

# TCMI: a non-parametric mutual-dependence estimator for multivariate continuous distributions

B. Regler · M. Scheffler · L. M. Ghiringhelli

Received: date / Accepted: date

**Abstract** The identification of relevant features, i.e., the driving variables that determine a process or the property of a system, is an essential part of the analysis of data sets whose entries are described by a large number of variables. The preferred measure for quantifying the relevance of nonlinear statistical dependencies is mutual information, which requires as input probability distributions. Probability distributions cannot be reliably sampled and estimated from limited data, especially for real-valued data samples such as lengths or energies. Here, we introduce total cumulative mutual information (TCMI), a measure of the relevance of mutual dependencies based on cumulative probability distributions. TCMI can be estimated directly from sample data and is a non-parametric, robust and deterministic measure that facilitates comparisons and rankings between feature sets with different cardinality. The ranking induced by TCMI allows for feature selection, i.e. the identification of the set of relevant features that are statistical related to the process or the property of a system, while taking into account the number of data samples as well as the cardinality of the feature subsets. We evaluate the performance of our measure with simulated data, compare its performance with similar multivariate dependence measures, and demonstrate the effectiveness of our feature selection method on a set of standard data sets and a typical scenario in materials science.

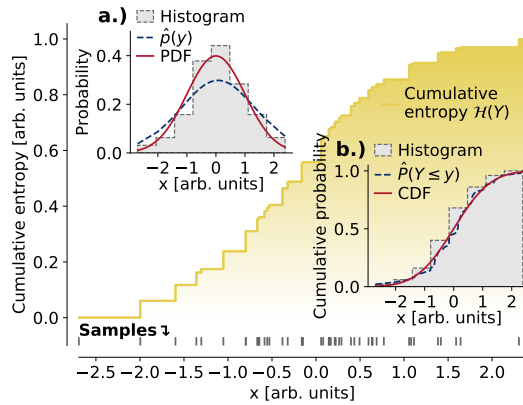
**Keywords** Dependence measure · Information theory · Mutual information · Feature selection · Machine learning · Materials science

## 1 Introduction

The past two decades have been marked by an explosion in the availability of scientific data and significant improvements in statistical analysis. Particularly in the physical sciences, an unparalleled surge in the exploration of data has been witnessed,

---

Fritz Haber Institute of the Max Planck Society, Berlin, Germany  
E-mail: regler@fhi-berlin.mpg.de



**Fig. 1** Empirical cumulative entropy  $\mathcal{H}(Y)$  of normal distribution for 50 data samples, which are shown as ticks in the bottom of the figure. Insets a.) and b.) show the (ground-truth) probability density (PDF) and cumulative probability (CDF) of the normal distribution, empirical cumulative distribution,  $\hat{P}(Y \leq y)$ , and estimated probability density,  $\hat{p}(y)$ . The estimated probability density was obtained by optimizing the bandwidth of the kernel density estimator with 10-fold cross-validation. Further, histograms of PDF and CDF are shown to underline the difficulty to approximate continuous distributions by discrete discontinuous functions even for low-dimensional data.

aiming at the data-driven identification of dependencies among physical variables. These observations have been even interpreted as the dawn of a new paradigm in science, the “big-data driven” science [1].

The identification of relevant features, i.e., properties or driving variables of a process or system’s property and often referred to as actuators or descriptors, has propelled investigations for an understanding of the underlying processes that generated the data [2]. In this context, a feature  $X \in \mathbf{X}$  may be an attribute, variable, parameter, or a combination of the above that has been measured or obtained from experiments or simulations. The main task is to estimate a statistical dependency,  $f : \mathbf{X} \mapsto Y$ , between a set of features  $\mathbf{X}$  and an output  $Y$  (target, response function), called functional dependence, subject to a feature selection criterion  $\mathcal{Q}$  [3,4],

$$\mathbf{X}^* = \arg \max_{\mathbf{X}' \subseteq \mathbf{X}} \mathcal{Q}(Y; \mathbf{X}'). \quad (1)$$

Feature selection screens the initial feature set  $\mathbf{X}$  for relevant features  $\mathbf{X}^*$  and ranks feature subsets based on their relevance. Therefore, it simplifies subsequent statistical data analysis and reduces the chances of detecting false dependencies. Used as a preliminary step for screening given feature sets, a robust and reliable feature-selection algorithm could for instance improve the performance and efficiency of data-analytics techniques such as symbolic regression, both in the genetic-programming [5] and compressed-sensing implementation [6,7], or regression-tree-based approaches [8].

Feature selection comprises two parts: (i) the choice of a search strategy and (ii) the feature selection criterion  $\mathcal{Q}$  for relevance. Search strategies for feature selection are either optimal (exhaustive) or sub-optimal (greedy). Optimal search strategies include exhaustive search and accelerated methods based on the monotonic property

Abbreviation	Explanation	Reference
CMI	Cumulative mutual information	[25]
MAC	Multivariate maximal correlation analysis	[26]
UDS	Universal dependency analysis	[28, 29]
MCDE	Monte Carlo dependency estimation	[35]
TCMI	Total cumulative mutual information	this work

**Table 1** Abbreviations used in the manuscript.

of a feature selection criterion, such as the branch and bound algorithm [9, 10, 11, 12, 13]. By construction, optimal search strategies explore the complete set of features for a global optimum. As such, they are impractical for high-dimensional data sets due to cost and time constraints in computer resources. Therefore, sub-optimal strategies have been developed to find local optima while balancing accuracy and speed: sequential (floating) forward selection [14, 15], sequential backward elimination [16], and minimal-redundancy-maximal-relevance criterion [17].

The academic community has extensively explored several feature selection criteria to evaluate a feature’s relevance [18], including distance measures [19, 20], dependency measures [21], consistency measures [22] and information measures [23]. In brief, an appropriate feature selection criterion  $\mathcal{Q}$  must be robust and deterministic, i.e., such that the feature selection is consistent and reproducible for the same type of settings and data. The prevailing method for quantifying dependencies is mutual information [24]. Mutual information quantifies the amount of information variables  $\mathbf{X}$  and  $Y$  share about each other.

There are several reasons to consider mutual-information-based quantities for feature selection. The two most important reasons are: (i) mutual information quantifies nonlinear statistical dependencies and (ii) mutual information provides an intuitive quantification of relevance for a feature subset  $\mathbf{X}' \subseteq \mathbf{X}$  relative to an output  $Y$  [23]. However, mutual information requires to estimate probability densities, which are problematic for high-dimensional data sets and are difficult to obtain from continuous distributions.

Prior investigations have implemented diverse approaches to address the problem: cumulative mutual information CMI [25], multivariate maximal correlation analysis MAC [26], maximal information coefficient MIC [27], universal dependency analysis UDS [28, 29], and mutual-information-based feature selection algorithms originally developed for discrete data [30, 31, 32, 33, 34]. The list is far from complete. However, all aforementioned measures are based on clustering, discretization, or estimation of probability densities of continuous data distributions; and thus make feature selection extremely dependent on the technique being used and requires to compute the feature selection criterion for each possible dependency in the data.

Yet, other dependence measures such as Pearson  $R$  and Spearman’s rank  $\rho$  correlation coefficients [36, 37], distance correlation DCOR [38, 39], kernel density estimation KDE [40, 41], or  $k$ -nearest neighbor estimation  $k$ -NN [42] are limited to bivariate dependencies only (Pearson, Spearman), are limited to specific types of dependencies (Spearman, DCOR), or require assumptions about the functional form of  $f$  (KDE,  $k$ -NN). Recent approaches, such as subspace slicing via high-contrast

Symbol	Definition
$Y$	output, target, response function
$X, \mathbf{X}$	features, variables
$\hat{X}$	empirical estimator of $X$
$\mathcal{Q}(Y;X), \mathcal{Q}^*(Y;X)$	(adjusted) feature selection criteria
<i>Discrete data</i>	
$I(Y;X)$	Shannon mutual information
$D(Y;X)$	fraction of information
$p(y)$	(marginal) probability density of $y \in Y$
$p(x,y)$	joint probability density of $x \in X$ and $y \in Y$
$p(y x)$	conditional probability density of $y \in Y$ given $x \in X$
<i>Continuous data</i>	
$\mathcal{I}(Y;X), \mathcal{I}^*(Y;X)$	(adjusted) cumulative mutual information
$\mathcal{Q}(Y;X), \mathcal{Q}^*(Y;X)$	(adjusted) fraction of cumulative information
$P(y), P(Y \leq y)$	(marginal) cumulative probability density of $y \in Y$
$P(x,y), P(X \leq x, Y \leq y)$	joint cumulative probability density of $x \in X$ and $y \in Y$
$P(y x), P(Y \leq Y X \leq x)$	conditional cumulative probability density of $y \in Y$ given $x \in X$

**Table 2** The list of symbols and notations used in this paper.

subspaces for density-based outliers ranking HiCS [43] or Monte Carlo dependency estimation MCDE [35], may enable new kind of feature selection criteria. At the moment, they still require to enumerate all possible combinations of features and, therefore, are computationally intractable for feature selection tasks.

The approach presented in this paper, instead, employs a mutual-information like quantity – total cumulative mutual information (TCMI). TCMI is a non-parametric, robust, and deterministic measure for estimating multivariate dependencies. Similar to CMI, MAC, and UDS, it is based on cumulative entropy [44,45,46,47] and adjusts the relevance of the features depending on the number of data samples and the cardinality of the feature subsets. While CMI, MAC, and UDS estimate conditional probability through clustering or discretization, TCMI defines all quantities by cumulative probabilities. We combine TCMI with a feature selection criterion to find subsets of features that influence the output.

Our feature selection procedure can be roughly divided into three steps: First, we quantify the dependence between the set of features and an output as the difference between cumulative marginal and cumulative conditional distributions. Second, we assess the relevance of a feature subset by comparing the strength of dependence with the mean dependence of features under the assumption of independence to reliably estimate the importance of one’s feature [48,49,50]. And third, we identify relevant features with the branch-and-bound algorithm [9,10,11,12,13], which has proven to be efficient in the discovery of nonlinear functional dependencies [51,52] and is the prerequisite for an exhaustive search for feature subsets with guarantees of optimality.

The remainder of this work is organized as follows. Section 2 introduces the theoretical background of mutual information and the concept of cumulative mutual information. Section 3 addresses the problem of estimating cumulative mutual information for continuous distributions from a limited set of sample data. Section 4 explains the steps to adjust the cumulative mutual information to assess the relevance

of feature sets. The same section also explains the feature selection criterion and the branch and bound implementation in detail. Next, Section 5 demonstrates the performance of TCMI on generated data, standard data sets, and on a typical scenario in materials science along with a comparison to similar multivariate dependence measures, namely CMI, MAC, UDS, and MCDE. Finally, Sections 6 and 7 present the discussion and conclusions of this work. Abbreviations, notations, and terminologies are summarized in Tables 1 and 2.

## 2 Theoretical background

Mutual information and all concepts presented in the following quantify relevance in terms of the similarity between two distributions  $U(\mathbf{X}, Y)$  and  $V(\mathbf{X}, Y)$  with Kullback-Leibler divergence,  $D_{\text{KL}}(U(\mathbf{X}, Y)||V(\mathbf{X}, Y))$  [53]. Based on mutual independence, they require no explicit modeling to quantify linear and nonlinear dependencies. Further, they monotonically increase with the cardinality of the feature's subset  $\mathbf{X}' \subseteq \mathbf{X}$ ,

$$\min_{\mathbf{X} \in \mathbf{X}'} D(Y; \mathbf{X}' \setminus X) \leq D(Y; \mathbf{X}') , \quad (2)$$

and are invariant under invertible transformations such as translations and any reparameterizations that preserve the order of  $\mathbf{X}$  and  $Y$  [54, 23].

For illustration purposes, only the case with two variables  $X$  and  $Y$  will be discussed in this section. However, a generalization to multiple variables can be derived directly from the independence assumption of random variables (see below).

### 2.1 Mutual information

Mutual information [55, 56] relates the joint probability distribution  $p(x, y)$  of two discrete random variables (=features) with the product of their marginal distribution  $p(x)$  and  $p(y)$ ,

$$\begin{aligned} I(Y; X) &= \sum_{y \in Y} \sum_{x \in X} p(y, x) \log \frac{p(y, x)}{p(y)p(x)} \\ &\equiv D_{\text{KL}}(p(y, x)||p(y)p(x)) . \end{aligned} \quad (3)$$

Mutual information is non-negative, is zero if and only if the variables are statistically independent,  $p(x, y) = p(x)p(y)$  (independence assumption of random variables), and increases monotonically with the mutual interdependence of variables otherwise. Further, mutual information indicates the reduction in the uncertainty of  $Y$  given  $X$  as  $I(Y; X) = H(Y) - H(Y|X)$ , where  $H(Y)$  denotes the Shannon entropy and  $H(Y|X)$  the conditional entropy [24]. The Shannon entropy  $H(Y)$  is defined as the expected value of the negative logarithm of the probability density  $p(y)$ ,

$$H(Y) = - \sum_{y \in Y} p(y) \log p(y) , \quad (4)$$

and can be interpreted as a measure of the uncertainty on the occurrence of events  $y$  whose probability  $p(y)$  is described by a random variable  $Y$ .

The conditional entropy  $H(Y|X)$  quantifies the amount of uncertainty about the value of  $Y$ , provided the value of  $X$  is known. It is given by

$$H(Y|X) = - \sum_{y \in Y} \sum_{x \in X} p(y,x) \log p(y|x) , \quad (5)$$

where  $p(y|x) = p(y,x)/p(x)$  is the conditional probability of  $y$  given  $x$ . Clearly,  $0 \leq H(Y|X) \leq H(Y)$  with  $H(Y|X) = 0$  if variables  $X$  and  $Y$  are functional dependent and  $H(Y|X) = H(Y)$  if variables are independent of each other.

Although mutual information is restricted to the closed interval  $0 \leq I(Y;X) \leq H(Y)$ , the upper bound is still dependent on  $Y$ . To facilitate comparisons, mutual information is normalized,

$$D(Y;X) = \frac{I(Y;X)}{H(Y)} = \frac{H(Y) - H(Y|X)}{H(Y)} . \quad (6)$$

Normalized mutual information, hereafter referred to as fraction of information (also known as coefficients of constraint [57], uncertainty coefficient [58], or proficiency [59]) quantifies the proportional reduction in uncertainty about  $Y$  when  $X$  is known. It is in the range  $D : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ , where 0 and 1 represent statistical independence and functional dependence, respectively [60].

## 2.2 Limitations of mutual information

Although mutual information was originally introduced for discrete data, it is a well-defined measure in the continuous domain. Defined as,

$$I(Y;X) = \int_{y \in Y} \int_{x \in X} p(y,x) \log \frac{p(y,x)}{p(y)p(x)} dx dy , \quad (7)$$

it requires to estimate probability densities from sample data. Common algorithms for probability density estimation are clustering [61,62,25], discretization [63,64,65], and density estimation [66,67,43,68,69]. All aforementioned methods necessarily introduce adjustable parameters. As a result, the arbitrariness of assigning these parameters has an impact on the identification of the optimal subset of features and on the ranking induced by the relevance function and the feature selection criterion (cf. Fig 1).

Other approaches to estimating probability densities are quantities that can be computed directly from sample data. An example is cumulative probability distributions, the anti-derivatives of probability densities:

$$P(x) := P(X \leq x) = \int_{-\infty}^x p(x') dx' . \quad (8)$$

Cumulative probability distributions of a random variable  $X$  evaluated at  $x$  describe the probability that  $X$  takes a value less than or equal to  $x$ . They are based on accumulated statistics and are more regular and less sensitive to statistical noise than

probability distributions [46, 47]. In addition, they are invariant under translations and reparameterizations that preserve the order of the original elements of the variables, i.e., positive monotonic transformations  $\mathcal{T} : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$P(x) = P(\mathcal{T}(x)) \quad \forall x \in X : x \mapsto \mathcal{T}(x) \quad \text{such that} \quad x_2 > x_1 \Rightarrow \mathcal{T}(x_2) > \mathcal{T}(x_1) . \quad (9)$$

However, cumulative probability distributions are not invariant under inversions as inversions transform them into either  $\mathcal{T}(X) \mapsto X$  or  $\mathcal{T}(X) \mapsto -X$ . Also, non-bijective transformations, e.g.,  $\mathcal{T}(X) = \pm|X|$ , do not preserve the order of the original elements of the variables either. To account for such transformations, they must be performed on the initial set of features, while creating additional features to be considered in the feature selection procedure.

Apart from this step, no further considerations are needed to introduce a new measure of relevance, called cumulative mutual information, analogous to mutual information.

### 2.3 Cumulative mutual information

Cumulative mutual information describes the inherent dependence expressed in the joint cumulative distribution  $P(x, y) = P(X \leq x, Y \leq Y)$  of two random variables relative to the product of their marginal cumulative distribution  $P(x)$  and  $P(y)$ ,

$$\begin{aligned} \mathcal{I}(Y; X) &= \int_{y \in Y} \int_{x \in X} P(y, x) \log \frac{P(y, x)}{P(y)P(x)} dx dy \\ &= D_{\text{KL}}(P(y, x) || P(y)P(x)) . \end{aligned} \quad (10)$$

Here, the independence assumption of random variables,  $P(y, x) = P(y)P(x)$ , results in a measure, that is again only zero if the variables are statistically independent and non-negative otherwise. Similarly to mutual information, cumulative mutual information quantifies the degree of dependency as the reduction in the uncertainty of  $Y$  given  $X$ , i.e.,  $\mathcal{I}(Y; X) = \mathcal{H}(Y) - \mathcal{H}(Y|X)$ . It is a function of cumulative entropy  $\mathcal{H}(Y)$  and conditional cumulative entropy  $\mathcal{H}(Y|X)$ ,

$$\mathcal{H}(Y) = - \int_{y \in Y} \int_{x \in X} P(y, x) \log P(y) dx dy \quad (11)$$

$$\mathcal{H}(Y|X) = - \int_{y \in Y} \int_{x \in X} P(y, x) \log P(y|x) dx dy , \quad (12)$$

where  $P(y|x) = P(y, x)/P(x)$  is the conditional cumulative distribution of  $y \leq Y$  given  $x \leq X$  (cf. Tab. 2). Again,  $\mathcal{H}(Y|X) = 0$  if variables  $X$  and  $Y$  are functional dependent and  $\mathcal{H}(Y|X) = \mathcal{H}(Y)$  if variables  $X$  and  $Y$  are independent of each other.

Bounds restrict cumulative mutual information to a closed interval  $0 \leq \mathcal{I}(Y; X) \leq \mathcal{H}(Y)$  with an upper bound still dependent on  $Y$ . For this reason, cumulative mutual information is normalized,

$$\mathcal{D}(Y; X) = \frac{\mathcal{I}(Y; X)}{\mathcal{H}(Y)} = \frac{\mathcal{H}(Y) - \mathcal{H}(Y|X)}{\mathcal{H}(Y)} , \quad (13)$$

and is hereafter referred to as fraction of cumulative information. Further, cumulative mutual information as well as fraction of cumulative information have the same properties as mutual information and fraction of information to be monotonically increasing with the cardinality of the feature set  $X$  (Eq. 2).

### 3 Empirical estimations of cumulative entropy and cumulative mutual information

The closed-form expression of cumulative mutual information (Eq. 10) is only applicable in the limit of large data samples,  $N \rightarrow \infty$ , and does not specify how to empirically derive the quantity in practical applications on a limited set of sample data.

For this reason, assume an empirical sample  $\{(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)\}$  drawn independently and identically distributed (i.i.d.) according to the joint distribution of  $X$  and  $Y$ . Such sample data induces empirical cumulative probability estimates for all variables  $Z \in \{Y, X\}$ , which lead to empirical estimators denoted by a hat (Tab. 2). For example, the empirical version of cumulative probability distribution is given by the sum of indicator functions,

$$\hat{P}(z) = \frac{1}{n} |\{i \mid z_i \leq z\}|, \quad \forall z_i \in Z, \quad (14)$$

that asymptotically converges to  $P(z)$  as  $n \rightarrow \infty$  for every value of  $z \in Z$  (Glivenko-Cantelli theorem [70, 71]). Thus, any empirical estimate,  $\hat{\mathcal{E}}$ , based on empirical cumulative probability converges pointwise as  $n \rightarrow \infty$  to the actual value of  $\mathcal{E}$ , i.e.,  $\hat{\mathcal{E}}(Z) \rightarrow \mathcal{E}(Z)$  [44, 47].

#### 3.1 Empirical cumulative entropy

The cumulative entropy is estimated by means of empirical probability distributions. For i.i.d. random samples that contain repeated values, empirical cumulative entropy has the form

$$\hat{\mathcal{H}}(Z) = - \sum_{i=1}^{k-1} \Delta z_i \hat{P}(z) \log \hat{P}(z) = - \sum_{i=1}^{k-1} (z_{(i+1)} - z_{(i)}) \frac{n_i}{n} \log \frac{n_i}{n}, \quad (15)$$

where  $z_{(i)}$  denotes the values  $z_{(0)} < z_{(1)} < \dots < z_{(k)}$  in sorted order of  $Z$  with  $z_{(0)} = -\infty$ , multiplicity  $n_i = |\{j \in n : z_{(i-1)} < z_j \leq z_{(i)}\}|$ , and constraint  $n = \sum_{i=1}^k n_i$ .

#### 3.2 Empirical conditional cumulative entropy

By construction, cumulative entropy is sensitive to the range of  $Z$ . The same is true for conditional cumulative entropy  $\mathcal{H}(Y|X)$  and its empirical estimate  $\hat{\mathcal{H}}(Y|X)$ . However, we are not interested in estimating  $\hat{\mathcal{H}}(Y|X)$ , but rather in estimating the fraction



of cumulative mutual information (Eq. 13). More precisely, we are interested in the residual fraction of cumulative information,

$$\hat{\mathcal{G}}_r(Y;X) = 1 - \hat{\mathcal{G}}(Y;X) = \frac{\hat{\mathcal{H}}(Y|X)}{\hat{\mathcal{H}}(Y)}. \quad (16)$$

We apply the following trick: to eliminate the scale dependence of  $X$ , we use the fact that features are invariant under rank-order preserving transformations  $\mathcal{T}$  (Eq. 9). Then, all features can be scaled to  $x' = \mathcal{T}(x)$  such that  $\Delta x'_i = x_{i+1} - x_i$  is constant and the volume element  $dx'$  in the integrals cancels out (cf. Eq. 12). Such a transformation is always possible and effectively removes the range dependence of variables from the residual fraction of cumulative information.

In this way, residual fraction of cumulative mutual information can be computed independently from the scaling of variables and is given by,

$$\hat{\mathcal{G}}_r(Y;X) = \frac{1}{m} \sum_{j=1}^m \frac{\sum_{i=1}^{n-1} \Delta y_i P(y_i, x_j) \log P(y_i|x_j)}{\sum_{i=1}^{n-1} \Delta y_i P(y_i, x_j) \log P(y_i)}. \quad (17)$$

## 4 Implementation details

Feature selection (Eq. 1) is an optimization problem that either requires a non-convex dependence measure or additional criteria to judge the optimality of a feature set [72]. Measures based on mutual information do not meet either requirement (cf. Eq. 2). However, mutual information can be turned into a convex measure by relating the strength of a dependence with the dependence of the same set of features under the independence assumption of random variables [48, 49]. This is because for a limited number of samples, independent variables tend to appear related to each other and can therefore be used to adjust (cumulative) mutual information.

### 4.1 Baseline adjustment

The limited availability of data makes it in practice challenging to estimate or calculate dependencies on empirical estimators. For example, empirical estimators need to assign a value (dependence score) close to zero for statistical independent features and a score close to one for functional dependent features. However, it is known that empirical estimators for mutual information never reach the theoretical maximum (functional dependence) or minimum (statistical independence), respectively and that mutual information tends to assign stronger dependencies for larger subsets of features regardless of the underlying relationship [35]. If the relevance of feature's subset cannot be evaluated directly by the information value, we need a baseline to actually compare dependence measures between subsets and different sizes of feature sets. One solution is to estimate the relevance of a feature subset by an adjusted measure,

$$\mathcal{Q}^*(Y;X) = \mathcal{Q}(Y;X) - \mathcal{Q}_0(Y;X), \quad (18)$$

where the relevance  $\mathcal{Q}$  of a feature subset  $X$  and the output vector  $Y$  is compared with the relevance  $\mathcal{Q}_0$  under the independence assumption of random variables [48, 49]. It requires that the adjustment term vanishes for large number of sample data  $\mathcal{Q}_0(Y;X) \rightarrow 0$  as  $n \rightarrow \infty$  and becomes zero if features are proportional to the output,  $\mathcal{Q}_0(Y;X) \rightarrow 0$  as  $X \rightarrow Y$ . While the baseline adjustment for mutual information has already been addressed [48, 49, 50, 52], we define the baseline adjustment for empirically cumulative mutual information as follows,

$$\hat{\mathcal{I}}^*(Y;X) = \hat{\mathcal{I}}(Y;X) - \hat{\mathcal{I}}_0(Y;X) , \quad (19)$$

$$\hat{\mathcal{D}}^*(Y;X) = \frac{\hat{\mathcal{I}}^*(Y;X)}{\mathcal{H}(Y)} = \hat{\mathcal{D}}(Y;X) - \hat{\mathcal{D}}_0(Y;X) \quad (20)$$

where  $\hat{\mathcal{I}}^*(Y;X)$  is the adjusted empirical cumulative mutual information,  $\hat{\mathcal{D}}^*(Y;X)$  is the adjusted fraction of empirical cumulative information, and  $\hat{\mathcal{I}}_0(Y;X)$  is the expected cumulative mutual information under the independence assumption of random variables.

Expected cumulative mutual information is given by permuting the values of each feature independently,  $X \rightarrow X(M)$ , and by taking the average of all particular realizations  $M \in \mathcal{M}$  of cumulative mutual information  $\hat{\mathcal{I}}(Y;X|M)$ ,

$$\hat{\mathcal{I}}_0(Y;X) = \sum_{M \in \mathcal{M}} \hat{\mathcal{I}}(Y;X|M) \mathcal{P}(Y;X|M) , \quad (21)$$

with  $\mathcal{M}$  being the set of all permutations of  $X$  and  $\mathcal{P}(Y;X|M)$  the probability to find a particular realization  $M$  for a given data set  $(Y,X)$ .

Although expected cumulative mutual information entails computationally intensive permutations, it can be rewritten in the form of a hypergeometric model of randomness [73, 48, 50] with quadratic complexity. The details can be found in the appendix and are analogous to the baseline adjustment for mutual information [48].

#### 4.2 Total cumulative mutual information

Besides defining a dependence measure based on cumulative probability distributions (Eq. 8), a similar measure for complementary cumulative probability distributions,  $P'(X) := P(X \geq x) = 1 - P(X \leq x)$ , can be derived, i.e.,

$$\mathcal{D}'(Y;X) = \frac{\mathcal{H}'(Y) - \mathcal{H}'(Y|X)}{\mathcal{H}'(Y)} , \quad (22)$$

$$\hat{\mathcal{D}}'^*(Y;X) = \hat{\mathcal{D}}'(Y;X) - \hat{\mathcal{D}}'_0(Y;X) , \quad (23)$$

where the cumulative entropies  $\mathcal{H}'(Y)$  and  $\mathcal{H}'(Y|X)$  are defined by  $P'(X)$ . Both measures  $\mathcal{D}(Y;X)$  and  $\mathcal{D}'(Y;X)$  quantify the dependence between features and an output from different sides of the distribution and impose lower and upper bounds on the information contained. The question arises of how to construct a single measure from these two.

In the context of feature selection, a natural choice is to instantiate total cumulative mutual information (TCMI) as the minimum contribution of fraction of empirical cumulative information for either of the two measures,

$$\hat{\mathcal{G}}_{\text{TCMI}}^*(Y;X) := \min(\hat{\mathcal{G}}^*(Y;X), \hat{\mathcal{G}}'^*(Y;X)) . \quad (24)$$

Thus, TCMI effectively quantifies the minimum strength of dependency between features and an output. While  $\hat{\mathcal{G}}_{\text{TCMI}}^*(Y;X)$  is extremely helpful in feature selection tasks, we use the average fraction of total cumulative information

$$\langle \hat{\mathcal{G}}_{\text{TCMI}}^*(Y;X) \rangle := \langle \hat{\mathcal{G}}_{\text{TCMI}}(Y;X) \rangle - \langle \hat{\mathcal{G}}_{\text{TCMI},0}(Y;X) \rangle \quad (25)$$

and

$$\begin{aligned} \langle \hat{\mathcal{G}}_{\text{TCMI}}(Y;X) \rangle &= \frac{1}{2} [\hat{\mathcal{G}}(Y;X) + \hat{\mathcal{G}}'(Y;X)] \\ \langle \hat{\mathcal{G}}_{\text{TCMI},0}(Y;X) \rangle &= \frac{1}{2} [\hat{\mathcal{G}}_0(Y;X) + \hat{\mathcal{G}}'_0(Y;X)] , \end{aligned} \quad (26)$$

for assessment tasks, i.e., for evaluating the strength of dependence, that compensates for imbalances in the contributions of total cumulative mutual information (Eq. 24).

#### 4.3 Feature selection

Having defined our feature selection criterion, we now briefly discuss the feature selection search strategy. As already mentioned in the introduction, the optimal search strategy (subset selection) of  $k$  features from an initial set of features  $\mathbf{X} = \{X_1, \dots, X_d\}$  is a combinatorial and exhaustive search procedure that is only applicable to low-dimensional problems. An efficient alternative to the exhaustive search is the depth-first branch-and-bound algorithm [9, 10, 11, 12]. It is an exponential search method and guarantees to find the optimal subset of feature variables without evaluating all possible subsets. However, prior knowledge of potential interdependencies between features and the target always leads to faster convergence and earlier termination of the algorithm. The performance therefore depends crucially on the features of the data set, which influence not only the performance of the feature-selection task, but also the maximum strength of the interdependence between features and the target. It may be that the given features only lead to a weak dependency. In these cases, it is recommended to include and consider as many features as possible in the feature-selection task and perform the analysis again.

In short, branch and bound maximizes an objective function  $\mathcal{Q}^* : \mathbf{X}' \rightarrow \mathbb{R}$  defined on a subset of features  $\mathbf{X}' \subseteq \mathbf{X}$  by making use of the monotonicity condition of a feature selection criterion,  $\mathcal{Q} : \mathbf{X}' \rightarrow \mathbb{R}$ , and a bounding function,  $\bar{\mathcal{Q}} : \mathbf{X}' \rightarrow \mathbb{R}$ . The monotonicity condition assumes that feature subsets  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k$  obtained by adding  $k$  features from the set of feature variables  $\mathbf{X}$  satisfy

$$\mathbf{X}_1 \subseteq \mathbf{X}_2 \subseteq \dots \subseteq \mathbf{X}_k , \quad (27)$$

---

```

Data: features  $\mathbf{X}$ , target  $Y$ 
Result: Optimal features  $\mathbf{X} \supseteq \mathbf{X}^* \leftarrow \text{optimal}$ 

1 begin
2    $\mathbf{S}_0 = \emptyset$ ;
3   subsets =  $\{\mathbf{S}_0\}$ ;
4   optimal =  $\mathbf{S}_0$ ;
5   while subsets do
6     for  $X_i \in \mathbf{X} \setminus \mathbf{S}_{k-1}$  do
7        $\mathbf{S}_k = \mathbf{S}_{k-1} \otimes X_i$ ;
8       Compute  $\mathcal{Q}(Y; \mathbf{S}_k)$  and  $\bar{\mathcal{Q}}(Y; \mathbf{S}_k)$ ;
9       if  $\mathcal{Q}(Y; \mathbf{S}_k) < \bar{\mathcal{Q}}(Y; \mathbf{S}_k)$  then
10        subsets = subsets  $\cup \{\mathbf{S}_k\}$ ;
11        if  $\mathcal{Q}(Y; \mathbf{S}_k) > \mathcal{Q}(Y; \text{optimal})$  then
12          optimal =  $\mathbf{S}_k$ ;
13        end
14      end
15    end
16  end
17 end

```

---

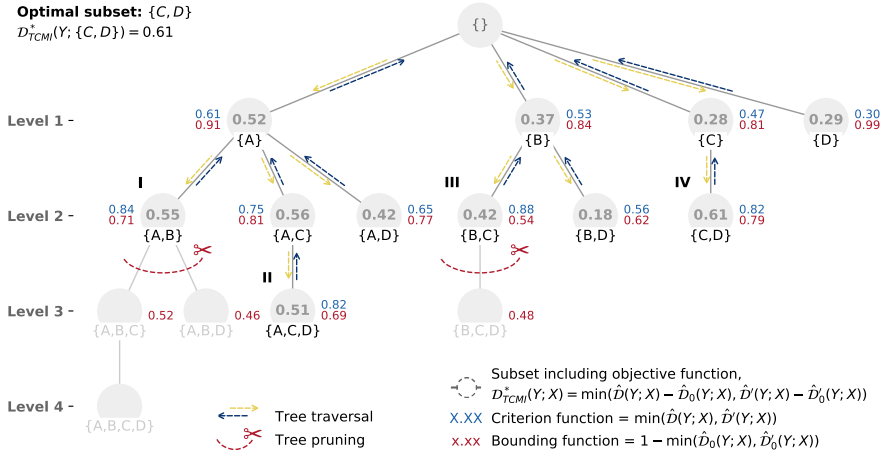
**Algorithm 1:** Simplified branch and bound algorithm for feature selection.

so that feature selection criterion  $\mathcal{Q}$  and bounding function  $\bar{\mathcal{Q}}$  are monotonically increasing and decreasing respectively,

$$\begin{aligned}
 \mathcal{Q}(\mathbf{X}_1) &\leq \mathcal{Q}(\mathbf{X}_2) \leq \dots \leq \mathcal{Q}(\mathbf{X}_k) \\
 \bar{\mathcal{Q}}(\mathbf{X}_1) &\geq \bar{\mathcal{Q}}(\mathbf{X}_2) \geq \dots \geq \bar{\mathcal{Q}}(\mathbf{X}_k).
 \end{aligned} \tag{28}$$

The branch and bound algorithm, sketched in Alg. 1 and Fig. 2, constructs a search tree where the root represents the empty subset and leaves represent subsets of  $k$  features. While traversing the tree down to leaves from left to right, a limited number of (non-redundant) sub-trees is generated by augmenting the subset by one feature from the initial set of features  $\mathbf{X}$  (branching step). The algorithm keeps the information about the currently best subset  $X^* := \mathbf{X}_k$  and the corresponding objective function it yields (the current maximum). Anytime the objective function  $\mathcal{Q}^*$  in some internal nodes exceeds the bounding function  $\bar{\mathcal{Q}}$  of sub-trees, it decreases – either due to the condition Eq. 28 or the bounding function is lower than the current maximum value of the objective function, sub-trees can be pruned and computations be skipped (bounding step). On termination of the algorithm, the bound contains the optimum objective function value and found subsets of features are ranked in descending order of the objective function values.

As objective and criterion function we set  $\mathcal{Q}^* = \mathcal{Q}_{\text{TCMI}}^*(Y; X)$ , the criterion function to be  $\mathcal{Q} = \min(\mathcal{Q}(Y; X), \mathcal{Q}'(Y; X))$ , and, as a pruning rule, the bounding function to be  $\bar{\mathcal{Q}} = 1 - \min(\hat{\mathcal{G}}_0(Y; X), \hat{\mathcal{G}}'_0(Y; X))$  (Eq. 24). Proofs for the monotonicity conditions for  $\mathcal{Q}$  and  $\bar{\mathcal{Q}}$  follow similar arguments as for Shannon entropy [52] and are provided in the appendix.













**Fig. 2** Example of a depth-first tree search strategy of the branch and bound algorithm [9, 10, 11, 12] to search for the optimal subset of features. Shown is the tree traversal going from top to down and left to right by dashed arrows, the estimated fraction of total cumulative information (objective function inside circles), subsets of features (labels at the bottom of the circles), fraction of cumulative information (criterion function, first number, right or left the circles), and the expected fraction of cumulative information contribution (bounding function, second number, right or left the circles). Capital roman symbols indicate applied pruning rules or updates of the current maximum objective function. Anytime the objective function in some internal nodes exceeds the bounding function of sub-trees (I), it decreases (II) – either due to the condition Eq. 28 or the bounding function is lower than the current maximum value of the objective function (III), sub-trees can be pruned and computations be skipped. On termination of the algorithm, the bound contains the optimum objective function value (IV).

#### 4.4 Complexity Analysis

Finally, we analyze the computational complexity of TCMI. In the worst case, for  $n$  number of example data and  $d$  features, cumulative mutual information must evaluate the integral  $\mathcal{O}(n^d)$  times and  $\mathcal{O}(n^2)$  times to calculate the baseline adjustment term. Thus, TCMI has time complexity  $\mathcal{O}(n^d)$  and suffers from the curse of dimensionality [74]. Second, branch and bound evaluates  $\binom{d}{1}$  features in the first level,  $\binom{d}{2}$  features in the second level,  $\binom{d}{3}$  features in the third level, and so on until all features are explored. Thus, a total of  $\sum_{k=1}^d \binom{d}{k} = 2^d - 1$  subsets of features with a total time complexity of about  $\mathcal{O}(2^d)$  are evaluated. Third, the ranking of subsets involves  $\mathcal{O}((n \log n)^d)$  sorting operations in case all subsets are relevant.

As a result, the total time complexity of the feature selection algorithm is non-deterministic polynomial-time (NP)-hard and, in general, the search strategy of examining all possible subsets is not viable. In the vast majority of cases, however, dependencies are relatively simple relationships of only a small number of features. In addition, feature selection can be restricted at any time to examine subsets of features that are less than or equal to a predefined dimensionality of a feature subset. Then the time complexity is greatly reduced and the feature selection can be solved in polynomial time. Whether the assumptions apply to arbitrary data sets is a case-by-case

Name	$\rho^2$	$\langle \hat{\mathcal{S}}_{\text{TCMI}}^* \rangle$	$\langle \hat{\mathcal{S}}_{\text{TCMI}} \rangle$	$\langle \hat{\mathcal{S}}_{\text{TCMI},0} \rangle$	CMI	MAC	UDS	MCDE
 linear	1.0000	0.97	1.00	0.03	1.00	1.00	0.67	1.00
 exponential	1.0000	0.97	1.00	0.03	1.00	1.00	0.65	1.00
 step-2	0.9999	0.96	0.98	0.02	1.00	1.00	0.67	1.00
 step-4	0.9996	0.93	0.95	0.02	1.00	1.00	0.67	1.00
 step-8	0.9984	0.87	0.88	0.01	1.00	1.00	0.67	1.00
 random	0.0091	0.33	0.65	0.32	0.02	0.34	0.00	0.54
 sawtooth-8	0.0016	0.23	0.31	0.07	0.03	0.03	0.00	0.14
 sawtooth-4	0.0004	0.17	0.27	0.10	0.00	0.01	0.00	0.09
 sawtooth-2	0.0001	0.09	0.19	0.09	0.00	0.00	0.00	0.03
 constant	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	1.00

**Table 3** Dependence scores,  $\langle \hat{\mathcal{S}}_{\text{TCMI}}^*(Y;X) \rangle$ , between a linear data distribution and a linear, exponential, step, sawtooth and uniform (random) distribution. The data sample size is  $n = 200$ . Step-like distributions were generated by discretization of the linear distribution with each value repeating  $r$ -times. Sawtooth-like distributions have 2, 4, or 8 number of steps per ramp and  $\lceil n/level \rceil$  ramps in total. The table also shows Spearman’s rank correlation coefficient squared  $\rho^2$ , total cumulative mutual information contributions,  $\langle \hat{\mathcal{S}}_{\text{TCMI}}(Y;X) \rangle$  and  $\langle \hat{\mathcal{S}}_{\text{TCMI},0}(Y;X) \rangle$ , and the scores from similar dependence measures such as CMI [25], MAC [26], UDS [28,29], and MCDE [35].

study. However, indicators such as the convergence rate of the TCMI approaching the maximum value or the estimated strength of the relationships are helpful in exploratory data analysis to search for relevant features.

## 5 Experiments

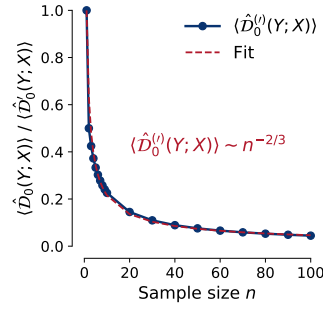
To demonstrate the performance of TCMI in different settings, we first look at generated data and show that our method can detect both univariate and multivariate dependencies. Then, we discuss applications of TCMI on data sets from KEEL and UCI Machine Learning Repository [75, 76, 77] and a typical scenario from the materials science community, namely to predict the crystal structure of octet-binary compound semiconductors [6, 78].

### 5.1 Case study on generated data

In a number of experiments, we test the theoretical properties of TCMI, i.e., its invariance properties and performance statistics. We also study an exemplified feature-selection task to find a bivariate normal distribution defined in a multi-dimensional space.

#### 5.1.1 Interpretability of TCMI

In the first experiment we investigate TCMI with respect to a linear data distribution  $Y$  of size  $n = 200$  and different distributions  $X$  as features of different similarity to the



**Fig. 3** Expected empirical cumulative mutual information,  $\langle \hat{D}_0^{(l)}(Y; X) \rangle$ , with respect to the number of sample data. Shown is the dependency (solid line) and a heuristic derived analytic functional relationship (dashed line).

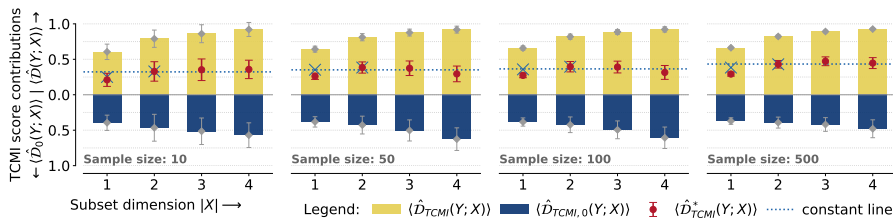
output (Tab. 3). Besides linear, exponential, and constant distributions (zero vector), we consider stepwise distributions generated by discretization of the linear distribution, where each value is repeated 2, 4 or 8 times. Furthermore, we consider uniform (random) and sawtooth distributions with 2, 4 or 8 steps per ramp. The results show that (i) there is a clear relationship between the TCMI values in terms of the similarity between the feature and the output, (ii) TCMI is zero for a constant distribution, or approaching one for an exact dependency (see also Fig. 3), and that (iii) dependence measures such as CMI [25], MAC [26], UDS [28, 29] and MCDE [35] are less sensitive in the distributions than TCMI.

Further, results show that MCDE cannot differentiate between a linear and a constant distribution and that the random distribution has a higher TCMI score, i.e., stronger dependency, than sawtooth distributions. In fact, the strength of the TCMI dependency score in this bivariate example is consistent with Spearman’s rank coefficient of determination  $\rho^2$  [37], which is well suited to assessing the similarity between bivariate monotonic distributions.

There is one remark, we would like to point out here: Due to the limited availability in the data, random variables lead to spurious dependencies that are clearly reflected in TCMI and MCDE, but not in CMI, MAC and UDS. Especially, the baseline adjustment  $\langle \hat{\mathcal{G}}_{\text{TCMI}, 0} \rangle$  for a random variable is larger than any other tested dependence of Tab. 3. A large baseline adjustment results in smaller TCMI values, so it is unlikely that random variables will be part of the feature selection. However, if the dependencies are of the same strength as spurious dependencies induced by random variables, TCMI may select features that do not influence the output. Therefore, the strength of the dependency must always be compared with dependencies of random variables.

### 5.1.2 Properties of the baseline correction term

In the second experiment, we take a closer look at the baseline adjustment term that decreases monotonically with respect to the number of sample data. Baseline adjustment is given by expected empirical cumulative mutual information (Eqs. 20 and 21).



**Fig. 4** Fraction of cumulative information scores against increasing dimensionality for  $\{Y, \mathbf{X}\}$  using 10, 50, 100, and 500 data samples generated from mutually independent and uniform distributions of size  $\mathbf{X} = \{Y, X_1, \dots, X_4\}$ . Contributions of average fraction of total cumulative mutual information,  $\langle D_{\text{TCMI}}(Y;X) \rangle$  and  $\langle D_{\text{TCMI},0}(Y;X) \rangle$  are shown on either side of the plot and the resulting score  $\langle \mathcal{D}_{\text{TCMI}}^*(Y;X) \rangle$  as points. Error bars indicate standard deviations from repeating the experiment 50 times. Since  $X$  and  $Y$  are independent, average total cumulative mutual information  $\langle \mathcal{D}_{\text{TCMI}}^*(Y;X) \rangle$  should be constant across subsets of features independent of sample size and subset dimensionality. While  $\langle D_{\text{TCMI}}(Y;X) \rangle$  is increasing with the cardinality of the feature set and  $\langle D_{\text{TCMI},0}(Y;X) \rangle$  decreasing,  $\langle \mathcal{D}_{\text{TCMI}}^*(Y;X) \rangle$  is approximately constant for a wide range of data samples 10...500 and subset dimensionality 1...4. The crosses represent the deviation of the TCMI from the constant baseline. By enlarging the feature subset with a shuffled version of the feature, TCMI can be corrected.

As our evaluation of the baseline adjustment shows, expected empirical cumulative mutual information decreases with increasing number of sample sizes in all our test cases presented below. For instance, the baseline adjustment for linear dependencies roughly follows a  $\langle \hat{\mathcal{D}}_0^{(l)}(Y;X) \rangle \sim n^{-2/3}$  scaling law that vanishes as  $n \rightarrow \infty$  (Fig. 3).

### 5.1.3 Invariance properties of TCMI

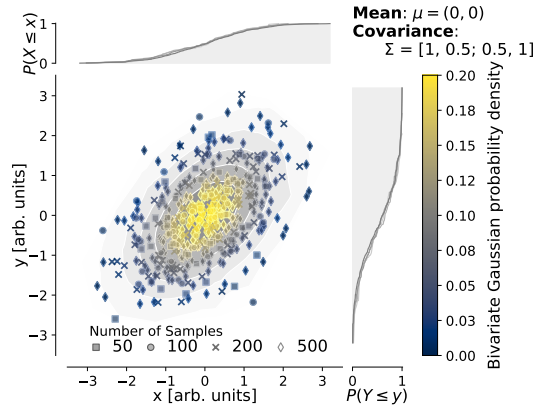
In the third experiment, we investigate the invariance properties of TCMI. By construction, TCMI is invariant under positive monotonic transformations (Eq. 9). To provide a comprehensive evaluation, we generated random distributions  $X$  of different sizes and reparameterize variables using invertible transformation (cf. Sec. 2.2). Results on monotonic transformations, e.g.,  $\mathcal{T}(X) = aX^k + b$  where  $a, b, k \in \mathbb{R}$ , or compositions, e.g.,  $\mathcal{T}(X) = \mathcal{T}_1(X) \pm \dots \pm \mathcal{T}_m(X)$ , show that as long as the order of the original elements of the features is preserved, TCMI is invariant. Furthermore, experiments with different feature subset sizes of random distributions show the invariance of TCMI by exchanging features, namely  $\hat{\mathcal{D}}_{\text{TCMI}}^*(Y; \mathbf{X}) = \hat{\mathcal{D}}_{\text{TCMI}}^*(Y; \mathbf{X}')$  for all  $\mathbf{X}' \in \text{perm}(\mathbf{X})$ , without having to define a sorted order of feature set as compared to CMI, MAC or UDS [25, 65, 28, 29].

### 5.1.4 Baseline adjustment of TCMI

In the fourth experiment, we investigate baseline adjustment (Sec. 4.1). We generated mutually independent and uniform distributions  $Z = \{Y, X_1, \dots, X_d\}$  of dimensionality  $d$  with sample sizes 10, 50, 100, and 500 and compared TCMI across subsets of feature variables of different subspace dimensionality while repeating the experiment 50 times. Fig. 4 summarizes the results.

While scores are constantly zero in the discrete case [48], scores are zero for TCMI only if the distribution is exactly uniform. In general, uniform distributions are





**Fig. 5** Bivariate normal probability distribution with mean  $\mu = (0, 0)$  and covariance matrix  $\Sigma = [1, 0.5; 0.5, 1]$ . Shown is a scatter plot with 50, 100, 200, and 500 data samples, its cumulative probability distributions,  $P(Z \leq z)$ ,  $Z \in \{X, Y\}$ , and contour lines of equal probability densities  $\in \{0.01, 0.02, 0.05, 0.08, 0.13\}$ .

generated pseudo-randomly and due to random sampling, we expect constant scores approaching  $\langle \mathcal{D}_{\text{TCMI}}^*(Y; X) \rangle \rightarrow 0.5$  as  $n \rightarrow \infty$  independent of the dimensionality of  $X = \{X_1, \dots, X_d\}$  in the case if none of the features is relevant.

Indeed, TCMI is approximately constant for a wide range of data samples  $10 \dots 500$  and subset dimensionality  $1 \dots 4$  and approaches  $\langle \mathcal{D}_{\text{TCMI}}^*(Y; X) \rangle \rightarrow 0.5$  as  $n \rightarrow \infty$ . Only in the one-dimensional case, TCMI under-estimates the dependency as CMI, MAC, UDS, and MCDE does even for higher dimensional subsets. The reason for this is that TCMI does not contain the self-correlation of the statistical noise of a feature. By enlarging the feature subset with a shuffled version of the feature, TCMI can be corrected (Fig. 4). As a result, TCMI shows to have a clear comparison mechanism of scores across different subsets of features independent of the number of data samples.

### 5.1.5 Bivariate normal distribution

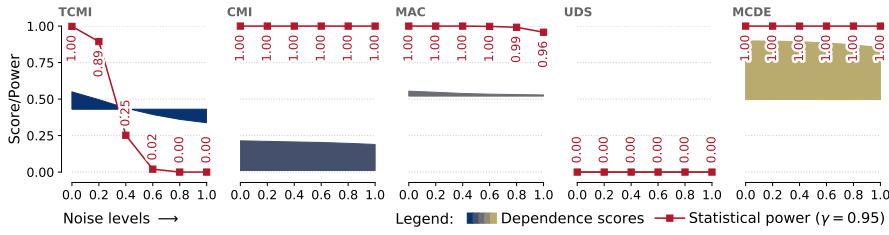
At last, we consider a simple feature-selection task with known ground truth, namely to find a bivariate normal distribution defined in a high-dimensional space. For this purpose, we generated a bivariate normal distribution of size  $n = 500$  from features  $x$  and  $y$ , added features such as normal, exponential, logistic, triangular, uniform, Laplace, Rayleigh, and Weibull distributions all with zero mean  $\mu = 0$  and identity covariance matrix  $\sigma = 1$ , and augmented the feature space as described in Section 2.2. In terms of Pearson or Spearman's correlation coefficient, none of the features have coefficients of determinations higher than 1% with respect to the bivariate normal distribution, so the data set appears to be completely uncorrelated. However, since the ground truth is known, there are two features, namely  $x$  and  $y$ , to describe the bivariate normal distribution of the data set.

Dependence Measure	Sample size			
	50	100	200	500
TCMI	{logistic,x}=0.54 {rayleigh,weibull}=0.53 {x,y}=0.52 {x}=0.35 {rayleigh}=0.35 {laplace}=0.35 {triangular}=0.34	{y,rayleigh}=0.57 {y,laplace}=0.55 {y,uniform}=0.55 {y}=0.46	{x,y}=0.58 {y,exponential}=0.55 {y}=0.48 {y,poisson}=0.47	{y,x}=0.60 {x,normal}=0.57 {normal,triangular}=0.57 {x}=0.38
CMI [25]	{y}=1.00 {logistic}=1.00 {triangular}=1.00 {laplace}=1.00	{y}=1.00 {logistic}=1.00 {normal}=1.00 {triangular}=1.00 {laplace}=1.00	{y}=1.00 {x}=1.00 {triangular}=1.00 {laplace}=1.00 {logistic}=1.00 {normal}=1.00	{y}=1.00 {x}=1.00 {logistic}=1.00 {triangular}=1.00 {laplace}=1.00 {normal}=1.00
MAC [26]	{y,laplace}=0.90 {y,x}=0.89 {y,triangular}=0.89 {y,exponential}=0.89 {y,normal}=0.89 {y,rayleigh}=0.89 {y,uniform}=0.89 {y,weibull}=0.89 {y,logistic}=0.89 {y}=0.88	{x,laplace}=0.82 {x,logistic}=0.82 {x,weibull}=0.82 {x,triangular}=0.82 {x,exponential}=0.82 {x,normal}=0.82 {x,rayleigh}=0.82 {x,uniform}=0.82 {x,y}=0.82 {y}=0.81	{y}=0.83 {x}=0.83	{y}=0.81 {weibull}=0.78
UDS [28, 29]	{laplace}=0.52	{y}=0.49 {normal}=0.48	{normal}=0.47	{normal}=0.45 {logistic}=0.44
MCDE [35]	{y}=0.84	{x}=0.88 {y}=0.86	{x}=0.92 {y}=0.88	{y}=0.94 {x}=0.92

**Table 4** Topmost feature subsets in the order of identification from the bivariate normal distribution data set with 50, 100, 200, and 500 data samples as being restricted to subset dimensionality  $\leq 2$  and selected by the following dependence measures: total cumulative mutual information (TCMI), cumulative mutual information (CMI), multivariate maximal correlation analysis (MAC), universal dependency analysis (UDS), and Monte Carlo dependency estimation (MCDE).

*Subspace search* In order to find the two most relevant features from the high-dimensional data set, subspace search is performed up to the second subset dimensionality. Further, feature selection is being performed for four sets of 50, 100, 200, and 500 data samples (Fig. 5). Results are reported in Table 4.

Overall, almost all dependence measures find at least one of the two relevant features  $x$ ,  $y$ , both, or at least similar distributions, such as the normal distribution. However, scores and subset sizes of relevant features decrease with larger sample sizes for MAC and UDS, and CMI identifies exact dependencies even between distributions where none dependency exists, e.g., between a Laplacian and bivariate normal distribution. In contrast, MCDE robustly finds one of the relevant features  $x$  or  $y$ , but never finds two of them being jointly relevant. TCMI also finds relevant features, but scores and relevance are more determined by sample size as it is being reflected in the score. Hence, with sample sizes greater than 200, TCMI is the only dependence measure that correctly identifies the optimal feature subset to be  $\{x, y\}$ . Still, TCMI scores are



**Fig. 6** Statistical power analysis with 95% confidence of dependence measures at different noise levels  $\sigma = 0 \dots 1$ : total cumulative mutual information (TCMI), cumulative mutual information (CMI), multi-variate maximal correlation analysis (MAC), universal dependency analysis (UDS), and Monte Carlo dependency estimation (MCDE). The diagrams also show the trends in the dependence scores of the optimal feature subset  $\{x, y\}$  of the bivariate normal distribution.

lower than of the other dependence measures, even though the score increases for larger sample sizes.

*Statistical power analysis* To assess the robustness of dependence measures, we performed statistical power analysis of CMI, MAC, UDS, MCDE, and TCMI and added Gaussian noise with increasing standard deviation  $\sigma$  [26, 28, 35]. We considered 5 + 1 noise levels, distributed linearly from 0 to 1, inclusive. We computed the score of the bivariate normal distribution for each dependency  $\Lambda = \{\text{CMI, MAC, UDS, MCDE, TCMI}\}$ , i.e.,  $\langle \Lambda(Y; X) \rangle_\sigma$ , with  $n = 500$  data samples and feature subset  $\{x, y\}$  and compared it with the score of independently drawn random data samples,  $\langle \Lambda(Y; I) \rangle_0$ , of the same size ( $n = 500$ ) and dimension ( $d = 1 + 2$ ). The power of a dependence measure  $\Lambda$ , was then evaluated as the probability  $P$  of a dependence score to be larger than the  $\gamma$ -th percentile of the score with respect to the independence,

$$\text{Power}_{\Lambda, \sigma}^\gamma(Y; X) := P(\langle \Lambda(Y; X) \rangle_\sigma > \langle \Lambda(Y; I) \rangle_0^\gamma). \quad (29)$$

Essentially, the power of a dependence measure quantifies the contrast, i.e., difference, between dependence  $X$  and independence  $I$  at noise level  $\sigma$  with  $\gamma\%$  confidence. It is a relative statistical measure and depends on the strength of the dependency. Therefore, dependence strengths that are close to independence are likely to be more sensitive to analysis than stronger dependencies.

For our experiments, we set  $\gamma = 95\%$ , repeated the experiment 500 times, while shuffling the data samples at each iteration, computed the scores  $\langle \Lambda(Y; X) \rangle_\sigma$  and  $\langle \Lambda(Y; I) \rangle_0^\gamma$  for every dependence measure at noise level  $\sigma$ , and recorded the average and standard deviation of the respective dependence measures. The results of statistical power analysis, the average score of the dependence measures and independence as well as the contrast are summarized in Figure 6.

With the exception of MAC, the statistical power of all dependence measures tends to be constant or to decrease with increasing noise level. It is remarkable that MCDE is the only dependence measure that has a high statistical power, offers a high contrast and assesses a strong dependency. In particular, the contrast of MCDE provides excellent statistics, even at noise levels much higher than TCMI. Although MAC and CMI also have high statistical power, their contrasts or dependence scores

are low. While a low contrast raises difficulties in identifying feature subsets and is a serious problem, a low dependence score is not a problem as long as all other subsets assess dependencies of the same or smaller strength. Thus, in our analysis, UDS completely fails to detect dependencies in line with observations [35] and TCMI shows some peculiar features: In general, TCMI is dependent on the number of samples (Eq. 19) and its contrast generally increases with more data samples. However, TCMI is more sensitive and, therefore, less robust as compared to the other dependence measures. An in-depth analysis shows: the sensitivity is merely due to the moderate strength of the dependency as the statistical power is much more robust for stronger dependencies in other data sets we tested.

## 5.2 Case study on real-world data

Next, we study selected real-world data sets from KEEL and UCI Machine Learning Repository [75,76,77], and highlight TCMI for one, not restricted to, typical application of the materials science community, namely crystal-structure predictions of octet-binary compound semiconductors [6,78].

### 5.2.1 KEEL and UCI regression data sets

We investigate how TCMI and similar dependence measures perform in real-world problems developed for multivariate regression tasks. Unfortunately, in practice, not every data set is known to have relevant features. Therefore, we compare our results with analyzed data sets with known relevant features. All in all, we consider one simulated data set from the KEEL database [75,76] and two data sets from the UCI Machine Learning Repository [77]:

#### 1. **Friedman #1 regression** [79]

This data set is used for modeling computer outputs. Inputs  $X_1$  to  $X_5$  are independent features that are uniformly distributed over the interval  $[0, 1]$ . The output  $Y$  is created according to the formula:

$$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon \quad (30)$$

where  $\varepsilon$  is the standard normal deviate  $N(0, 1)$ . In addition, the data set has five redundant variables  $X_6 \dots X_{10}$  that are i.i.d random samples. Further, we enlarge the number of features by adding four variables  $X_{11} \dots X_{14}$  each very strongly correlated with  $X_1 \dots X_4$  and generated by  $f(x) = x + N(0, 0.01)$ .

#### 2. **Concrete compressive strength** [80]

The aim of this data set is to predict the compressive strength of high performance concrete. Compressive strength is the ability of a material or structure to withstand loads that tend to reduce size. It is a highly nonlinear function of age and ingredients. These ingredients include cement, water, blast furnace slag (a by-product of iron and steel production), fly ash (a coal combustion product), superplasticizer (additive to improve the flow characteristics of concrete), coarse aggregate (e.g., crushed stone or gravel), and fine aggregate (e.g., sand).

### 3. Forest fires [81]

This data set focuses on wildfires in the Montesinho Natural Park, which is located at the northern border of Portugal. It includes features such as local coordinates  $x$  and  $y$  where a fire occurred, the time (day, month, and year), temperature (temp), relative humidity (RH), wind, rain, and derived forest fire features such as fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC), and initial spread index (ISI) to estimate the propagation speed of fire.

For each data set, we performed feature selection using all aforementioned dependence measures (TCMI, CMI, MAC, UDS, MCDE) and compared resulting feature subsets with potentially relevant features reported from the original references. Results are summarized in Table 5.

As our results show, even in the simplest example of the Friedman regression data set, two dependence measures show extreme behavior: UDS selects no features and MAC selects all features of the data set and therefore does not perform any selection at all. Both dependence measures do not only completely fail to identify the actual dependencies of the Friedman regression data set, but also fail in the concrete compressive strength and forest fires data set. Therefore, it is likely that these dependence measures report incorrect results in other data sets and are therefore inappropriate for feature selection and dependence assessment tasks.

In contrast, CMI and MCDE tend to identify low-dimensional feature subsets and, thus, can only detect a few dependencies of a data set. In this case, both dependence measures are able to find relevant features, in the sense that they partly agree with potentially relevant features reported from the respective references. The only exception is TCMI, which effectively selects all relevant features of the Friedman regression data set. However, TCMI it is not free from selecting non-relevant features in other feature subsets as it reports  $X_7$  or  $X_8$  in the fourth or fifth best feature subset. Therefore, dependence scores need to be directly related with the baseline adjustment term, and the lower the dependence scores are, the more likely non-relevant features are in the subsets (cf. Sec. 5.1.1).

Found feature subsets for the Friedman regression data set as well as for the concrete compressive strength data set have high dependence scores. They agree well with relevant features as reported by the references, even though TCMI misses slag in the concrete compressive strength example: It is likely that features such as fine and coarse aggregate or superplasticizer serve as a substitute for slag due to the limited number of data samples. However, we cannot test this assumption as all data samples were used for the analysis and none are available for further tests.

The only difference in the selection of feature subsets is in the forest fires data set. Apart from weather conditions, TCMI also includes some of the derived forest fires features such as duff moisture (CMD) and drought code (DC), that are indirectly related to rainfall and estimate the lower and deeper soil moisture content. Since temperature and relative humidity as well as duff moisture and drought code are not only reported by TCMI, but also by CMI and MCDE, indicates the relevance of these features in the forest fire prediction, although they were not mentioned as in the reference [81]. Admittedly, the TCMI scores are moderate, which indicates difficulties

in assessing the interdependencies between the features and the burnt area of forest fires as a whole.

### 5.2.2 Octet-binary compound semiconductors

Our last example is dedicated to a typical, well characterized, and canonical materials-science problem, namely the crystal-structure stability prediction of octet-binary compound semiconductors [6, 78]. Octet-binary compound semiconductors are materials consisting of two elements formed by groups of I/VII, II/VI, III/V, or IV/IV elements leading to a full valence shell. They crystallize in rock salt (RS) or zinc blende (ZB) structures, i.e., either with ionic or covalent bindings and were already studied in the 1970's [82, 83], followed by further studies [84, 85], and recent work using machine learning [86, 6, 78, 7].

The data set is composed of 82 materials with two atomic species in the unit cell. The objective is to accurately predict the energy difference  $\Delta E$  between RS and ZB structures based on 8 electro-chemical atomic properties for each atomic species  $A/B$  (in total 16) such as atomic ionization potential IP, electron affinity EA, the energies of the highest-occupied and lowest-unoccupied Kohn-Sham levels, H and L, and the expectation value of the radial probability densities of the valence  $s$ -,  $p$ -, and  $d$ -orbitals,  $r_s$ ,  $r_p$ , and  $r_d$ , respectively [6]. As a reference, we added Mulliken electronegativity  $EN = -(\text{IP} + \text{EA})/2$  to the data set and also studied the best two features from the publication [6]

$$D_1 = \frac{\text{IP}(B) - \text{EA}(B)}{r_p(A)^2}, \quad D_2 = \frac{|r_s(A) - r_p(B)|}{\exp[r_s(A)]}, \quad (31)$$

as known dependencies to show the consistency of the method as well as to probe TCMI with linearly dependent features [6].

To predict the energy difference  $\Delta E$  between RS and ZB structures, we conducted a subspace search with TCMI to identify the subset of features that have the strongest dependence on  $\Delta E$ . Results are summarized in Table 6. In total, the strongest dependence on  $\Delta E$  was found with six features from both atomic species,  $A$  and  $B$ , before TCMI decreased again with seven features.

Results reveal that there are several feature subsets that are found to be optimal among different cardinalities. We note that TCMI never selects Mulliken electronegativity EN together with either electron affinity EA or ionization potential IP for the same atomic species. We also notice that EN can be replaced by IP (see bold feature subsets in Tab. 6), whereas EN replaced by EA result in slightly smaller TCMI values (by at least 0.02 in case of the optimal subsets, not shown in the table) as EN is found to be stronger linearly correlated with IP than with EA. Thus, results do not only corroborate the functional relationship between EN, IP, and EA, but also the consistency of TCMI.

Furthermore, TCMI indicates that features, like the atomic radii  $r_s(B)$  and  $r_p(B)$  or the energies  $EN(B)$ ,  $H(B)$ ,  $H(B)$  and  $IP(B)$  of IV to VIII elements, can be used interchangeably without reducing the dependence scores. Indeed, by assessing dependencies between pairwise feature combinations, TCMI identifies  $r_s(B)$  and  $r_p(B)$

<b>Dependence Measure</b>	<b>Relevant feature subsets</b> (Data set, reported relevant features, feature subsets by dependence measures)
Friedman #1 regression [79]: $X_1 \dots X_{14}$	
Potentially relevant features: $X_1 \dots X_5$ and $X_{11} \dots X_{14}$ [500 data samples]	
TCMI	$\{X_{14}, X_{12}, X_1, X_5, X_3\} = 0.79$ , $\{X_{14}, X_{12}, X_1, X_5\} = 0.77$ , $\{X_4, X_2, X_1, X_3\} = 0.77$ , $\{X_4, X_2, X_1, X_8\} = 0.76$ , $\{X_{14}, X_{12}, X_1, X_7\} = 0.75$
CMI	$\{X_{14}\} = 1.00$ , $\{X_4\} = 1.00$
MAC	$\{X_{14}, X_8, X_9, X_7, X_{11}, X_3, X_6, X_{10}, X_{12}, X_5\} = 0.89$ , ... (+ 119.981 subsets = 0.89)
UDS	–
MCDE	$\{X_2\} = 0.78$ , $\{X_{12}\} = 0.77$ , $\{X_{11}\} = 0.77$ , $\{X_1\} = 0.77$
Concrete compressive strength [80]: age, cement, water, blast furnace slag (slag), fly ash, superplasticizer (sp), coarse aggregate (coarse_aggr), fine aggregate (fine_aggr)	
Potentially relevant features: age, cement, water, slag [1030 data samples]	
TCMI	$\{\text{cement, sp, water, coarse\_aggr, fine\_aggr}\} = 0.68$ $\{\text{fine\_aggr, water, sp, coarse\_aggr, fly\_ash}\} = 0.68$ $\{\text{fine\_aggr, water, sp, coarse\_aggr, age}\} = 0.68$ $\{\text{cement, coarse\_aggr, water, slag, fine\_aggr}\} = 0.68$ $\{\text{fine\_aggr, slag, water, coarse\_aggr, age}\} = 0.67$ $\{\text{cement, coarse\_aggr, water, sp, age}\} = 0.67$ $\{\text{cement, coarse\_aggr, fine\_aggr, sp, age}\} = 0.67$ $\{\text{coarse\_aggr, cement, fine\_aggr, water, sp}\} = 0.66$
CMI	$\{\text{age}\} = 1.00$ , $\{\text{cement}\} = 1.00$ , $\{\text{coarse\_aggr}\} = 1.00$ , $\{\text{fine\_aggr}\} = 1.00$ , $\{\text{slag}\} = 1.00$ , $\{\text{water}\} = 1.00$ , $\{\text{sp}\} = 0.98$
MAC	$\{\text{water, coarse\_aggr, fine\_aggr, cement, sp, slag, fly\_ash, age}\} = 0.76$
UDS	–
MCDE	$\{\text{age}\} = 0.90$
Forest fires [81]: $x$ , $y$ , time (day, month, and year), temperature (temp), relative humidity (RH), wind, rain, fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC), initial spread index (ISI)	
Potentially relevant features: temp, rain, RH, wind [517 data samples]	
TCMI	$\{\text{DMC, RH, ISI, temp, wind, DC}\} = 0.53$ , $\{\text{DMC, RH, DC, temp, FFMC, wind}\} = 0.51$
CMI	$\{\text{temp, DC}\} = 1.00$ , $\{\text{temp, DMC}\} = 1.00$ , $\{\text{temp, RH}\} = 1.00$ , $\{\text{temp, FFMC}\} = 1.00$ , $\{\text{FFMC, DC}\} = 1.00$ , $\{\text{FFMC, DMC}\} = 1.00$ , $\{\text{FFMC, RH}\} = 1.00$ , $\{\text{FFMC, temp}\} = 1.00$ , $\{\text{DMC, DC}\} = 1.00$ , $\{\text{DMC, ISI}\} = 1.00$ , $\{\text{DMC, RH}\} = 1.00$ , $\{\text{DMC, month}\} = 1.00$ , $\{\text{DC, DMC}\} = 1.00$ , $\{\text{DC, ISI}\} = 1.00$ , $\{\text{DC, RH}\} = 1.00$ , $\{\text{DC, month}\} = 1.00$ , $\{\text{RH, DMC}\} = 1.00$ , $\{\text{RH, DC}\} = 1.00$ , $\{\text{ISI, DMC}\} = 1.00$ , $\{\text{ISI, DC}\} = 1.00$ , $\{\text{temp, month}\} = 1.00$
MAC	$\{\text{temp, RH, DMC, FFMC, DC, ISI, wind, day, x}\} = 0.85$ , $\{\text{temp, RH, DMC, FFMC, DC, ISI, wind, day}\} = 0.83$ , $\{\text{temp, RH, DMC, FFMC, DC, ISI, wind, x}\} = 0.83$
UDS	$\{\text{rain}\} = 0.35$
MCDE	$\{\text{DMC, temp, RH}\} = 0.84$ , $\{\text{DMC, temp, DC}\} = 0.82$ , $\{\text{DMC, temp, FFMC}\} = 0.81$

**Table 5** Relevant feature subsets for selected data sets from the KEEL database [75,76] and UCI Machine Learning Repository [77], designed for multivariate regression tasks and feature selection as found out by total cumulative mutual information (TCMI), cumulative mutual information (CMI), multivariate maximal correlation analysis (MAC), universal dependency analysis (UDS), and Monte Carlo dependency estimation (MCDE). For comparison, potentially relevant feature subsets mentioned in the references are also included.

Subset dimension	Feature subsets and dependence score (TCMI)	Metrics (GBDT)			
		RMSE	MAE	MaxAE	$r^2$
6	* $\{D_2, EA(A), r_p(A), r_s(A), r_p(B), L(B)\} = 0.84$	0.15	0.10	0.43	0.86
	$\{EA(A), r_p(A), r_s(A), EN(B), L(B), r_s(B)\} = \mathbf{0.82}$	0.12	0.08	0.32	0.91
	$\{EA(A), r_p(A), r_s(A), EN(B), L(B), r_p(B)\} = \mathbf{0.82}$	0.12	0.08	0.32	0.91
	$\{EA(A), r_p(A), r_s(A), IP(B), L(B), r_s(B)\} = \mathbf{0.82}$	0.13	0.09	0.33	0.90
	$\{EA(A), r_p(A), r_s(A), IP(B), L(B), r_p(B)\} = \mathbf{0.82}$	0.13	0.09	0.33	0.90
	$\{EA(A), r_p(A), r_s(A), H(B), L(B), r_s(B)\} = 0.82$	0.14	0.10	0.36	0.87
	$\{EA(A), r_p(A), r_s(A), H(B), L(B), r_p(B)\} = 0.82$	0.14	0.10	0.36	0.87
	$\{EA(A), H(A), r_d(A), r_p(A), L(B), r_d(B)\} = 0.82$	0.15	0.10	0.45	0.86
	$\{EA(A), H(A), r_d(A), r_s(A), L(B), r_d(B)\} = 0.82$	0.16	0.10	0.46	0.85
	$\{EA(A), H(A), r_p(A), L(B), r_d(B), r_p(B)\} = 0.81$	0.14	0.10	0.37	0.88
	$\{EA(A), H(A), r_p(A), L(B), r_d(B), r_s(B)\} = 0.81$	0.14	0.10	0.37	0.87
	5	$\{EA(A), r_p(A), r_s(A), IP(B), L(B)\} = \mathbf{0.79}$	0.13	0.08	0.40
$\{EA(A), r_p(A), r_s(A), EN(B), L(B)\} = \mathbf{0.79}$		0.14	0.08	0.46	0.88
$\{EA(A), r_p(A), r_s(A), H(B), L(B)\} = 0.79$		0.15	0.09	0.42	0.86
* $\{D_1, D_2, r_p(A), r_s(A), r_s(B)\} = 0.79$		0.17	0.10	0.50	0.83
$\{EN(A), r_p(A), r_s(A), IP(B), L(B)\} = 0.78$		0.14	0.08	0.43	0.88
$\{EA(A), H(A), r_p(A), L(B), r_s(B)\} = 0.78$		0.14	0.10	0.37	0.88
$\{EA(A), H(A), r_p(A), L(B), r_p(B)\} = 0.78$		0.14	0.10	0.37	0.88
$\{EA(A), H(A), r_d(A), r_p(A), L(B)\} = 0.78$		0.17	0.09	0.51	0.84
$\{EA(A), H(A), r_d(A), r_s(A), L(B)\} = 0.78$		0.17	0.10	0.53	0.83
$\{EA(A), H(A), L(A), r_s(A), L(B)\} = 0.78$		0.18	0.10	0.55	0.82
4	$\{EA(A), r_p(A), r_s(A), L(B)\} = 0.78$	0.16	0.09	0.49	0.85
	$\{L(A), r_p(A), r_s(A), r_p(B)\} = 0.76$	0.13	0.09	0.35	0.90
	$\{L(A), r_p(A), r_s(A), r_s(B)\} = 0.76$	0.13	0.09	0.33	0.90
	$\{EN(A), r_p(A), r_s(A), L(B)\} = 0.76$	0.17	0.10	0.52	0.83
	* $\{D_1, r_p(A), r_s(A), r_s(B)\} = 0.75$	0.15	0.11	0.37	0.87
3	$\{r_p(A), r_s(A), r_s(B)\} = 0.73$	0.13	0.10	0.31	0.89
	$\{IP(A), r_p(A), L(B)\} = \mathbf{0.73}$	0.16	0.10	0.49	0.84
	$\{r_p(A), r_s(A), L(B)\} = 0.73$	0.16	0.10	0.48	0.84
	$\{EN(A), r_p(A), L(B)\} = \mathbf{0.73}$	0.18	0.11	0.53	0.80
	$\{r_p(A), r_s(A), r_p(B)\} = 0.72$	0.13	0.10	0.31	0.89
	$\{IP(A), r_s(A), L(B)\} = \mathbf{0.72}$	0.17	0.10	0.49	0.82
	$\{EN(A), r_s(A), L(B)\} = \mathbf{0.72}$	0.18	0.11	0.52	0.80
	* $\{D_1, r_s(A), r_p(B)\} = 0.70$	0.15	0.11	0.40	0.86
2	* $\{D_1, r_s(B)\} = 0.71$	0.19	0.14	0.52	0.76
	$\{r_s(A), L(B)\} = 0.69$	0.18	0.12	0.49	0.80
	$\{r_s(A), r_s(B)\} = 0.67$	0.14	0.10	0.34	0.88
	* $\{D_1, D_2\} = 0.62$	0.19	0.14	0.53	0.77
1	* $\{D_1\} = 0.57$	0.23	0.18	0.56	0.69
	$\{r_s(A)\} = 0.56$	0.21	0.15	0.53	0.75
	$\{r_p(A)\} = 0.55$	0.21	0.15	0.54	0.75
All 16 features (GBDT reference):		0.15	0.09	0.45	0.86

**Table 6** Relevant feature subsets for the octet-binary compound semiconductors data set as found out by total cumulative mutual information (TCMI) showing the two most relevant feature subsets of each cardinality. For comparison, feature best subsets for  $D_1 = D_1(IP(B), EA(B), r_p(A))$  and  $D_2 = D_2(r_s(A), r_p(B))$  from reference [6] (entries with a star  $\star$ ) are also listed. Bold feature subsets mark subsets with interchangeable features EN and IP. The table also shows statistics of constructed machine-learning models using the gradient boosting decision tree (GBDT) algorithm [87] with 10-fold cross-validation: root-mean-squared error (RMSE), mean absolute error (MAE), maximum absolute error (MaxAE), and Pearson coefficient of determination ( $r^2$ ).



to be strongly dependent and  $\text{EN}(B)$ ,  $\text{H}(B)$ , and  $\text{IP}(B)$  strongly dependent, consistent with bivariate correlation measures such as Pearson or Spearman. In numbers, the Pearson coefficient of determination ( $r^2$ ) between the atomic radii  $r_s$  and  $r_p$  are  $r^2(r_s(A), r_p(A)) = 0.94$ ,  $r^2(r_s(B), r_p(B)) = 0.99$  and the Pearson coefficient of determination between Mulliken electronegativity and ionization potential or electron affinity is  $r^2(\text{EN}(B), \text{IP}(B)) = 0.96$ , or  $r^2(\text{EN}(B), \text{H}(B)) = 0.99$ , respectively. These findings illustrate that TCMI assigns similar scores to collinear features.

Features  $D_1$  and  $D_2$  (Eq. 31) from the reference [6], are combinations of atomic properties that best represent  $\Delta E$  linearly,

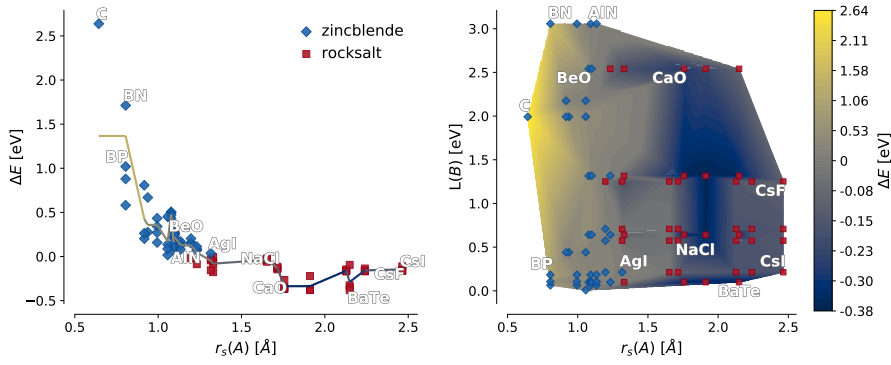
$$D_1 = D_1(\text{IP}(B), \text{EA}(B), r_p(A)) , \quad (32)$$

$$D_2 = D_2(r_s(A), r_p(B)) . \quad (33)$$

As such, they incorporate knowledge that generally lead to higher TCMI scores for the same feature subset cardinality. While this applies to the first and second subset dimensions, feature subsets with the aforementioned features  $D_1$ ,  $D_2$  are on par with feature subsets based on atomic properties at higher dimensions. However,  $D_1$  and  $D_2$  are not selected consistently by TCMI because TCMI does not make any assumption about the linearity of the dependency  $(D_1, D_2) \mapsto \Delta E$ . This suggests that the linear combination of  $D_1$  and  $D_2$  is a good, but a not complete, description of  $\Delta E$ .

A visualization of relevant subsets also reveals clear monotonous relationships in one and two dimensions (Fig. 7). In addition, we construct machine-learning models for each feature subset and report model statistics for the prediction of  $\Delta E$  along with statistics of the full feature set (Tab. 6). The details can be found in the appendix. We partitioned the data set into  $k = 10$  groups (so-called folds) and generated  $k$  machine-learning models, using 9 folds to generate the model and the  $k$ -th fold to test the model (10-fold cross validation). To reduce variability, we performed five rounds of cross-validation with different partitions and averaged the rounds to obtain an estimate of the model’s predictive performance. For the machine-learning models we used the gradient boosting decision tree algorithm (GBDT) [87]. GBDT is resilient to feature scaling (Eq. 9) just like TCMI and is one of the best available, award-winning, and versatile machine-learning algorithm for classification and regression [88, 89, 90]. Notwithstanding this, traditional methods sensitive to feature scaling may show superior performances for data sets with sample sizes larger than number of considered features [91] (compare also model performances in Tab. 6 with references [6, 78, 7]).

Machine-learning models are designed to improve with more data and a feature subset that best represents the data for the machine learning algorithm [87, 92]. Therefore, we expect a general trend of higher model performances with larger subset cardinalities. Furthermore, we do not expect that the optimal feature subset of TCMI performs best for every machine-learning model (“No free lunch” theorem [93, 94, 95, 96]) as an optimal feature subset identified by the feature selection criterion TCMI may not be same according to other evaluation criteria such as root-mean-squared error (RMSE), mean absolute error (MAE), maximum absolute error (MaxAE), or Pearson coefficient of determination ( $r^2$ ). This fact is evident in our analysis. The choice of GBDT may not be optimal because its predictive performance generally decreases with the number of noisy features (compare the model performance with



**Fig. 7** Feature spaces of the topmost selected feature subsets for one (left) and two dimensions (right). Shown are the two classes of crystal-lattice structures as diamonds (zinc blende) and squares (rock salt), their distribution, and the trend line/manifold in the prediction of the energy difference  $\Delta E$  between rock salt and zinc blende. The trend line/manifold was computed from with the gradient boosting decision tree algorithm [87] and 10-fold cross validation. For reference, some octet-binary compound semiconductors are labeled.

all 16 features to a subset, Tab. 6). However, to the best of our knowledge, there is no other machine-learning algorithm that models data without making assumptions about the functional form of dependency, is independent of an intrinsic metric, and can operate on a small number of data samples. Therefore, our focus is only on the predictive performance of the found subsets relative to the predictive performance of identified subsets and all features in the data set (Tab. 6).

Results confirm the general trend of higher model performances with larger feature subset cardinalities and show that the initial subset of 16 features can be reduced down to 6 features without decreasing model performances. Essentially, feature subsets with three to four features are already as good as a machine-learning model with all 16 features, where noisy features already start to degrade the prediction performance of the GBDT model. The overall performance gradually increases with the subset cardinality. However, our analysis identifies significant variability in performance with a higher dispersion for feature subsets at smaller dependence scores than for larger values.

An exhaustive search for the best GBDT model yields an optimum of seven features to best predict the energy difference between rock salt and zinc blende crystal structures with  $D1$  and  $D2$  neglected,

$$\{EA(A), IP(A), r_d(A), r_p(A), IP(B), r_s(B), r_p(B)\}$$

$$\text{RMSE} : 0.11, \text{MAE} : 0.08, \text{MaxAE} : 0.27, r^2 : 0.92 .$$

In contrast to the optimal feature subsets of TCMI (cf. Tab. 6), the optimal GBDT feature set is a variation of optimal feature subsets of TCMI with highest-occupied Kohn-Sham level and ionization potential interchanged,  $H(A) \leftrightarrow IP(A)$ , and lowest-unoccupied Kohn-Sham level,  $L(B)$ , missing. Model performances demonstrate that the optimal feature subsets of TCMI are close to the model's optimum and corroborate the usefulness of TCMI in finding relevant feature subsets for machine-learning

predictions. Slight differences in performances are mainly due to the variances of the cross-validation procedure and the small number of 82 data samples, which effectively limited the reliable identification of larger feature subsets in the case of TCMI (Tab. 4).

## 6 Discussion

Although TCMI is a non-parametric, robust, and deterministic measure, the biggest limitation is its computational complexity. For small data sets ( $n < 500$ ) and feature subsets ( $d < 5$ ) feature selection finishes in minutes to hours on a modern computer. For larger data sets, however, TCMI scales with  $\mathcal{O}(n^d)$  and quickly exceeds any realizable runtime. Furthermore, the search for the optimal feature subset also needs to be improved. Even though in our analysis only a fraction of less than one percent of the possible search space had to be evaluated, TCMI was evaluated hundreds of thousands of times. Future research towards pairwise evaluations [17], Monte Carlo sampling [35], or gradual evaluation of features based on iterative refinement strategies of sampling will show to what extent the computational costs of TCMI can be reduced.

A further limitation is that non-relevant features may be selected in the optimal feature subsets, when only a limited amount of data points is available (see Sec. 5.1.5). By construction, the identification of feature subsets is dependent on the feature selection search strategy (cf. Sec. 1). The results show that it is critical to use optimal search strategies because sub-optimal search strategies can report subsets of features that are not related to the output. Even if the exhaustive search for feature subsets is computationally intensive, it can be implemented efficiently, e.g., by using the branch-and-bound algorithm. In our implementation, the branch-and-bound algorithm was used to search for optimal, i.e. minimal non-redundant feature subsets. However, as our results demonstrate, different feature subsets with few or no common features may lead to similar dependence scores. The main rationale for this outcome is that the features may be correlated with each other and therefore contain redundant information about dependencies. Including these redundant features will surely lead to a higher stability of the method, more consistent results, and better insights into the actual dependency. If a machine-learning algorithm is given, the best option at present is to generate predictive models for each of the found feature subsets and select the one that works best.

## 7 Conclusions

We constructed a non-parametric and deterministic dependence measure based on cumulative probability distribution [44,45] to propose fraction of cumulative mutual information  $\mathcal{D}(Y; \mathbf{X})$ , an information-theoretic divergence measure to quantify dependencies of multivariate continuous distributions. Our measure can be directly estimated from sample data using well-defined empirical estimates (Sec. 2). Fraction of cumulative mutual information quantifies dependencies without breaking permutation invariance of feature exchanges, i.e.,  $\mathcal{D}(Y; \mathbf{X}) = \mathcal{D}(Y; \mathbf{X}')$  for all  $\mathbf{X}' \in \text{perm}(\mathbf{X})$ ,

while being invariant under invertible transformations. Measures based on mutual information are monotonously increasing with respect to the cardinality of feature subsets and sample size. To turn fraction of cumulative mutual information into a convex measure, we related the strength of a dependence with the dependence of the same set of features under the independence assumption of random variables [48, 49]. We further constructed a measure based on complementary cumulative probability distributions and introduced total cumulative mutual information  $\langle \hat{\mathcal{G}}_{\text{TCMI}}^*(Y; \mathbf{X}) \rangle$ .

Tests with simulated and real data confirm that total cumulative mutual information is capable of identifying relevant features of linear and nonlinear dependencies. The main application of total cumulative mutual information is to assess dependencies, to reduce an initial set of features before processing scientific data, and to identify relevant subset of features, which jointly have the largest dependency and minimum redundancy on the output. The performance of the total cumulative mutual information is still exponential and thus outweighs potential benefits of TCMI. In future works, we will address the performance issues of TCMI, the stability of identified feature subsets, and provide a feature selection framework that is also suitable for discrete, continuous, and mixed data types. We will also apply TCMI to current problems in the physical sciences with a practical focus on the identification of feature subsets to simplify subsequent data-analysis tasks.

Since total cumulative mutual information identifies dependencies with strong mutual contributions, it is applicable to a wide range of problems directly operating on multivariate continuous data distributions and does not need to require probability density estimation, clustering, or discretization. Thus, total cumulative mutual information has the potential to promote an information-theoretic understanding of functional dependencies in different research areas and can be used to gain more insights from data.

**Acknowledgements** This research received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (Grant Agreement No. 676580: the NOMAD Laboratory and European Center of Excellence and Grant Agreement No. 740233: TEC1p) and from BiGmax, the Max Planck Society’s Research Network on Big-Data-Driven Materials Science. B.R. acknowledges financial support from the Max Planck Society. L.M.G. acknowledges support from Berlin Big-Data Center (Grant Agreement No. 01IS14013E). The authors thank J. Vreeken, M. Boley and P. Mandros for inspiring discussions and for carefully reading the manuscript.

## Appendix

### A Baseline adjustment

Dependency measurements that assign stronger dependencies for larger subsets of features independently of the underlying relationship are considered biased [48]. To actually compare dependence measures across subsets of features and different cardinality, dependence measures must have a baseline. Baseline adjustment is addressed by eliminating the inherent bias of the measure, so that the baseline becomes constant under the random assumption of variables. The baseline adjustment was discussed for

$\mathbf{Y} \setminus \mathbf{X}$	$\tilde{X}_1$	$\cdots$	$\tilde{X}_j$	$\cdots$	$\tilde{X}_c$	
$\tilde{Y}_1$	$n_{11}$	$\cdots$	$\cdot$	$\cdots$	$n_{1c}$	$a_1$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\tilde{Y}_i$	$\cdot$		$n_{ij}$		$\cdot$	$a_i$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\tilde{Y}_r$	$n_{r1}$	$\cdots$	$\cdot$	$\cdots$	$n_{rc}$	$a_r$
	$b_1$	$\cdots$	$b_j$	$\cdots$	$b_c$	

**Fig. 8** A  $r \times c$  cumulative contingency table  $\mathcal{M}$  related to two clusterings  $\tilde{X}$  and  $\tilde{Y}$  with row marginals,  $a_i = \sum_{j=1}^c n_{ij}$ , and column marginals,  $b_j = \sum_{i=1}^r n_{ij}$ . The two marginal sum vectors  $a = [a_i]$  and  $b = [b_j]$  are constant and satisfy the fixed marginals condition,  $\sum_{i=1}^r a_i = \sum_{j=1}^c b_j = N$ .

mutual information in [48, 49, 50, 52]. By following the notation of *Vinh, Epps et. al.* [48], we propose the baseline adjustment for cumulative mutual information.

A common model of randomness is the hypergeometric model [73, 48, 50] (also called permutation model). It uniformly and randomly generates  $m$  distinct permutations of pairs  $M$  with probability  $\mathcal{P}(Y; X|M)$  by permuting all values of each variable in the data set,

$$\hat{\mathcal{I}}_0(Y; X) = \sum_{M \in \mathcal{M}} \hat{\mathcal{I}}(Y; X|M) \mathcal{P}(Y; X|M). \quad (34)$$

Then the baseline-adjusted cumulative fraction of the information can be obtained by subtracting fraction of the cumulative information (Eq. 10) from the expected fraction of the cumulative information under the assumption of independent and identical distributed random variables,

$$\hat{\mathcal{I}}^*(Y; X) = \hat{\mathcal{I}}(Y; X) - \hat{\mathcal{I}}_0(Y; X), \quad (35)$$

$$\hat{\mathcal{D}}^*(Y; X) = \hat{\mathcal{D}}(Y; X) - \hat{\mathcal{D}}_0(Y; X) = \frac{\hat{\mathcal{I}}^*(Y; X)}{\mathcal{H}(Y)}. \quad (36)$$

Specifically, the average cumulative mutual information between all different permutations with  $|X_i| = a_i$ ,  $i = 1, \dots, r$  and  $|Y_j| = b_j$ ,  $j = 1, \dots, c$  has constant marginal sum vectors  $a = [a_i]$  and  $b = [b_j]$ . Therefore, the cumulative information overlap between  $X$  and  $Y$ ,

$$\hat{\mathcal{I}}_0(Y; X|M) = \hat{\mathcal{I}}_0(a, b|M = [n_{ij}]_{j=1 \dots c}^{i=1 \dots r}) = - \sum_{i=1}^{r-1} \sum_{j=1}^c \Delta y_i(M) \frac{n_{ij}}{n} \log \frac{n_{ij}}{b_j}, \quad (37)$$

can be summarized in the form of a  $r \times c$  cumulative contingency table,  $M = [n_{ij}]_{j=1 \dots c}^{i=1 \dots r}$  (Tab. 8), with  $n_{ij}$  as being a specific realization of the joint cumulative probability given row marginal  $a_i$  and column marginal  $b_j$ .

By rearranging the sums in Eq. 37 and writing the sum over the entire permutation of variable values as a sum over all permutations of possible values of  $n_{ij}$ , we get

$$\begin{aligned} \hat{\mathcal{I}}_0(Y; X) &= - \sum_{M \in \mathcal{M}} \sum_{i=1}^{r-1} \sum_{j=1}^c \Delta y_i(M) \frac{n_{ij}}{n} \log \frac{n_{ij}}{b_j} \mathcal{P}(Y; X|M) \\ &= - \sum_{i=1}^{r-1} \sum_{j=1}^c \sum_{n_{ij}} \Delta y_i(n_{ij}, a_i, b_j|M) \cdot \frac{n_{ij}}{n} \log \frac{n_{ij}}{b_j} \mathcal{P}(n_{ij}, a_i, b_j|M), \end{aligned} \quad (38)$$

$$\hat{\mathcal{J}}_0(Y;X) = - \sum_{i=1}^{r-1} \sum_{j=1}^c \sum_{n_{ij}} \Delta y_i(M|n_{ij}, a_i, b_j) \frac{n_{ij}}{n} \log \left( \frac{n_{ij}}{b_j} \right) \cdot \frac{(r-i)!(i-1)!(b_j-1)!(r-b_j)!}{(b_j-n_{ij})!(r-i-b_j+n_{ij})!(n_{ij}-1)!(i-n_{ij})!(r-1)!} \quad (42)$$

where  $\mathcal{P}(n_{ij}, a_i, b_j|M)$  is the probability to encounter an associative cumulative contingency table subject to fixed marginals.

The probability to encounter an associative cumulative contingency table subject to fixed marginals, with the cell at the  $i$ -th row and  $j$ -th column equals to  $n_{ij}$ , is given by the hypergeometric distribution,

$$\begin{aligned} \mathcal{P}(n_{ij}, a_i, b_j|M) &= \mathcal{P}(b_j - n_{ij}, r - 1, r - i, b_j - 1) \\ &= \binom{r-i}{b_j - n_{ij}} \binom{i-1}{n_{ij} - 1} / \binom{r-1}{b_j - 1}. \end{aligned} \quad (39)$$

The hypergeometric distribution describes the probability of  $b_j - n_{ij}$  successes in  $b_j - 1$  draws without replacement where the finite population consists of  $r - 1$  elements, of which  $r - i$  are classified as successes. It is limited by the number of successes that must not exceed the limit of  $\max(0, i + b_j - r) \leq n_{ij} \leq \min(i, b_j)$ .

Similar, the distance  $\Delta y_i(M)$  between two consecutive ordered values is described by a binomial distribution,

$$\Delta y_i(n_{ij}, a_i, b_j|M) = \frac{1}{\mathcal{N}} \sum_{k=1}^{k_{\max}} \binom{r-k-1}{b_j-2} (y_{(i+k)} - y_{(i)}), \quad (40)$$

where the upper limit is given by  $k_{\max} = \min(n - b_j + 1, r - i)$  and  $\mathcal{N}$  is the normalization constant:

$$\mathcal{N} = \sum_{k=1}^{k_{\max}} \binom{r-k-1}{b_j-2}. \quad (41)$$

Summarizing all the single parts of Eq. 37, the final formula for the expected fraction of cumulative information under the assumption of the hypergeometric model of randomness is given by Eq. 42.

## B Monotonicity conditions for total cumulative mutual information

In the following we will prove that expected cumulative mutual information under the randomness assumption of variables  $\hat{\mathcal{J}}_0(Y;X)$  is monotonically increasing with respect to the number of features in the subset, i.e.,

$$\hat{\mathcal{J}}_0(Y;X) \leq \hat{\mathcal{J}}_0(Y;X') \text{ for } X \subset X' \subseteq \mathbf{X} \quad (43)$$

with  $X' = X \cup \{\chi\}$  and some  $\chi \notin X$ . For reference, we will closely follow the proof for the baseline correction term in the discrete case with mutual information [52].

Let the row and column marginals of  $Y, X, X'$  be  $a_i$  for  $i = 1 \dots R$ ,  $b_j$  for  $j = 1 \dots C$  and  $b'_j$  for  $j = 1 \dots C'$ , respectively. We note that  $C' > C$ . In order to show that

$$\sum_{M \in \mathcal{M}} \hat{\mathcal{J}}(Y; X|M) \mathcal{P}(Y; X|M) \leq \sum_{M' \in \mathcal{M}'} \hat{\mathcal{J}}(Y; X|M') \mathcal{P}(Y; X|M'). \quad (44)$$

we define a relation between the cumulative contingency tables  $\mathcal{M} = \mathcal{M}(Y; X)$  and  $\mathcal{M}' = \mathcal{M}(Y; X')$  via the projection operator  $\pi : \mathcal{M}' \rightarrow \mathcal{M}$ . The projection operator links the projection  $\pi : V(X') \rightarrow V(X)$  of values from  $X'$  to values of  $X$  defined by  $\pi(X') = X$  with the projection to the sets of cumulative contingency tables by finding the counts in the column corresponding to  $X \in V(X)$  of  $\pi(M')$  as the sum of the columns in  $M'$  corresponding to  $\pi^{-1}(X)$ . Therefore, it remains to show that for all  $M \in \mathcal{M}$  holds:

$$\hat{\mathcal{J}}(Y; X|M) \mathcal{P}(Y; X|M) \leq \sum_{M' \in \pi(M)} \hat{\mathcal{J}}(Y; X|M') \mathcal{P}(Y; X|M'). \quad (45)$$

Then, from the chain rule of cumulative mutual information [44, 97, 45], it follows that  $\hat{\mathcal{J}}(Y; X|M) \leq \hat{\mathcal{J}}(Y; X|M')$  for  $M = \pi(M')$ . Thus, showing the relation  $\mathcal{P}(Y; X|M) = \sum_{M' \in \pi(M)} \mathcal{P}(Y; X|M')$  concludes the proof. We will show the proof by contradiction.

Formally, let  $S_n$  denote the symmetric group of degree  $n$ , i.e.,  $S_n$  consists of all  $n!$  bijections  $\sigma : \{1 \dots n\} \rightarrow \{1 \dots n\}$ . For a bijection  $\sigma \in S_n$ , we denote the permuted version of  $Y$  as  $Y_\sigma$ . Then, for any cumulative contingency table  $N \in \mathcal{M}(Y; Z)$   $S_n[N] = \{\sigma \in S_n : M(Y_\sigma; Z) = N\}$  denotes the permutations that result in  $Z$ . Let  $\sigma \in S_n \setminus S_n[M]$ . This means that  $M_{ij}(Y; X) \neq M_{ij}(Y_\sigma; X)$  for at least one cell  $i, j$ . Further, denote the set of all indices of values of  $X'$  that are projected down to  $X$  by

$$\pi^{-1}(j) = \{j' : 1 \leq j' \leq C', \pi(X'_{j'}) = X_j\}, \quad (46)$$

for which, by definition, follows that

$$\sum_{j' \in \pi^{-1}(j)} M'_{ij'}(Y; X') \neq \sum_{j' \in \pi^{-1}(j)} M'_{ij'}(Y_\sigma; X'). \quad (47)$$

Since for at least one index  $j' \in \pi^{-1}(j)$  we get  $M'_{ij'}(Y; X') \neq M'_{ij'}(Y_\sigma; X')$ , we also find  $\sigma \notin S_n[M']$  and can conclude

$$S_n[M] \supseteq \bigcup_{M' \in \pi^{-1}(M)} S_n[M']. \quad (48)$$

Now let  $N' \in \mathcal{M}(Y; X')$  with  $\pi(N') \neq M$  and assume that  $S_n[M] \supset S_n[M']$ , i.e., there is a  $\sigma \in S_n[M] \cap S_n[N']$ . Let us denote  $N = \pi(N')$ . Since  $S_n[M] \cap S_n[N] = \emptyset$ , we know that  $\sigma \notin S_n[N]$ . However, it follows from Eq. 48 that  $\sigma \notin S_n[N']$  – a contradiction and, hence,

$$S_n[M] = \bigcup_{M' \in \pi^{-1}(M)} S_n[M'] \quad (49)$$

and

$$\mathcal{P}(Y; X|M) = \frac{|S_n[M]|}{|S_n|} = \sum_{M' \in \pi^{-1}(M)} \frac{|S_n[M']|}{|S_n|} = \sum_{M' \in \pi(M)} \mathcal{P}(Y; X|M'). \quad (50)$$

□

## C Gradient boosting decision trees

We used LightGBM [98], a recent modification of the gradient boosting decision trees algorithm [87]. LightGBM improves the efficiency and scalability without sacrificing performance. The following settings were used and were found by hyper-parameter tuning: number of leaves (`num_leaves`, 1% of the number of samples), number of iterations (`n_estimators`, 2000), and model depth (`max_depth`, -1).

During the training, i.e., the model optimization, we performed a regularization to automatically select the inflection point at which the performance of the test data set begins to decrease while the performance of the training data set continues to improve. The data set was partitioned into 10 groups (so-called folds), using 9 folds to generate the model and the remaining fold to test the model (10-fold cross validation). To reduce variability, we performed five rounds of cross-validation with different partitions and averaged the rounds to obtain an estimate of the model's predictive performance. We monitored the L1 and L2 norms [87,92] and simultaneously penalized the model optimization ("learning") process on the 9 folds to minimize the squared residuals and the complexity of the model (`eval_metric`, ['l1', 'l2\_root']), while stopping the learning process as soon as one metric of the remaining fold in the last  $n = 50$  rounds did not improved (`early_stopping_rounds`, 50).

### Conflict of interest

The authors declare that they have no conflict of interest.

### Software license

We implemented total cumulative mutual information in Python. Our Python-based implementation is part of B.R.'s doctoral thesis and is made publicly available under a Apache License 2.0.

### Data availability

All data and scripts involved in producing the results can be downloaded from GitHub (<https://github.com/sommerregen/tcmi>). An online tutorial to reproduce the main results presented in this work can also be found in the NOMAD Analytics Toolkit (<https://labdev-nomad.esc.rzg.mpg.de/jupyterhub/hub/user-redirect/tree/tutorials/tcmi/tcmi.ipynb>).

### Corresponding authors

Correspondence to Benjamin Regler or Luca M. Ghiringhelli.



## References

1. T. Hey, S. Tansley, K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009). The concept of a fourth paradigm was probably first discussed by J. Gray at a workshop 22on Januar 11, 2007 before he went missing at the Pacific on January 28, 2007.
2. I. Guyon, A. Elisseeff, *Journal of Machine Learning Research* **3**, 1157 (2003)
3. A.L. Blum, P. Langley, *Artificial Intelligence* **97**(1), 245 (1997). DOI 10.1016/S0004-3702(97)00063-5
4. R. Kohavi, G.H. John, *Artificial Intelligence* **97**(1), 273 (1997). DOI 10.1016/S0004-3702(97)00043-X
5. J.R. Koza, *Statistics and Computing* **4**(2), 87 (1994). DOI 10.1007/BF00175355
6. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, *Physical Review Letters* **114**, 105503 (2015). DOI 10.1103/PhysRevLett.114.105503
7. R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, *Phys. Rev. Materials* **2**(8), 083802 (11) (2018). DOI 10.1103/PhysRevMaterials.2.083802
8. L. Breiman, J. Friedman, C.J. Stone, R. Olshen, *Classification and regression trees* (Chapman and Hall/CRC, Boca Raton, FL, 1984)
9. A.H. Land, A.G. Doig, *Econometrica* **28**(3), 497 (1960). DOI 10.2307/22910129
10. Narendra, Fukunaga, *IEEE Transactions on Computers* **C-26**(9), 917 (1977). DOI 10.1109/TC.1977.1674939
11. J. Clausen, Branch and bound algorithms – principles and examples. Tech. rep., Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK2100 Copenhagen, Denmark (1999)
12. D.R. Morrison, S.H. Jacobson, J.J. Sauppe, E.C. Sewell, *Discrete Optimization* **19**, 79 (2016). DOI 10.1016/j.disopt.2016.01.005
13. Y. Huhtala, J. Krkkinen, P. Porkka, H. Toivonen, *The Computer Journal* **42**(2), 100 (1999). DOI 10.1093/comjnl/42.2.100
14. A.W. Whitney, *IEEE Transactions on Computers* **C-20**(9), 1100 (1971). DOI 10.1109/T-C.1971.223410
15. P. Pudil, J. Novoviov, J. Kittler, *Pattern Recognition Letters* **15**(11), 1119 (1994). DOI 10.1016/0167-8655(94)90127-9
16. T. Marill, D. Green, *IEEE Transactions on Information Theory* **9**(1), 11 (1963). DOI 10.1109/TIT.1963.1057810
17. H. Peng, F. Long, C. Ding, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(8), 1226 (2005). DOI 10.1109/TPAMI.2005.159
18. U.M. Khaire, R. Dhanalakshmi, *Journal of King Saud University - Computer and Information Sciences* (2019). DOI 10.1016/j.jksuci.2019.06.012
19. M. Basseville, *Signal Processing* **18**(4), 349 (1989). DOI 10.1016/0165-1684(89)90079-0
20. H. Almuallim, T.G. Dietterich, *Artificial Intelligence* **69**(1), 279 (1994). DOI 10.1016/0004-3702(94)90084-1
21. M. Modrzejewski, in *Machine Learning: ECML-93*, ed. by P.B. Brazdil (Springer Berlin Heidelberg, Berlin, Heidelberg, 1993), pp. 213–226. DOI 10.1007/3-540-56602-3\_138
22. A. Arauzo-Azofra, J.M. Benitez, J.L. Castro, *Journal of Intelligent Information Systems* **30**(3), 273 (2008). DOI 10.1007/s10844-007-0037-0
23. J.R. Vergara, P.A. Estévez, *Neural Computing and Applications* **24**(1), 175 (2014). DOI 10.1007/s00521-013-1368-0
24. C.E. Shannon, *The Bell System Technical Journal* **27**(3), 379 (1948). DOI 10.1002/j.1538-7305.1948.tb01338.x
25. H.V. Nguyen, E. Müller, J. Vreeken, F. Keller, K. Böhm, *CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection* (Proc. SIAM International Conference on Data Mining, 2013), chap. 21, pp. 198–206. DOI 10.1137/1.9781611972832.22
26. H.V. Nguyen, E. Müller, J. Vreeken, P. Efron, K. Bhm, in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, vol. 32, ed. by T. Jebara, E.P. Xing (JMLR Workshop and Conference Proceedings, Beijing, China, 2014), vol. 32, pp. 775–783
27. D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, *Science* **334**(6062), 1518 (2011). DOI 10.1126/science.1205438
28. H.V. Nguyen, P. Mandros, J. Vreeken, *Universal Dependency Analysis* (Society for Industrial and Applied Mathematics, 2016), pp. 792–800. Proceedings. DOI 10.1137/1.9781611974348.89

29. Y. Wang, S. Romano, V. Nguyen, J. Bailey, X. Ma, S.T. Xia, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (2017)
30. N. Kwak, Chong-Ho Choi, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(12), 1667 (2002). DOI 10.1109/TPAMI.2002.1114861
31. T.W.S. Chow, D. Huang, *IEEE Transactions on Neural Networks* **16**(1), 213 (2005). DOI 10.1109/TNN.2004.841414
32. P.A. Estevez, M. Tesmer, C.A. Perez, J.M. Zurada, *IEEE Transactions on Neural Networks* **20**(2), 189 (2009). DOI 10.1109/TNN.2008.2005601
33. Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, W. Pedrycz, *Expert Systems with Applications* **38**(9), 10737 (2011). DOI 10.1016/j.eswa.2011.01.023
34. M. Bennisar, Y. Hicks, R. Setchi, *Expert Systems with Applications* **42**(22), 8520 (2015). DOI 10.1016/j.eswa.2015.07.007
35. E. Fouché, K. Böhm, in *Proceedings of the 31st International Conference on Scientific and Statistical Database Management (ACM, New York, NY, USA, 2019)*, SSDBM '19, pp. 13–24. DOI 10.1145/3335783.3335795
36. K. Pearson, *Philosophical Transactions of the Royal Society of London Series A* **187**, 253 (1896). DOI 10.1098/rsta.1896.0007
37. C. Spearman, *The American Journal of Psychology* **15**(1), 72 (1904). DOI 10.2307/1412159
38. G.J. Székely, M.L. Rizzo, N.K. Bakirov, *The Annals of Statistics* **35**(6), 2769 (2007). DOI 10.1214/009053607000000505
39. G.J. Székely, M.L. Rizzo, *The Annals of Statistics* **42**(6), 2382 (2014). DOI 10.1214/14-AOS1255
40. D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (Wiley, New York, 1982). DOI 10.1002/9780470316849
41. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, vol. 1 (Chapman and Hall/CRC, New York, 1986)
42. L.F. Kozachenko, N.N. Leonenko, *Problemy Peredachi Informatsii* **23**(2), 9 (1987)
43. F. Keller, E. Müller, K. Böhm, in *28th International Conference on Data Engineering (IEEE, 2012)*, pp. 1037–1048. DOI 10.1109/ICDE.2012.88
44. M. Rao, Y. Chen, B.C. Vemuri, F. Wang, *IEEE Transactions on Information Theory* **50**(6), 1220 (2004). DOI 10.1109/TIT.2004.828057
45. M. Rao, *Journal of Theoretical Probability* **18**(4), 967 (2005). DOI 10.1007/s10959-005-7541-3
46. A.D. Crescenzo, M. Longobardi, in *Methods and Models in Artificial and Natural Computation. A Homage to Professor Mira's Scientific Legacy: Third International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2009, Santiago de Compostela, Spain, June 22-26, 2009, Proceedings, Part I*, ed. by J. Mira, J.M. Ferrández, J.R. Álvarez, F. de la Paz, F.J. Toledo (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009), pp. 132–141. DOI 10.1007/978-3-642-02264-7\_15
47. A.D. Crescenzo, M. Longobardi, *Journal of Statistical Planning and Inference* **139**(12), 4072 (2009). DOI 10.1016/j.jspi.2009.05.038
48. N.X. Vinh, J. Epps, J. Bailey, in *Proceedings of the 26th Annual International Conference on Machine Learning (ACM, New York, NY, USA, 2009)*, ICML '09, pp. 1073–1080. DOI 10.1145/1553374.1553511
49. N.X. Vinh, J. Epps, J. Bailey, *Journal of Machine Learning Research* **11**, 2837 (2010)
50. S. Romano, J. Bailey, V. Nguyen, K. Verspoor, in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, vol. 32, ed. by T. Jebara, E.P. Xing (JMLR Workshop and Conference Proceedings, Beijing, China, 2014), vol. 32, pp. 1143–1151
51. Y. Zheng, C.K. Kwok, *Entropy* **13**(4), 860 (2011). DOI 10.3390/e13040860
52. P. Mandros, M. Boley, J. Vreeken, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York, NY, USA, 2017)*, KDD '17, pp. 355–363. DOI 10.1145/3097983.3098062
53. S. Kullback, R.A. Leibler, *The Annals of Mathematical Statistics* **22**(1), 79 (1951)
54. S. Kullback, *Information Theory and Statistics* (John Wiley and Sons, New York, 1959)
55. C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, vol. III (Illinois Press, 1949)
56. T.M. Cover, J.A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006)
57. C. Coombs, R. Dawes, A. Tversky, *Mathematical Psychology: An Elementary Introduction* (Prentice-Hall, Englewood Cliffs, NJ, 1970)
58. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1988). DOI 10.1137/1031025

59. J.V. White, S. Steingold, C. Fournelle, in *Computing Science and Statistics: Computational Biology and Informatics - Proceedings of the 36th Symposium on the Interface*, vol. 36, ed. by Y.H. Said, D.J. Marchette, J.L. Solka (Baltimore, Maryland, 2004), vol. 36
60. M. Reimherr, D.L. Nicolae, *Statistical Science* **28**(1), 116 (2013). DOI 10.1214/12-STS405
61. D. Pfizner, R. Leibbrandt, D. Powers, *Knowledge and Information Systems* **19**(3), 361 (2008). DOI 10.1007/s10115-008-0150-6
62. D. Xu, Y. Tian, *Annals of Data Science* **2**(2), 165 (2015). DOI 10.1007/s40745-015-0040-1
63. U. Fayyad, K. Irani, in *Proceedings of the 13th Int. Joint Conference on Artificial Intelligence* (Morgan Kaufmann, 1993), pp. 1022–1027
64. J. Dougherty, R. Kohavi, M. Sahami, in *Machine Learning: Proceedings of the Twelfth International Conference*, ed. by A. Prieditis, S.J. Russell (Morgan Kaufmann, 1995), pp. 194–202
65. H.V. Nguyen, E. Müller, J. Vreeken, K. Böhm, *Data Mining and Knowledge Discovery* **28**(5), 1366 (2014). DOI 10.1007/s10618-014-0350-5
66. D. Garcia, *Computational Statistics & Data Analysis* **54**(4), 1167 (2010). DOI 10.1016/j.csda.2009.09.020
67. A. Bernacchia, S. Pigolotti, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(3), 407 (2011). DOI 10.1111/j.1467-9868.2011.00772.x
68. T.A. OBrien, W.D. Collins, S.A. Rauscher, T.D. Ringler, *Computational Statistics & Data Analysis* **79**, 222 (2014). DOI 10.1016/j.csda.2014.06.002
69. T.A. OBrien, K. Kashinath, N.R. Cavanaugh, W.D. Collins, J.P. OBrien, *Computational Statistics & Data Analysis* **101**, 148 (2016). DOI 10.1016/j.csda.2016.02.014
70. V. Glivenko, *Gion. Ist. Ital. Attuari*, **4**, 92 (1933)
71. F.P. Cantelli, *Giorn. Ist. Ital. Attuari* **4**(421–424) (1933)
72. S. Yu, J.C. Principe, *Entropy* **21**(1) (2019). DOI 10.3390/e21010099
73. H.O. Lancaster, *The Chi-squared Distribution* (Wiley & Sons, Inc., New York, 1969)
74. D. Koller, M. Sahami, in *Proceedings of the 13th International Conference on Machine Learning* (Bari, Italy, 1996), pp. 284–292
75. J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, *Soft Computing* **13**(3), 307 (2009). DOI 10.1007/s00500-008-0323-y
76. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garca, L. Snchez, F. Herrera, *Journal of Multiple-Valued Logic and Soft Computing* **17**(2–3), 255 (2011)
77. D. Dua, C. Graff. UCI machine learning repository (2017)
78. L.M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S.V. Levchenko, C. Draxl, M. Scheffler, *New Journal of Physics* **19**(2), 023017 (2017). DOI 10.1088/1367-2630/aa57bf
79. J.H. Friedman, *The Annals of Statistics* **19**(1), 1 (1991). DOI doi.org/10.1214/aos/1176347963
80. I.C. Yeh, *Cement and Concrete Research* **28**(12), 1797 (1998). DOI 10.1016/S0008-8846(98)00165-3
81. P. Cortez, A. Morais, in *New Trends in Artificial Intelligence*, ed. by J. Neves, M.F. Santos, J. Machado (Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, Guimaraes, Portugal, 2007), pp. 512–523
82. J.A. Van Vechten, *Physical Review* **182**, 891 (1969). DOI 10.1103/PhysRev.182.891
83. J.C. Phillips, *Reviews of Modern Physics* **42**, 317 (1970). DOI 10.1103/RevModPhys.42.317
84. A. Zunger, *Physical Review B* **22**, 5839 (1980)
85. D. Pettifor, *Solid State Communications* **51**(1), 31 (1984). DOI 10.1016/0038-1098(84)90765-8
86. Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J.R. Chelikowsky, W. Andreoni, *Physical Review B* **85**, 104104 (2012). DOI 10.1103/PhysRevB.85.104104
87. J.H. Friedman, *The Annals of Statistics* **29**(5), 1189 (2001)
88. A. Natekin, A. Knoll, *Frontiers in neurorobotics* **7**, 21 (2013). DOI 10.3389/fnbot.2013.00021.24409142[pmid]
89. M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, *J. Mach. Learn. Res.* **15**(1), 3133 (2014)
90. R. Couronné, P. Probst, A.L. Boulesteix, *BMC Bioinformatics* **19**(1), 270 (2018). DOI 10.1186/s12859-018-2264-5
91. F. Lu, E. Petkova, *Statistics in Medicine* **33**(3), 401 (2014). DOI 10.1002/sim.5937
92. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning, Springer Texts in Statistics*, vol. 103 (Springer, New York, 2013). DOI 10.1007/978-1-4614-7138-7
93. D.H. Wolpert, *Neural Computation* **8**(7), 1341 (1996). DOI 10.1162/neco.1996.8.7.1341
94. D.H. Wolpert, *Neural Computation* **8**(7), 1391 (1996). DOI 10.1162/neco.1996.8.7.1391
95. D.H. Wolpert, W.G. Macready, No free lunch theorems for search. Technical Report SFI-TR-95-02-010 10, Santa Fe Institute (1995)

- 
96. D.H. Wolpert, W.G. Macready, *IEEE Transactions on Evolutionary Computation* **1**(1), 67 (1997). DOI 10.1109/4235.585893
  97. F. Wang, B.C. Vemuri, M. Rao, Y. Chen, *A New & Robust Information Theoretic Measure and Its Application to Image Alignment* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003), pp. 388–400. DOI 10.1007/978-3-540-45087-0\_33
  98. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, in *Advances in Neural Information Processing Systems 30*, ed. by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Curran Associates, Inc., 2017), pp. 3146–3154