimensionality [20], which affects power.

To avoid tests that rely on asymptotic null distributions, **sampling strategies** consider explicitly estimating the data distribution under the null assumption $\mathcal{H}_0$. Permutation-based methods [9, 21, 4, 23] follow this approach. To induce conditional independence, they select permutations of the data that preserve the marginal structure between X and Z, and between Y and Z. For a set of continuous conditioning variables and for sizes of the conditioning set above a few variables, the "similar" examples (in Z) that they seek to permute are hard to define as common notions of distance increase exponentially in magnitude with the number of variables. The approximated permutation will be inaccurate and its computational complexity will not be manageable for use in practical scenarios. As an example, [9] constructs a permutation P that enforces invariance in Z ($PZ \approx Z$) while [21] uses nearest neighbors to define suitable permutation sets.

We propose a different sampling strategy building on the ideas proposed by [5] that introduce the conditional randomization test (CRT). It assumes that the conditional distribution of X given Z is known under the null hypothesis (in our experiments we will assume it to be Gaussian for use in practice). The CRT then compares the known conditional distribution to the distribution of the observed samples of the original data using summary statistics. Instead we require a weaker assumption, namely having access to a viable approximation, and give an approximately valid test that does not depend on the dimensionality of the data or the distribution of the response Y; resulting in a
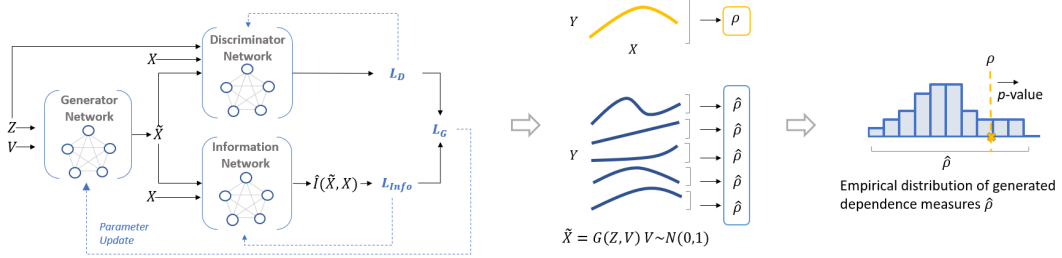
Figure 1: Illustration of conditional independence testing with the GCIT. A generator G is optimized by adversarial training to estimate the conditional distribution $X|Z$ under $\mathcal{H}_0$. We then use G to generate synthetic samples of $\tilde{X}$ under the estimated conditional distribution. Multiple draws are taken for each configuration Z and a measure of dependence between generated $\tilde{X}$ and Y, $\hat{\rho}$, is computed. The sequence of synthetic $\hat{\rho}$ is subsequently compared to the original sample statistic ρ to get a p-value.

non-parametric alternative to the CRT. [4] also expands the CRT by proposing a permutation-based approach to density estimation. **Generative adversarial networks** have been used for hypothesis testing in [22]. In this work, the authors use GANs to model to the data distribution and fit a classification model to discriminate between the true and estimated samples. The difference with our test is that they provide only a loose characterization of their test statistic's distribution under $\mathcal{H}_0$ using Hoeffding's inequality. As an example of how this might impact performance is that Hoeffding's inequality does not account for the variance in the data sample which biases the resulting test. A second contrast with our work is that we avoid estimating the distribution exactly but rather use the generating mechanism directly to inform our test.

3 Generative Conditional Independence Test

Our test for conditional independence, the GCIT (short for Generative Conditional Independence Test), compares an observed sample with a generated sample equal in distribution if and only if the null hypothesis holds. We use the following representation under $\mathcal{H}_0$,

$$Pr(X|Z,Y) = Pr(X|Z) \sim q_{\mathcal{H}_0}(X) \tag{1}$$

On the right hand side the null model preserves the dependence structure of $Pr(X,Z)$ but breaks any dependency between X and Y. If actually there exists a direct causal link between X and Y then replacing X with a null sample $\tilde{X} \sim q_{\mathcal{H}_0}$ is likely to break this relationship.

Sampling repeatedly $\tilde{X}$ conditioned on the observed confounders Z results in an exchangeable sequence of generated triples $(\tilde{X}, Y, Z)$ and original data (X, Y, Z) under $\mathcal{H}_0$. In this context, any function ρ - such as a statistic $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ - chosen independently of the values of X applied to the real and generated samples preserves exchangeability. Hence the sequence,

$$\rho(X,Y,Z), \rho(\tilde{X}^{(1)},Y,Z), ..., \rho(\tilde{X}^{(M)},Y,Z) \tag{2}$$

is exchangeable under the null hypothesis $\mathcal{H}_0$, deriving from the fact that the observed data is equally likely to have arisen from any of the above. Without loss of generality, we assume that larger values of ρ are more extreme. The p-value of the test can be approximated by comparing the generated samples with the observed sample,

$$\sum_{m=1}^{M} \mathbf{1}\{\rho(\tilde{X}^{(m)},Y,Z) \geq \rho(X,Y,Z)\}/M \tag{3}$$

which can be made arbitrarily close to the true probability, $\mathbb{E}_{\tilde{X} \sim q_{\mathcal{H}_0}} \mathbf{1}\{\rho(\tilde{X},Y,Z) \geq \rho(X,Y,Z)\}$, by sampling additional features $\tilde{X}$ from $q_{\mathcal{H}_0}$. $\mathbf{1}$ is the indicator function. Figure 1 gives a graphical overview of the GCIT.

3.1 Generating samples from $q_{\mathcal{H}_0}$

In this section we describe a sampling algorithm that adapts generative adversarial networks [11] to generate samples $\tilde{X}$ conditional on high dimensional confounding variables Z. GANs provide a powerful method for general-purpose generative modeling of datasets by designing a discriminator D explicitly used as an adversary to train a generator G responsible for estimating $q_{\mathcal{H}_0} := Pr(X|Z)$. Over successive iterations both functions improve based on the performance of the adversarial player.

Our implementation is based on Energy-based generative neural networks introduced in [30] which if trained optimally, can be shown to minimize a measure of divergence between probability measures that directly relates to a theoretical bound shown in this section that underlies our method. Pseudo-code for the GCIT and full details on the implementation are given in Supplement C.

Discriminator. We define the discriminator as a function $D_\eta : \mathcal{X} \times \mathcal{Z} \to [0,1]$ parameterized by η that judges whether a generated sample $\tilde{X}$ from G is likely to be distributed as its real counterpart X or not, conditional on Z. We train the discriminator by gradient descent to minimize the following loss function,

$$\mathcal{L}_D := \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D_\eta(x,z) + \mathbb{E}_{\tilde{v} \sim p(v)} (1 - D_\eta(G_\phi(v,z),z) \tag{4}$$

where $G_\phi(z,v), v \sim p(v)$ is a synthetic sample from the generator (described below) and $x \sim q_{\mathcal{H}_0}$ is a sample from the data distribution under $\mathcal{H}_0$. Note that in contrast to [30] we set the image of D to lie in $(0,1)$ and include conditional data generation.

Generator. The generator, G, takes (realizations of) Z and a noise variable, V, as inputs and returns $\tilde{X}$, a sample from an estimated distribution $X|Z$. Formally, we define $G : \mathcal{Z} \times [0,1]^d \to \mathcal{X}$ to be a measurable function (specifically a neural network) parameterized by ϕ, and V to be d-dimensional noise variable (independent of all other variables). For the remainder of the paper, let us denote $\tilde{x} \sim \hat{q}_{\mathcal{H}_0}$ the generated sample under the model distribution implicitly defined by $\hat{x} = G_\phi(v,z), v \sim p(v)$. In opposition to the discriminator, G is trained to minimize

$$\mathcal{L}_G(D) := \mathbb{E}_{\tilde{x} \sim \hat{q}_{\mathcal{H}_0}} D_\eta(\tilde{x},z) - \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D_\eta(x,z) \tag{5}$$

We estimate the expectations empirically from real and generated samples.

3.2 Validity of the GCIT

The following result ensures that our sampling mechanism leads to a valid test for the null hypothesis of conditional independence.

Proposition 1 (Exchangeability) *Under the assumption that $X \perp\!\!\!\perp Y | Z$, any sequence of statistics $(\rho_i)_{i=1}^M$ functions of the generated triples $(\tilde{X}^{(m)}, Y, Z)_{m=1}^M$ is exchangeable.* □

Proof. All proofs are given in Supplement B.

Generating conditionally independent samples with a neural network preserves exchangeability of input samples and thus leads to a valid p-value, defined in eq. (3), for the hypothesis of conditional independence. Under the assumption that the conditional distribution $q_{\mathcal{H}_0}$ can be estimated exactly, this implies that we maintain an exact control of the type I error in finite samples. In practice however, limited amounts of data and noise will prevent us from learning the conditional distribution exactly.

In such circumstances we show below that the excess type I error - that is the proportion of false negatives reported above a specified tolerated level α - is bounded by the loss function $\mathcal{L}_G$; which, moreover, can be made arbitrarily close to 0 for a generator with sufficient capacity. We give this second result as a corollary of the GAN's convergence properties in Supplement B.

Theorem 1 *An optimal discriminator D^* minimizing $\mathcal{L}_D$ exists; and, for any statistic $\hat{\rho} = \rho(X,Y,Z)$, the excess type I error over a desired level α is bounded by $\mathcal{L}_G(D^*)$,*

$$Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0) - \alpha \leq \mathcal{L}_G(D^*) \tag{6}$$

where $c_\alpha := \inf\{c \in \mathbb{R} : Pr(\hat{\rho} > c) \leq \alpha\}$ is the critical value on the test's distribution and $Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0)$ is the probability of making a type I error. □

Theorem 1 shows that the GCIT has an increase in type I error dependent only on the quality of our conditional density approximation, given by the loss function with respect to the generator, even in the worst-case under *any* statistic ρ. For reasonable choices of ρ, robust to errors in the estimation of the conditional distribution, this bound is expected to be tighter. The *key* assumption to ensure control of the type I error, and therefore to ensure the validity of the GCIT, thus rests solely on our ability to find a viable approximation to the conditional distribution of $X|Z$. The capacity of deep neural networks and their success in estimating heterogeneous conditional distributions even in high-dimensional samples make this a reasonable assumption, and the GCIT applicable in a large number of scenarios previously unexplored.

3.3 Maximizing power

For a fixed sample size, conditional dependence $\mathcal{H}_1 : X \not\perp Y|Z$, is increasingly difficult to detect with larger conditioning sets (Z) as spurious correlations due to sample size make X and Y appear independent. To maximize power it will be desirable that differences between generated samples $\tilde{X}$ (under the model $Pr(X|Z)$) and observed samples X (distributed according to $Pr(X|Z,Y)$) be as apparent as possible. In order to achieve this we will encourage $\hat{X}$ and X to have low mutual information because irrespective of dimensionality, mutual information between distributions in the null and alternative relates directly to the hardness of hypothesis testing problems, which can be seen for example via Fano's inequality (section 2.11 in [27]). To do so, we investigate the use of the information network proposed in [2] and used in the context of feature selection in [13]. [2] propose a neural architecture and training procedure for estimating the mutual information between two random variables. We approximate the mutual information with a neural network $T_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, parameterized by θ, with the following objective function (to be maximized),

$$\mathcal{L}_{Info} := \sup_\theta \mathbb{E}_{p_{x,\tilde{x}}^{(n)}}[T_\theta] - \log \mathbb{E}_{p_x^{(n)} \times p_{\tilde{x}}^{(n)}}[\exp(T_\theta)] \tag{7}$$

We estimate T_θ in alternation with the discriminator and generator given samples from the generator in every iteration. We modify the loss function for the generator to include the mutual information and perform gradient descent to optimize the generator on the following objective,

$$\mathcal{L}_G(D) + \lambda \mathcal{L}_{Info} \tag{8}$$

$\lambda > 0$ is a hyperparameter controlling the influence of the information network. This additional term ($\lambda \mathcal{L}_{Info}$) encourages the generation of samples $\tilde{X}$ as independent as possible from the observed variables X such that the resulting differences (between $\tilde{X}$ and X) are truly a consequence of the *direct* dependence between X and Y rather than spurious correlations with confounders Z.

To provide some further intuition, one can see why generating data different than the sample observed in the alternative $\mathcal{H}_1$ might be beneficial by considering the following bound (proven in Supplement B),

$$\text{Type I error} + \text{Type II error} \geq 1 - \delta_{TV}(\hat{q}_{\mathcal{H}_0}, q_{\mathcal{H}_1}) \tag{9}$$

where $\hat{q}_{\mathcal{H}_0}$ is the estimated null distribution with the GCIT, $q_{\mathcal{H}_1}$ is the distribution under $\mathcal{H}_1$ and where δ_{TV} is the total variation distance between probability measures. This result suggests that when emphasizing the differences between the estimated samples and true samples from $\mathcal{H}_1$, which increases the total variation, can improve the overall performance profile of our test by reducing a lower bound on type I and type II errors.

Remark. The GCIT aims at generating samples whose conditional distribution matches the distribution of its real counterparts, but can be independent otherwise. It is that gap that the power maximizing procedure intends to exploit. In practice, there will be a trade-off between the objectives of the discriminator and information network but we found that setting $\lambda = 10$ in our experiments achieved good performance. It should be noted also that hyperparameter selection cannot be performed using cross-validation as we do not have access to ground truth and so the hyperparameters must be fixed a priori. We test configurations of λ in section 4.2.

3.4 Choice of statistic ρ

The bound on the type I error given in Theorem 1 holds for any choice of statistic ρ as it depends solely on the conditional distribution estimation. For choices of ρ less sensitive to spurious differences

between generated and true samples when the null $\mathcal{H}_0$ holds, the type I error is expected to be below this bound. We experimented with various dependence measures (between two samples) as choices for ρ. We consider the Maximum Mean Discrepancy [12], Pearson's correlation coefficient, the distance correlation (which measures both linear and nonlinear association, in contrast to Pearson's correlation), the Kolmogorov-Smirnov distance between two samples and the randomized dependence coefficient [17]. In our experiments we use the distance correlation and analyze performance using all other measures in Supplement A.

4   Synthetic data example

In this section we analyse the performance of the GCIT in a controlled fashion with synthetic data against a wide range of competing algorithms, illustrating the effects of different components of our method. We consider the CRT [5] with pre-specified Gaussian sampling distribution, whose parameters are estimated from data; the kernel-methods KCIT [29] and RCoT [25] with bandwith parameter estimated with the median of all pairwise distances between X and Y, a common choice in the literature; and the CCIT [23], which does not make prior assumptions on data distributions but was also not specifically designed for high-dimensional data.

When testing at level α, type I error should be as close as possible to α even though this might not be the case because of violated assumptions or approximations. An important consideration in our discussion of power as we increase the dimensionality of Z, is the choice of alternatives $\mathcal{H}_1$. For instance, if the strength of the dependency between X and Y increases, the hypothesis testing problem will be made artificially easier and bias our conclusions with regards to data dimensionality, as observed also in [20]. In every synthetic experiment, we maintain the mutual information between X and Y approximately constant by first generating data and second estimating the mutual information before deciding to draw a new dataset, if the mutual information disagrees with the previous draw, or otherwise proceed with testing. We estimate the mutual information with a Gaussian approximation, $MI(X, Y) = -\frac{1}{2}\log(1 - \hat{\rho}^2)$, where $\hat{\rho}$ is the linear correlation between X and Y.

4.1   Setup

We generate synthetic data according to the "post non-linear noise model" similarly to [29, 9, 25] that defines (X, Y, Z) under $\mathcal{H}_0$ and $\mathcal{H}_1$ as follows,

$$\mathcal{H}_0: \quad X = f(A_f Z + \epsilon_f), \quad Y = g(A_g Z + \epsilon_g) \tag{10}$$

$$\mathcal{H}_1: \quad Y = h(A_h Z + \alpha X + \epsilon_h) \tag{11}$$

The matrix dimensions of $A_{(.)}$ are such that X and Y are univariate, matrix entries as well as parameter α are generated at random in the interval $[0, 1]$, and lastly, the noise variables $\epsilon_{(.)}$ are 0 on average with variance 0.025. The distributions of X, Y and ϵ, and the complexity of dependencies via f, g and g will be tuned carefully to make performance comparisons in three settings:

(1) Multivariate Gaussian
We set f, g and h to be the identity functions which induces linear dependencies, $Z \sim \mathcal{N}(0, \sigma^2)$, and $X \sim \mathcal{N}(0, \sigma^2)$ under $\mathcal{H}_1$ which results in jointly Gaussian data under the null and the alternative. Such a setting matches the assumptions of all methods and the interest of this study will be to provide a baseline for more complex scenarios.

(2) Multivariate Laplace
Kernel choice has a large impact on power, as we demonstrate in this setting. In this case, we set f, g and h as before but use a Laplace distribution to generate Z and X. The RBF kernel in this case overestimates the "smoothness" of the data. This study highlights the robustness of the GCIT in comparison to kernel-based methods which is important since hyperparameters cannot be tuned by cross-validation.

(3) Arbitrary distributions
We set f, g and h to be randomly sampled from $\{x^3, \tanh x, \exp(-x)\}$, resulting in more complex distributions and variable dependencies. Here $Z \sim \mathcal{N}(0, \sigma^2)$, and $X \sim \mathcal{N}(0, \sigma^2)$ under $\mathcal{H}_1$. This is our most general setting which most faithfully resembles the complexities we can expect in real applications.
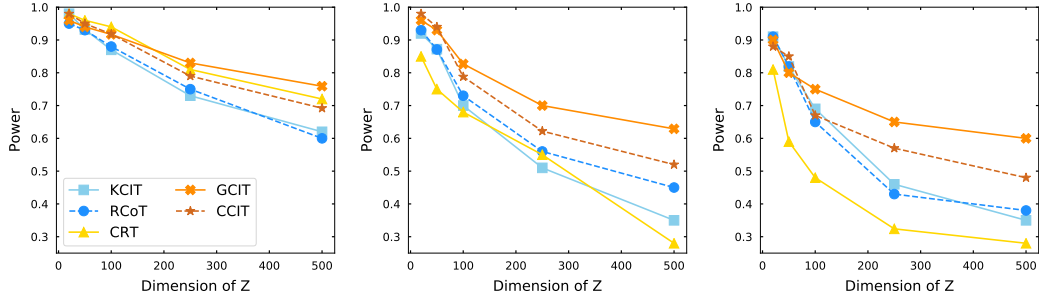
Figure 2: Power results of the synthetic simulations. (Higher better). **Left panel:** (1) Multivariate Gaussian, **Middle panel:** (2) Multivariate Laplace, **Right panel:** (3) Arbitrary distributions.

Results: Power as a function of the dimensionality of Z is shown in Figure 2. Each point on the curves is computed by taking averages over 1000 random experiments with sample size equal to 500 examples. The results from scenario **(1)** are consistent with our expectations; all methods perform comparably, the CRT and kernel-based methods achieving high power in lower dimensions while slightly under-performing in higher dimensions. In scenario **(2)** and **(3)**, the failure of the CRT and kernel-based methods is apparent while the GCIT maintains high power, even with increasing dimensionality, which demonstrates the robustness of our sampling mechanism to arbitrary complex data distributions. The CCIT outperforms kernel-based methods in these cases also. An important contrast of the GCIT with respect to the CCIT is our addition of the information network, which we argue contributes to the higher power observed across all experiments. We analyze this empirically below.

Figure 2 in Supplement A shows that type I error is approximately controlled at a level α for all methods. Observe also that even though the GCIT requires training a new GAN in every iteration, in Figure 3 Supplement A we show empirically that running times for the GCIT scale much better with dimensionality and sample size in comparison with the best benchmark, the CCIT: its running times are prohibitive in practice with more than 1000 samples or 500 dimensions in Z, with each test taking over $600s$ versus $60s$ for the GCIT.
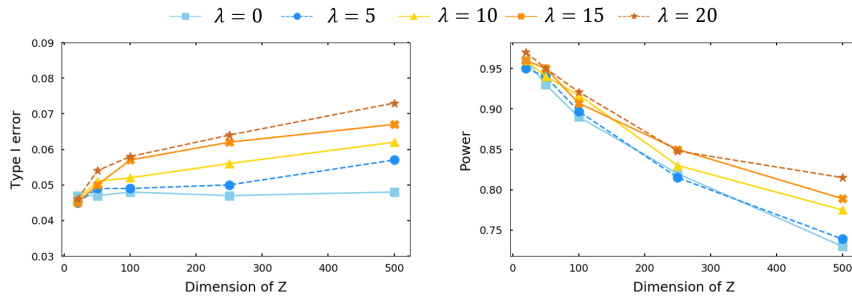

Figure 3: Type I error and power for different values of λ.

4.2 Source of gain: consequences of the information network

The information network aims to encourage maximum power in high-dimensional data. We control for its influence by varying λ in the loss function of the GCIT given in eq. (8). Higher values of λ encourage the generation of independent samples which improves power even though it might decrease the accuracy of the density approximation in the GAN optimization when the null in fact holds. We notice this trade-off between power and type I error for higher values of λ in Figure 3. The underlying data was generated from setting **(1)**, each curve in the two panels corresponds to a different value of λ. Lastly, we computed the lower-bound from GCIT generated samples and observed samples (by numerical integration) in eq. (9) to conclude that higher values of λ did decrease the lower bound, as expected.

7

5 Genetic data example

There is compelling evidence that the likelihood of a patient's cancer responding to treatment can be strongly influenced by alterations in the cancer genome [10]. We study the response of cancer cell lines to an anti-cancer drug where the problem is to distinguish between genetic mutations that influence directly the cancer cell line response from those that are not directly relevant [1, 26]. We use the subset of the CCLE data [1] relating to the drug PLX4720; it contains 474 cancer cell lines described by 466 genetic mutations. More details on the data can be found in Supplement D.

	AGREEMENT							DISAGREEMENT		
	BRAF.V600E	BRAF.MC	HIP1	CDC42BPA	THBS3	DNMT1	PRKD1	FLT3	PIP5K1A	MAP3K5
EN	1	3	4	7	8	9	10	5	19	78
RF	1	2	3	8	34	28	18	14	7	9
CRT	<0.001	<0.001	0.008	0.009	0.017	0.022	0.002	0.017	0.024	0.012
GCIT	<0.001	<0.001	0.008	0.050	0.013	0.020	0.002	0.521	0.001	<0.001

Figure 4: Genetic experiment results. Each cell gives the p-value or importance rank (where appropriate) indicating the dependency between a mutation and drug response.

Evaluating conditional independence relations from real data is difficult as we do not have access to the ground truth causal links. Instead we give our results in comparison to those of [1], who proceeded by reporting discriminative features returned by the parameter values of a fitted elastic net regression model (EN). This is common practice in genetic studies, see for example also [10]. In addition, we compare with the rank of each feature given by a random forest model importance scores (RF) and the p-value assigned by the CRT. The results for 10 selected mutations can be found in Figure 4. The first two rows give ranks of heuristic methods and the last two rows give p-values of conditional independence tests. We distinguish between the mutations where all methods agree (in the leftmost columns), and the mutations where not all methods agree (in the rightmost columns).

The mutations on genes PIP5K1A and MAP3K5 are recognized to be discriminative by the random forest model (high rank) and the GCIT (low p-value), which highlights the significance of the GCIT for conditional independence testing, suggesting that non-linear dependencies occur which are not captured by the elastic net or the CRT. For further evaluation, in this case we were able to cross-reference with a previous study to find evidence of the PIP5K1A gene to have a differential response on cancer cell lines when PLX4720 is applied [26]. The MAP3K5 gene has not previously been reported in the literature as being directly linked to the PLX4720 drug response, however [19] did find a proliferation of these gene mutations to be of BRAF *type* in cancer patients. This is interesting because PLX4720 is precisely designed as a BRAF inhibitor, and thus we would expect it to have an impact also on MAP3K5 mutations of the BRAF type. FLT3 is an interesting gene, found to be dependent on cancer response by the EN, RF and CRT, but not by the GCIT. This finding by the GCIT was confirmed however by a posterior genetic study [6] that established no link between cancer response and FLT3 mutations in the presence of PLX4720. Such results encourage us to believe that the GCIT is able to better detect dependence for these problems.

6 Conclusions and future perspectives

We propose a generative approach to conditional independence testing using generative adversarial networks. We show this approach results in an approximately valid test for an arbitrary data distribution irrespective of the number of variables observed. We have demonstrated through simulated data significant gains in statistical power, and we illustrated the application of our method to discover genetic markers for cancer drug response on real high-dimensional data. This work opens the door to principled statistical testing with more heterogeneous data, and expand our ability to reason and test variable relationships in more challenging scenarios.

Supplement A: Hyperparameter and Complexity Analysis

In this section we analyze hyperparameter configurations of the GCIT, present results on the type I error of the GCIT and all baseline algorithms for the synthetic simulations considered in the main body of this paper, and analyze computational complexity as a function of sample size and data dimensionality.

Choice of statistic ρ

We start by analyzing potential choices for the summary statistic $\rho : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \to \mathbb{R}$ that summarizes the generated and observed samples into a real-valued scalar. Different choices for ρ encode in more or less detail the distributional differences in samples and thus we can expect them to influence the resulting performance of the test. We considered the following distance and correlation measures between two samples:

- The Maximum mean discrepancy (MMD) is defined as the largest difference between the mean function values on two samples in a reproducing kernel Hilbert space. When MMD is large, the samples are likely from different distributions. For a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a consistent empirical estimate of the MMD is given by [12],

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \frac{1}{n^2} \sum_{i,j} k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$

- The Pearson's correlation coefficient (PCC) is a measure of linear correlation between two variables. It is defined as,

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})} \sqrt{\sum_i (y_i - \bar{y})}}$$

- The distance correlation (DC) measures both linear and nonlinear association between two random variables or random vectors. It is defined as,

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \frac{dCov(\mathbf{x}, \mathbf{y})}{\sqrt{dVar(\mathbf{x})dVar(\mathbf{y})}}$$

- The Kolmogorov-Smirnov statistic (KS) is defined as the sup-norm between cumulative distribution functions of two samples as follows,

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \sup_w |F_x^{(n)}(w) - F_y^{(n)}(w)|$$

  where $F_x^{(n)}$ and $F_y^{(n)}$ are the empirical distribution functions of X and Y samples respectively.

- The Randomized Dependence Coefficient (RDC) measures the dependence between random samples X and Y as the largest canonical correlation between k randomly chosen nonlinear projections of their copula transformations. It is formally defined an analyzed in [17].

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \sup_{\alpha, \beta} PCC(\alpha^T \Phi_{\mathbf{x}}, \beta^T \Phi_{\mathbf{y}})$$

  where PCC is Pearson's correlation coefficient and Φ are nonlinear random projections, such as sine or cosine projections. See [17] for more details.

We tested the above metrics with simulated data under setting **(3)** described in the main paper. Type I error and power results for the GCIT implemented with each one of the above choices for ρ are given in Figure 5. Finer differences are given by the MMD, the RDC or the DC that all consider non-linear relationships between variables; we see in the power computations in the right column that this results in higher power of the GCIT since the underlying data generating mechanism in non-linear. However, these statistics will also encode spurious differences between samples when the null $\mathcal{H}_0$ is in fact true, resulting in higher type I error. We can see this behaviour in the type I error results on the panels in the left column. The PCC, for example, that encodes only linear differences between samples is more robust to type I error.
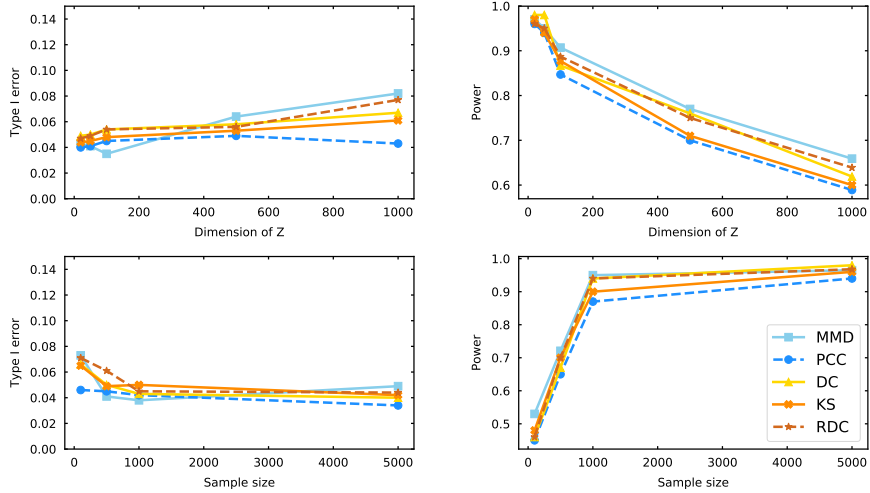
Figure 5: Power and type I error results for different choices of ρ.

Type I error versus dimensionality of Z

Next we show in Figure 6 type I error as a function of dimensionality of Z for each one of the three synthetic simulations considered in the main body of this paper. We observe that in all cases, type I error is approximately controlled at the chosen level $\alpha = 0.05$ when the distributional assumptions underlying each method holds. This is not the case otherwise, the CRT fails to control type I error in the non-linear setting when a Gaussian approximation to the joint distribution of the variables is not appropriate.
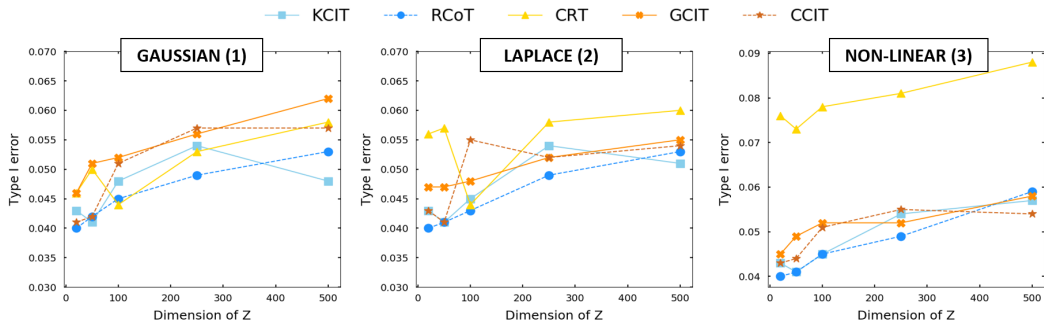

Figure 6: Type I error results for the synthetic simulations.

Computational complexity analysis

We give in Figure 7 the run times in seconds of all algorithms for a single conditional independence test for data generated under setting **(1)** in the main body of this paper. We vary both the number of samples (fixing the dimension of Z to 100) and the dimensionality of Z (fixing the sample size to 100). The GCIT scales very well with both sample size and conditioning set size, even if each iteration requires training a new GAN. In contrast, the running times of KCIT for sample sizes above 1000 and those of CCIT in higher dimensional samples are prohibitive in practice.
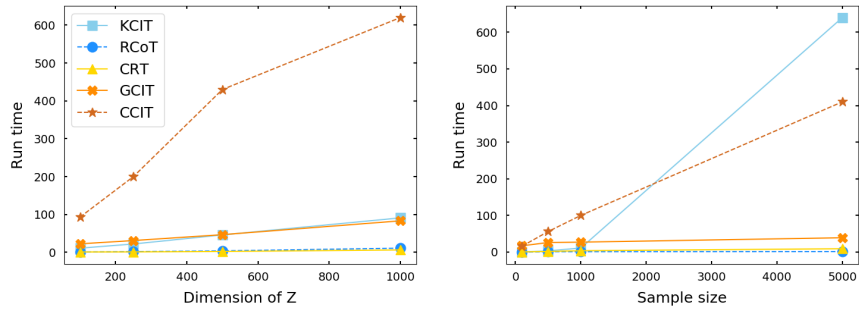
10

Figure 7: Running times in seconds as a function of sample size and dimension of Z.

Supplement B: Theoretical results

Proof of Proposition 1. A sequence of random variables is said to be exchangeable if its distribution is invariant under variable permutations. We make use of the "representation theorem" for exchangeable sequences of random variables, first stated by de Finetti and extended by Diaconis and Freedman for finite sequences [7, 8]. They show that every sequence of conditionally $i.i.d.$ random variables can be considered as a sequence of exchangeable random variables. With our definition of the generator we start from $i.i.d.$ sequence of noise random variables $\{V_m\}_{m=1}^M$ and define, for every m, $\tilde{X}^{(m)} = \phi(Z, V_m)$ where Z is a random variable independent of V_m and ϕ is a measurable function, such as a neural network in our case. By construction, the resulting random sequence of data sets $(\tilde{X}^{(m)}, Y, Z)_{m=1}^M$ is exchangeable and therefore also the sequence of statistics $(\rho_i)_{i=1}^M$ (measurable functions of $(\tilde{X}^{(m)}, Y, Z)_{m=1}^M$) is exchangeable. $\qquad\square$

The theoretical results that follow are proven only for the version of the generator loss given in equation (5) in the main body of this paper, $\mathcal{L}_G(D) := \mathbb{E}_{\tilde{x} \sim \hat{q}_{\mathcal{H}_0}} D_\eta(\tilde{x}) - \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D_\eta(x)$ though we do believe that the theorem holds more generally with the addition of the power maximizing procedure - this is backed up by our empirical results demonstrating Type I error control while using the power maximizing procedure. We prove the bound on the excess Type I error in two parts. First we show in the following lemma that an optimal discriminator exists, and second we prove the bound on the Type I error.

Lemma 1 *An optimal discriminator D^* minimizing $\mathcal{L}_D := \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D(x) + \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} (1 - D(x))$ over all measurable functions D such that $D \in (0,1)$ exists and it is given by $D^* = \frac{1}{2} sign \left(q_{\mathcal{H}_0} - \hat{q}_{\mathcal{H}_0} \right) + \frac{1}{2}$.*

Proof. To see this note first that the $sign(x)$ function is defined as $+1$ or -1 depending on the sign of x. Then,

$$\mathcal{L}_{D^*} = 1 + \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D^*(x) - \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} D^*(x) \tag{12}$$

$$= 1 + \int_{\mathcal{X}} q_{\mathcal{H}_0}(x) \frac{1}{2} \text{sign} \left(q_{\mathcal{H}_0} - \hat{q}_{\mathcal{H}_0} \right) dx - \int_{\mathcal{X}} \hat{q}_{\mathcal{H}_0}(x) \frac{1}{2} \text{sign} \left(q_{\mathcal{H}_0} - \hat{q}_{\mathcal{H}_0} \right) dx \tag{13}$$

$$= 1 - \frac{1}{2} \int_{x : \hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x) > 0} q_{\mathcal{H}_0}(x) - \hat{q}_{\mathcal{H}_0}(x) dx - \frac{1}{2} \int_{x : \hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x) < 0} \hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x) dx \tag{14}$$

$$= 1 - \frac{1}{2} \int_{\mathcal{X}} |\hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x)| dx \tag{15}$$

$$= 1 - \frac{1}{2} \sup_{||D||_\infty \leq 1} \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} D(x) - \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D(x) \tag{16}$$

$$= 1 - \sup_{||D||_\infty \leq 1/2} \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} D(x) - \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D(x) \tag{17}$$

$$= 1 - \sup_{0 \leq D \leq 1} \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} D(x) - \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D(x) \tag{18}$$

$$= \inf_{0 \leq D \leq 1} \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D(x) + \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} (1 - D(x)) \tag{19}$$

$$\leq \mathcal{L}_D \tag{20}$$

for any D in the mentioned space and with equality if and only if $D = D^*$. Eq (5) follows from the Kantorovich-Rubinstein dual representation for general f divergences, proven for example in [28]. $\qquad\square$

Corollary 1 For a generator G with infinite capacity converging to the true conditional distribution $q_{\mathcal{H}_0}(x)$, $\mathcal{L}_G(D^*)$ attains its minimum value of 0.

Proof By setting D^* in the loss of the generator $\mathcal{L}_G$ we observe that,

$$\mathcal{L}_G(D^*) = 1 - \mathcal{L}_{D^*} \tag{21}$$

$$= \frac{1}{2} \int_{\mathcal{X}} |\hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x)| dx \tag{22}$$

Hence, for a generator with infinite capacity converging to the true conditional distribution $q_{\mathcal{H}_0}(x)$, the last term is 0 which implies $\mathcal{L}_G(D^*) = 0$. $\qquad\square$

Proof of Theorem 1. Our derivation is similar to [4]. By definition the statistic $\hat{\rho}$ results in a p-value $p < \alpha$ if and only if the observed variable x is contained in the set $A_\alpha := \{x : \sum_{m=1}^M \mathbf{1}\{\rho(x^{(m)}, y, z) \geq \rho(x, y, z)\}/M < \alpha\}$. Consider generating a new sample $\tilde{x}$ from the generator G under the estimated conditional distribution and let $x \sim q_{\mathcal{H}_0}$ be sampled from the true conditional. Then it holds that,

$$\mathcal{L}_G(D^*) = \mathbb{E}_{x \sim \hat{q}_{\mathcal{H}_0}} D^*(x) - \mathbb{E}_{x \sim q_{\mathcal{H}_0}} D^*(x) \tag{23}$$

$$= \int_{\mathcal{X}} |q_{\mathcal{H}_0}(x) - \hat{q}_{\mathcal{H}_0}(x)| dx \tag{24}$$

$$= \sup_A |q_{\mathcal{H}_0}(A) - \hat{q}_{\mathcal{H}_0}(A)| \tag{25}$$

$$\geq Pr(x \in A_\alpha) - Pr(\tilde{x} \in A_\alpha) \tag{26}$$

$$\geq Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0) - \alpha \tag{27}$$

where by expanding the expectations, eq. (13) follows from similar arguments to those presented in Lemma 1. Next eq. (14) follows from a well known equivalent representation of the total variation divergence between probability measures, proven for example in Proposition 4.2, page 48 of [16]. Eq. (15) follows by standard properties of the supremum operator. Finally we arrive at eq. (16) given the fact that, by definition of A_α, $Pr(x \in A_\alpha) = Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0)$ and, since given y and z the set $(\tilde{x}, x^{(1)}, ..., x^{(M)})$ is conditionally independent and therefore exchangeable, we have that $Pr(\tilde{x} \in A_\alpha) \leq \alpha$. $\qquad\square$

Proof of equation (9). Let $q_{\mathcal{H}_0}$ and $q_{\mathcal{H}_1}$ be the true conditional distributions of $X|Z$ under the null hypothesis $H_0 : X \perp\!\!\!\perp Y | Z$ and its alternative $H_1 : X \not\perp\!\!\!\perp Y | Z$ respectively. Denote by A the event that samples x result in a p-value below the level α. Then,

$$\text{Type I error} + \text{Type II error} = q_{\mathcal{H}_0}(A) + q_{\mathcal{H}_1}(A^c) \tag{28}$$

$$= 1 + q_{\mathcal{H}_0}(A) - q_{\mathcal{H}_1}(A) \tag{29}$$

$$\geq 1 + \inf_A (q_{\mathcal{H}_0}(A) - q_{\mathcal{H}_1}(A)) \tag{30}$$

$$= 1 - \sup_A (q_{\mathcal{H}_1}(A) - q_{\mathcal{H}_0}(A)) \tag{31}$$

$$= 1 - \delta_{TV}(q_{\mathcal{H}_0}, q_{\mathcal{H}_1}) \tag{32}$$

$$\square$$

Supplement C: Implementation details

GCIT

In all our experiments we have set the depth of the generator, the discriminator and information network to 3. The number of hidden nodes in each layer is $d/10$ and $d/16$ for the generator and discriminator respectively (d the number of inputs). For the information network, we use 2 diagonal matrices for each layer to make two hidden nodes for each feature separately. We use ReLu and tanh as the activation functions for each layer except for the output layer where we use a linear activation function for the information network, and sigmoid activation function for the discriminator and generator network given that we require its output to be constrained in the $(0, 1)$ interval and re-scale the data in the $(0, 1)$ interval prior to training. The number of samples in each mini-batch is 128 for the synthetic experiments and 64 for the genetic experiment. The GCIT and all experiments have been implemented and carried out in tensorflow and python. Pseudocode for the GCIT is given in Algorithm 1.

Algorithm 1 GCIT

Input: batch size n_b, data $\mathcal{D} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ of size N, statistic ρ, iterations M, parameter λ
Initialize: neural network model parameters ϕ, η, θ
while convergence criteria not satisfied **do**
    1. **Update Discriminator**
    Sample $\mathbf{z}_1, ..., \mathbf{z}_{n_b}$ from $\mathcal{D}$ and $\mathbf{v}_1, ..., \mathbf{v}_{n_b} \sim p_v$ a batch from the real and latent samples
    $\tilde{\mathbf{x}}_i \leftarrow G_\phi(\mathbf{z}_i, \mathbf{v}_i)$ for $i = 1, ..., n_b$
    Update η by stochastic gradient descent with,

$$\nabla_\eta \frac{1}{n_b} \sum_{i=1}^{n_b} D_\eta(\mathbf{x}_i, \mathbf{z}_i) + (1 - D_\eta(\tilde{\mathbf{x}}_i, \mathbf{z}_i))$$

    2. **Update Information Network**
    Sample $\mathbf{z}_1, ..., \mathbf{z}_{n_b}$ from $\mathcal{D}$, $\mathbf{v}_1, ..., \mathbf{v}_{n_b} \sim p_v$ and κ a permutation of $1, ..., n_b$
    $\tilde{\mathbf{x}}_i \leftarrow G_\phi(\mathbf{z}_i, \mathbf{v}_i)$ for $i = 1, ..., n_b$
    Update θ by stochastic gradient ascent with,

$$\nabla_\theta \left(\frac{1}{n_b} \sum_{i=1}^{n_b} T_\theta(\tilde{\mathbf{x}}_i, \mathbf{x}_i) - \log[\frac{1}{n_b} \sum_{i=1}^{n_b} \exp(T_\theta(\tilde{\mathbf{x}}_i, \mathbf{x}_{\kappa(i)}))] \right)$$

    3. **Update Generator**
    Sample $\mathbf{z}_1, ..., \mathbf{z}_{n_b}$ from $\mathcal{D}$ and $\mathbf{v}_1, ..., \mathbf{v}_{n_b} \sim p_v$
    $\tilde{\mathbf{x}}_i \leftarrow G_\phi(\mathbf{z}_i, \mathbf{v}_i)$ for $i = 1, ..., n_b$
    Update ϕ by stochastic gradient descent with,

$$\nabla_\phi(\mathcal{L}_G(D) + \lambda \mathcal{L}_{Info.})$$

end while
for $m = 1, ..., M$ **do**
    Sample $v_1, ..., v_N \sim p_v$
    $\tilde{x}_j^{(m)} \leftarrow G_\phi(z_j, v_j)$ for $j = 1, ..., N$
    $\hat{\rho}^{(m)} \leftarrow \rho(\tilde{\mathbf{x}}^{(m)}, \mathbf{y}, \mathbf{z})$
end for
$\hat{\rho} \leftarrow \rho(\mathbf{x}, \mathbf{y}, \mathbf{z})$
$\hat{p} \leftarrow \sum_{m=1}^{M} \mathbf{1}\{\hat{\rho}^{(m)} \geq \hat{\rho}\}/M$
Output: p-value $\hat{p}$

Baseline algorithms

We implemented the KCIT and RCoT with code provided by the authors in [25] in their R package `RCIT`, available at `https://github.com/ericstrobl/RCIT`. The CCIT [23] was implemented with the code provided by the authors at

`https://github.com/rajatsen91/CCIT/blob/master/CCIT`. The CRT was implemented in python with our own code.

Supplement D: Genomics experiment details

The Cancer Cell Line Encyclopedia (CCLE) is a compilation of gene expression, chromosomal copy number and sequencing data from 947 human cancer cell lines. A cancer cell line can be understood as a string of cancer cells that keep dividing and growing over time under certain conditions in a laboratory. Then, using high-throughput sequencing technologies, the molecular characteristics of cancer cell lines, such as gene expression or mutation data, can be extracted. These genetic predictors were coupled with measures of drug sensitivity for $PLX4720$: a drug used against cancer whose response is available for 474 of the above cancer cell lines. By correlating the genetic information with the corresponding sensitivity to drug response, the data in principle allows for the identification of relevant genetic markers which could then lead to personalized treatment therapies depending on a patients genetic makeup. We illustrate this procedure in Figure 8.

Except for the conditional independence test to report significant genetic variables, our experiments followed similar procedures to those detailed in [26] and [1]. We choose to analyze dependence of drug response with 466 genetic mutations observed on each cancer line. We give summary statistics of the final data used in Table 1 below. This is a very high-dimensional problem that makes conditional independence testing unfeasible with traditional tests.

As in the original study in [1], we proceeded by fitting an elastic net model to predict drug response from genetic features with 10-fold cross-validation to optimize hyperparameters. Influential features were then ranked by their heuristic importance score given by the magnitude of fitted parameter values. The random forest model was used with default hyper-parameters in the python library `sklearn` and the CRT was implemented with a Gaussian approximation like in all other experiments. We ran the GCIT and the CRT considering each feature separately with drug response and all remaining features as confounders.

Remark. For a more systematic biological evaluation of features reported by the GCIT, we would use a more principled feature selection procedure such as Benjamini-Hochberg's correction for false discoveries [3].
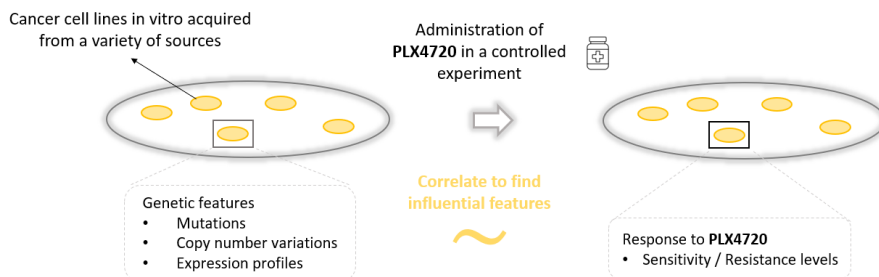

Figure 8: Diagram illustrating the data used in the Genetic experiment.

Table 1: Summary statistics of the final genetic data used from [1].

Statistics	Values
No. of cancer cell lines	474
No. of genetic mutations	466
Pearson correlation with drug response	**min:** 0.05, **max:** 0.51, **mean:** 0.07, **var:** 0.001
Drug response distribution	**min:** -97.9, **max:** 43.3, **mean:** -17.2, **var:** 633.3

References

[1] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[2] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. Mine: mutual information neural estimation. In *International Conference on Machine Learning*, 2018.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[4] Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test. *arXiv preprint arXiv:1807.05405*, 2018.

[5] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[6] Anindya Chatterjee, Joydeep Ghosh, Baskar Ramdas, Raghuveer Singh Mali, Holly Martin, Michihiro Kobayashi, Sasidhar Vemula, Victor H Canela, Emily R Waskow, Valeria Visconte, et al. Regulation of stat5 by fak and pak1 in oncogenic flt3-and kit-driven leukemogenesis. *Cell reports*, 9(4):1333–1348, 2014.

[7] Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

[8] Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

[9] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[10] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[13] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *ICLR*, 2019.

[14] Amit V Khera and Sekar Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature reviews Genetics*, 18(6):331, 2017.

[15] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[16] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[17] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In *Advances in neural information processing systems*, pages 1–9, 2013.

[18] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[19] Todd D Prickett, Brad Zerlanko, Jared J Gartner, Stephen CJ Parker, Ken Dutton-Regester, Jimmy C Lin, Jamie K Teer, Xiaomu Wei, Jiji Jiang, Guo Chen, et al. Somatic mutations in map3k5 attenuate its proapoptotic function in melanoma through increased binding to thioredoxin. *Journal of Investigative Dermatology*, 134(2):452–460, 2014.

[20] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

[21] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *AISTATS*, 2018.

[22] Rajat Sen, Karthikeyan Shanmugam, Himanshu Asnani, Arman Rahimzamani, and Sreeram Kannan. Mimic and classify: A meta-algorithm for conditional independence testing. *arXiv preprint arXiv:1806.09708*, 2018.

[23] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2951–2961, 2017.

[24] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

[25] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *arXiv preprint arXiv:1702.03877*, 2017.

[26] Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The hold-out randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*, 2018.

[27] Joy A Thomas and TM Cover. Elements of information theory. *John Wiley & Sons, Inc., New York. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, MPH (2009),"Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," Journal of the Royal Society Interface*, 6:187–202, 1991.

[28] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[29] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2012.

[30] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *ICLR*, 2017.

[31] Zhihong Zhu, Zhili Zheng, Futao Zhang, Yang Wu, Maciej Trzaskowski, Robert Maier, Matthew R Robinson, John J McGrath, Peter M Visscher, Naomi R Wray, et al. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications*, 9(1):224, 2018.