Bioinformatics Advances, 2024, **00**, vbad175
https://doi.org/10.1093/bioadv/vbad175
Advance Access Publication Date: 29 November 2023
Original Article

Systems biology

BannMI deciphers potential *n*-to-1 information transduction in signaling pathways to unravel message of intrinsic apoptosis

Bettina Schmidt [1,2,*], Christine Sers[3,4], Nadja Klein [1,5]

[1]Research Center Trustworthy Data Science and Security, Universitätsallianz Ruhr, 44227 Dortmund, North Rhine-Westphalia, Germany
[2]Department of Computer Science, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
[3]Institute of Pathology, Charité–Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, 10117 Berlin, Germany
[4]Department of Biology, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
[5]Department of Statistics, Technische Universität Dortmund, 44227 Dortmund, North Rhine-Westphalia, Germany

*Corresponding author. Research Center Trustworthy Data Science and Security, Universitätsallianz Ruhr and Department of Statistics, Technische Universität Dortmund, Joseph-von-Fraunhofer-Straße 25, 44227 Dortmund, North Rhine-Westphalia, Germany. E-mail: qwerty11.bs@gmx.de

Associate Editor: Thomas Lengauer

Abstract

Motivation: Cell fate decisions, such as apoptosis or proliferation, are communicated via signaling pathways. The pathways are heavily intertwined and often consist of sequential interaction of proteins (kinases). Information integration takes place on the protein level via *n*-to-1 interactions. A state-of-the-art procedure to quantify information flow (edges) between signaling proteins (nodes) is network inference. However, edge weight calculation typically refers to 1-to-1 interactions only and relies on mean protein phosphorylation levels instead of single cell distributions. Information theoretic measures such as the mutual information (MI) have the potential to overcome these shortcomings but are still rarely used.

Results: This work proposes a Bayesian nearest neighbor-based MI estimator (BannMI) to quantify *n*-to-1 kinase dependency in signaling pathways. BannMI outperforms the state-of-the-art MI estimator on protein-like data in terms of mean squared error and Pearson correlation. Using BannMI, we analyze apoptotic signaling in phosphoproteomic cancerous and noncancerous breast cell line data. Our work provides evidence for cooperative signaling of several kinases in programmed cell death and identifies a potential key role of the mitogen-activated protein kinase p38.

Availability and implementation: Source code and applications are available at: https://github.com/zuiop11/nn_info and can be downloaded via Pip as Python package: nn-info.

1 Introduction

Signal propagation in molecular networks can be abstracted to a set of interfaces of *n*-to-1 communication in which many senders concurrently “talk” to one receiver. Examples of such interfaces are the posttranslational modification of proteins or the orchestration of gene expression via signaling pathways. In more detail, this means that proteins, such as tumor suppressor p53 are phosphorylated at several residues simultaneously upon activation (Lavin and Gueven 2006, Liebl and Hofmann 2019). An *n*-to-1 example on the genetical level comes from the B-cell lymphoma 2 (Bcl-2) family, where the expression of various subsets of this family is governed by several signaling pathways. This can be fatal for a cell, as an imbalance of particular Bcl-2 members initiates the cell fate decision of apoptosis (Ramesh *et al.* 2009, Wang *et al.* 2022).

In this work, we provide evidence that this complex interplay cannot be investigated sufficiently by consideration of 1-to-1 signals only. However, predictions of state-of-the-art network inference methods such as `STASNet` (Dorel *et al.* 2018) or `CellOracle` (Kamimoto *et al.* 2023) are based on the analysis of 1-to-1 signals. In the following motivational example, we provide evidence how this can miss crucial interactions. To solve the issue for *n*-to-1 network interfaces in general, we present BannMI—a Bayesian nearest neighbor (NN)-based mutual information (MI) estimator. While BannMI can be applied on any data drawn from continuous variables, the focus of this work is its application on phosphoproteomic data.

1.1 Motivational example

Let $X = \{X_1, X_2, \ldots, X_n\}$, $X_i = (X_{i,1}, X_{i,2})$ be an *n*-sized sample of 2D uniform, independent and identically distributed (i.i.d.) random variables with no componentwise dependencies. Further, let $Z_i = \phi_2(X_i) + Y_i$, where ϕ_2 is the probability density function (pdf) of a standard 2D Gaussian distribution $\mathcal{N}_2(0, \Sigma_2)$, and $Y_i \overset{iid}{\sim} \mathcal{N}(0, 0.01)$, $i = 1, \ldots, n$. Intuitively speaking, we expect that X_i is more informative for Z_i if the

Received: August 29, 2023; **Revised:** October 28, 2023; **Editorial Decision:** November 24, 2023; **Accepted:** November 28, 2023

Table 1. Quantifying multivariate data dependency.[a]

Measure	Uncorrelated	Correlated
Pearson correlation	−0.633	−0.167
Univariate MI	0.673	0.711
Bivariate MI	1.184	3.316

[a] Linear correlation between $X_{\cdot,1}, Z$ (first row), univariate/bivariate MI between $X_{\cdot,1}, Z$ (second row)/X, Z (third row), respectively. MI values are derived via BannMI.

covariance matrix Σ_2 of ϕ_2 has componentwise correlations (plotted data of this example is shown in the Supplement). This intuition, which corresponds to a multivariate data dependency, is captured by the bivariate MI value, see Table 1 (last row). However, as the first two rows of the table illustrate, it is not possible to capture this effect via Pearson correlation or univariate MI. In this work, we take advantage of this new multivariate perspective to further understand potential information transduction in apoptosis.

Our case study on breast cancer data (Tognetti *et al.* 2021) provides evidence for such cooperative signaling in the apoptotic process. Further, via separate investigation of cell fate "phenotypes", we identify a potentially fateful role of phosphorylated p38 in the apoptotic process. In addition, when comparing the potential, apoptotic signals of control cells and cancer cells, we find that the latter is significantly reduced. By that, we directly address Zielińska and Katanaev (2019), who suggest information theory to analyze signal alterations in cancer cells.

The undeniable asset of multivariate quantification of dependency has recently been demonstrated by the work of Erman (2023). Here, the author analyzed the molecular dynamics of a structural domain found in a broad variety of signaling proteins. To do so, MI was approximated with help of tensor Hermite polynomials. Here, it was found that in particular the dynamics of triplets of residues and not pairs of residues were altered in the presence of mutations. Uda *et al.* (2013) provides a rare example where n-to-1 communication is quantified in systems biology. To do so, empirical MI estimation based on descriptive histograms was used. More recently, Wada *et al.* (2021) concluded in their review that intracellular cell-to-cell heterogeneity, which is caused by the stochasticity of signal propagation, serves as information and not as noise in signaling. For their analysis, once again, the authors used mutual information. In the review Uda (2020), the author discusses diverse use-cases of MI in systems biology in theory. Finally, Karolak *et al.* (2021) discuss applications of information theory in systems biology with focus on cancer.

The paper is structured as follows: Section 2 (Methods) is a formal introduction of the information theoretical measures entropy, KLD, MI and channel capacity (CC). In this section we also provide a derivation of BannMI. In Section 3, we benchmark BannMI against state-of-the-art NN methods for MI on synthetic data with data generating processes that try to mimic the characteristics of phosphoproteomic data. In our case study we perform 1-to-1 MI/n-to-1 MI signal analysis for apoptosis on breast cell lines *per se* (Section 4.1). Finally, we conclude with a cell fate phenotypic signal analysis (Section 4.2).

2 Methods

BannMI estimates MI via the KLD. Both quantities are closely related to entropy, which forms the core of information theory. This familiarity explains why state-of-the-art estimators for entropy (Kozachenko and Leonenko 1987), MI (KSG; Kraskov *et al.* 2004), and KLD (Wang *et al.* 2009) apply the same pdf approximation. This approximation is rooted in Lebesgue's differentiation theorem and applies NN distances. We describe Lebesgue's approximation and derive the three NN distance-based estimators in the Supplementary Material. In contrast, our method approximates the ratio of two pdf's and therefore is based on NN ratios.

In this section, we briefly introduce the information theoretical measures. Then, we review the KLD NN ratio estimator proposed by Noshad *et al.* (2017), which our BannMI extends to a Bayesian framework.

2.1 Preliminary

2.1.1 Notation

Let P be a continuous probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \in \mathbb{N} \setminus \{0\}$ with probability density function (pdf) p, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-Algebra on $\mathbb{R}^d$. Let $x \in \mathbb{R}^d$ and n_x be an open neighborhood around x. Then, the support of P is defined as

$$\mathcal{S} := \{x \in \mathbb{R}^d \,|\, \forall n_x \in \mathcal{B}(\mathbb{R}^d) : P(n_x) > 0\}.$$

Let Q be a continuous probability measure on the same support with pdf q. Then, the Radon–Nikodym derivative of p with respect to q is

$$\tau(x) := \frac{p(x)}{q(x)} = \left(\frac{q(x)}{p(x)}\right)^{-1}, \quad \forall x \in \mathcal{S}.$$

Furthermore, the differential entropy of P is $H(p) := -\int_{\mathcal{S}} \log(p)p(x)dx$, where dx is the Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Further, the KLD of P with respect to Q is

$$D(pq) := \int_{\mathcal{S}} \log\left(\frac{p(x)}{q(x)}\right)p(x)dx = \int_{\mathcal{S}} \log\left(\tau(x)\right)p(x)dx.$$

Now, let dimension $d \geq 2$ such that $P = P_1 \times P_2$. Note that if $d > 2$, at least one of the marginal distributions P_1, P_2 is multidimensional. Further, let p_1 and p_2 be the respective pdfs. The MI of P_1 with respect to P_2 (and vice versa) is

$$I(p_1, p_2) := \int_{\mathcal{S}} \log\left(\frac{p(x)}{p_1(x_1)p_2(x_2)}\right)p(x)dx. \tag{1}$$

It is easy to see, that for $q(x) = p_1(x_1)p_2(x_2)$, the KLD is equivalent to MI. Further, channel capacity (CC) is defined as the maximum MI, given all feasible marginal probability measures P_2. That is

$$C(p_1, p_2) := \max_{p_2} I(p_1, p_2).$$

2.2 Estimation via NN ratios

2.2.1 Notation

Let $X = \{X_1, X_2, \ldots, X_n\}$ with $n \in \mathbb{N}$ be an i.i.d. sample with $X_i \sim P$ and let $\|\cdot\|$ be a norm on $\mathbb{R}^d \supset \mathcal{S}$. Further, let $Y = \{Y_1, Y_2, \ldots, Y_n\}$ be another i.i.d. sample with $Y_i \sim Q$, with the same support as P. Then, $nn_{i,1} \in (X \cup Y) \setminus \{X_i\}$ is the first NN of X_i in the joint sample if and only if

$$\|X_i - nn_{i,1}\| \le \|X_i - Z_j\| \ \forall Z_j \in X \cup Y \setminus \{X_i\}.$$

For every positive integer $k < 2n$, the kth NN $nn_{i,k}$ of X_i is defined accordingly.

In the joint sample $X \cup Y$, let $R_{i,k} = \{nn_{i,1}, \dots, nn_{i,k}\}$ be the set of the k NNs of X_i. Further, define the number of NNs of X_i in X and Y as $N_{i,k} = |R_{i,k} \cap X|$ and $M_{i,k} = k - N_{i,k}$, respectively. For $0 < \epsilon \le 1$, an f-divergence estimator is

$$\hat{D}^N_{n,k}(X,Y) := \max\left\{\frac{1}{n}\sum_i^n \tilde{f}\left(\frac{M_{i,k}}{N_{i,k}+1}\right), 0\right\} \quad (2)$$
$$\text{with} \quad \tilde{f}(x) := \max\{f(x), f(\epsilon)\}.$$

If we insert $f(x) = -\log(x)$, an estimator for the inverse KLD is derived. If we allow $nn_{i,1} \in X \cup Y$, then the NN of X_i is X_i itself and the $+1$ in the denominator of [Equation (2)] becomes redundant. The interpretation of this NN ratio is as follows: Radon–Nikodyn derivative $\tau(X_i)$ is considered as odds of a 0–1 random variable B_i, which is drawn from P rather than Q for a fixed X_i. That is

$$B_i \sim \text{Ber}_{\theta_i} \quad \text{with} \quad \theta_i := \frac{p(X_i)}{p(X_i)+q(X_i)}.$$

Using $R_{i,k}$, the parameter θ_i can be estimated via its maximum likelihood estimator (MLE) given by $\hat{\theta}_{i,\text{MLE}} := N_{i,k}/k$. For $M_{i,k} \neq 0$, an estimator of the Radon–Nikodym derivative is thus

$$\hat{\tau}(X_i) = \frac{\hat{\theta}_{i,\text{MLE}}}{1-\hat{\theta}_{i,\text{MLE}}} = \frac{N_{i,k}}{M_{i,k}}. \quad (3)$$

We have shown that [Equation (2)] is based on MLE estimation. But the frequentist approach can be inferior to a Bayesian approach in cases of small sample size and biased data (Gelman *et al.* 2015). In the application considered, k defines $R_{i,k}$, which is the dataset used for MLE estimation for each $i = 1, \dots, n$. Here, the choice of k is a tradeoff between sample size and biasedness of the data: For a small k, $|R_{i,k}|$ is small but the neighboring points $nn_{i,j}$ with $j \in \{1, 2, \dots, k\}$ are likely to be "close" to X_i. This implies $\frac{p(nn_{i,j})}{p(nn_{i,j})+q(nn_{i,j})}$ to be close to θ_i, which is the parameter of interest. On the other hand, for a large k, sample size is increased at the cost of biased data. Those drawbacks motivated our Bayesian approach.

2.3 A Bayesian NN-based KLD estimator

As shown in [Equation (3)], the issue of KLD estimation can be reduced to estimation of a success-parameter θ_i, for every data point X_i in the sample X. In a Bayesian framework, estimation of θ_i is a statistic of the Beta-binomial distribution. We choose a natural, conjugate prior $\theta_i \sim \text{Beta}(\alpha, \beta)$. In the following, we refer to its parameters as $\hat{\alpha}$ and $\hat{\beta}$. This leads to a NN-based Bayesian KLD estimator

$$\hat{D}^B_{n,k}(X,Y) := \frac{1}{n}\sum_{i=1}^n \log\left(\frac{N_{i,k}+\hat{\alpha}}{M_{i,k}+\hat{\beta}}\right). \quad (4)$$

The posterior distribution for each θ_i is again Beta-distributed with parameters $\alpha_i = \hat{\alpha}+N_{i,k}$ and $\beta_i = \hat{\beta}+M_{i,k}$. By that, a posterior mean estimate of θ_i is

$$\hat{\theta}_{i,\text{B}} = \frac{\alpha_i}{\alpha_i+\beta_i} = \frac{\hat{\alpha}+N_{i,k}}{\hat{\alpha}+\hat{\beta}+k}. \quad (5)$$

Now, plugging $\hat{\theta}_{i,\text{B}}$ into the formula leads to

$$\hat{\tau}_{\text{B}}(X_i) = \frac{\hat{\theta}_{i,\text{B}}}{1-\hat{\theta}_{i,\text{B}}} = \frac{N_{i,k}+\hat{\alpha}}{M_{i,k}+\hat{\beta}}.$$

Remark 1 In Section 3, we test an empirically derived set of $\{\hat{\alpha}, \hat{\beta}\}$ using the method of moments as proposed by Gelman et al. (2015). See the Supplementary Material for further details.

2.4 BannMI—a Bayesian NN-based MI estimator

As derived in Section 2.1, the MI is an application of KLD, when $Q = P_1 \times P_2$ and $q(x_1, x_2) = p_1(x_1)p_2(x_2)$. Accordingly, one can estimate the MI between the lower dimensional subsets $X^1 = \{X_{1,1}, \dots, X_{n,1}\}$ and $X^2 = \{X_{1,2}, \dots, X_{n,2}\}$ of X via a random shuffling of elements in X^2 such that $\tilde{X}^2 = \{X_{(1),2}, X_{(2),2}, \dots, X_{(n),2}\}$ is independent from X^1. Our BannMI is then defined as

$$\hat{I}^B_{n,k}(X^1, X^2) := \hat{D}^B_{n,k}((X^1, X^2), (X^1, \tilde{X}^2)). \quad (6)$$

By making use of this equivalence between KLD and MI, our BannMI in [Equation (6)] allows us to quantify information propagation in molecular networks. Please see the Supplement for its detailed application in this context.

3 Benchmark study

In this section, we conduct three studies to benchmark the performance of BannMI against selected competitors. These are the state-of-the-art NN MI estimator KSG (Kraskov *et al.* 2004) and the KLD estimator based on NN distances (Wang *et al.* 2009); see the Supplementary Material for details on both approaches. As in Equation (1), we apply the latter as estimator for the MI and refer to it as WMI. Furthermore, we use the frequentist NN ratio-based KLD estimator as presented in Equation (2) and apply it for MI estimation (NMI). Here, we may note that to the best of our knowledge both KLD estimators have not been applied as MI estimators so far. In particular, the utilization of Equation (2) as MI estimator requires several steps, such as the choice of the f-function, the choice of a suited ϵ parameter (see Supplementary Material) as well as the approach of data shuffling (see Section 2.4). The reader finds our collection of new MI estimators as well as the established KLD estimators implemented in our Python package `nn-info`.

Due to our precise application goal of MI estimation on phosphoproteomic data, the aim of this paper is not to construct an estimator that performs well on any feasible pair of distributions, in any dimensionality and preferably on any sample size n. Rather, in this benchmark we focus on good (qualitative) performance on simulated data that shares the characteristics of phosphoproteomic data. Some of those characteristics are a sample size with a magnitude of 10^4, nonnegativity and skewness of the data and a high componentwise dependency.

Optimal parameter settings of the implemented algorithms were either derived in the KLD benchmark that can be found in the Supplementary Material. Or/and, the parameters were further tested for optimality in this study. Here, we point out the importance of the nearest neighbor parameter k. As NMI/BannMI are both nearest neighbor ratio estimators on the one hand, and WMI/KSG are both nearest neighbor distance estimators on the other hand, the optimal choice of k varies greatly. As in the KLD benchmark, performance of BannMI with empirically derived hyperparameters was comparable to a setting where $\hat{\alpha} = \hat{\beta} = 0.1$, we include both approaches in the following procedures and refer to the latter as "uBannMI". The following benchmark study was conducted on three data generating processes (DGP)s.

3.1 Multivariate Gaussian distribution

DGP: First, we derived 62 covariance matrices from phosphoproteomic data presented by Tognetti *et al.* (2021) for $d \in \{2, 5, 10\}$ (all cell lines, condition: EGF stimulation, time point $t = 0$). Next, we sampled centered Gaussian data from those covariance matrices $\Sigma \in \mathbb{R}^{d \times d}$. We tested performances for two different choices of k which produced best results in the KLD benchmark procedure (see Supplementary Material).

Main results: We find that in the Gaussian application, the empirical version of BannMI performs second best after KSG with respect to MSE/standard deviation and Pearson correlation. See extensive results in the Supplementary Material. Figure 1 (left) depicts an performance example of the estimators for $d = 5$ and $k = 10$ (BannMI, uBannMI, NMI), $k = 1$ (KSG) and $k = 2$ (WMI).

3.2 Multivariate skew Gaussian distribution

DGP: Phosphoproteomic data are likely to be skewed and on $\mathbb{R}^+$. To take this setting into account, skew Gaussian data are simulated for 2D as described in Azzalini and Valle (1996). We apply the dependency parameter $\delta = (\delta_1, \delta_2)^\top$ as follows: While $\delta_2 = 0.5$ remains fixed, we perform experiments for $\delta_1 \in \{0.0, 0.1, \ldots, 0.8\}$. Then, we sample Y_0 from a 1D standard Gaussian distribution and $Y = (Y_1, Y_2)^\top$ from a 2D Gaussian distribution with 0-mean vector and covariance matrix Σ_2. As in Section 3.1, Σ_2 is derived from a cell line phosphoprotein expressions. For $j \in \{1, 2\}$, we finally compute the skew Gaussian random variables $X_j = \delta_j|Y_0|+(1-\delta_j^2)^{\frac{1}{2}}Y_j$ with dispersion matrix Σ_s and skewness parameter α.

Main results: Figure 1 (center) shows results for increasing δ_1 (x-axis). The plot shows mean values and standard deviations for all estimators. Here, Pearson correlation to the numerically estimated MI is highest for BannMI/NMI (see Fig. 1, right). However due to its lower standard deviation, the choice of BannMI is favorable.

3.3 Bivariate exponential distributions

DGP: So far, we tested our MI estimators for componentwise correlated symmetric and skewed data. Now, we further approximate the characteristics of phosphoproteomic data by consideration of nonlinear correlated data. Figure 2 (right) shows expressions of phosphoproteins RB and 4E-BP1 of the Tognetti *et al.* (2021) dataset. Both phosphoproteins represent an "either-or" scenario (XOR), which resembles the joint expression of two exponential variables, as shown next to it.

Gumbel (1960) suggested a dependency parameter $\delta \in (0, 1)$ for two exponential variables, such that the joint cumulative distribution function is $F_\delta(x, y) = 1-e^{-x}-e^{-y}+e^{-x-y-\delta xy}$. To simulate data, we used the No-U-Turn-Sampler (NUTS; Hoffman and Gelman 2014) from the PyMC3 package. We sampled bivariate random variables from F_δ with exponential priors for $\delta \in \{0.0, 0.1, \ldots, 0.8\}$. An approximation of the true MI value is derived via numerical integration.

Main results: Figure 1 (center) shows that the two NN distance-based MI estimators, KSG and particularly WMI, perform poorly for increasing dependency parameter δ (x-axis). With respect to standard deviations and the high Pearson correlation toward the numerically derived estimate, BannMI once again performs superior to NMI (Fig. 1, right).

4 Case study on CyTOF data

Motivated by our biological reasoning, we have so far postulated n-to-1 signaling and suggested MI for its analysis in single cells. This led to our Bayesian MI estimator based on NN ratios which is well suited for MI estimation on phosphoproteomic data. Next, we apply the algorithm on suitable *in vitro* data. Suitable in this sense means that firstly phosphoproteins measured should provide information about the activation/deactivation of signaling pathways; and secondly

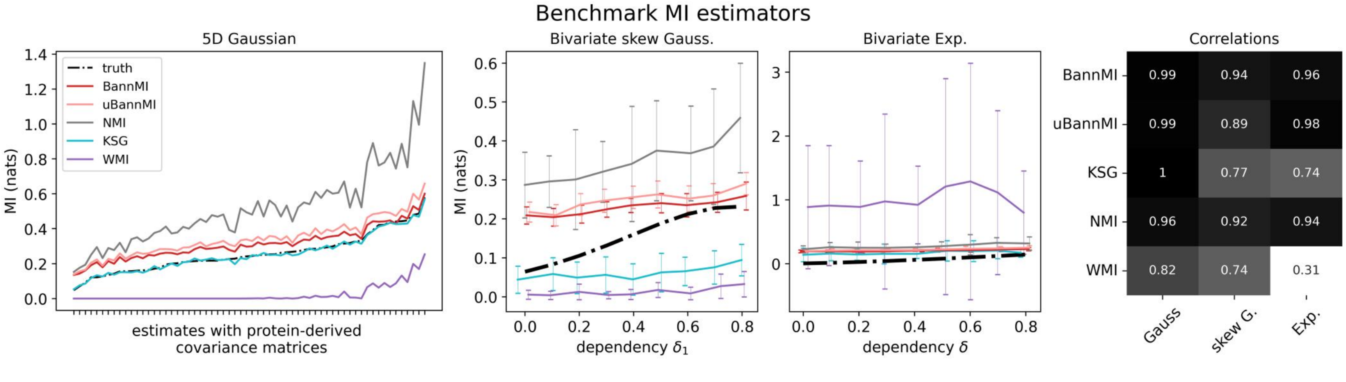


Figure 1. Left: Performance of MI estimators on 5D Gaussian data sampled from 62 protein-like covariance matrices (x-axis), shown together with the respective analytical MI values. Each estimate is a mean of 25 computations with sample size $n = 1000$. Center: Performance of MI estimators on bivariate skew Gaussian data (left) and bivariate exponential data. X-axis show values for increasing componentwise dependency, as indicated by dependency parameters δ_1 and δ. Depicted are means and standard deviations of 25 computations with sample size n=1000 each. Approximations of the true values were derived by numerical integration via nquad in Python. Right: Pearson correlation between MI estimates and analytical value (Gaussian)/numerical estimates (others).

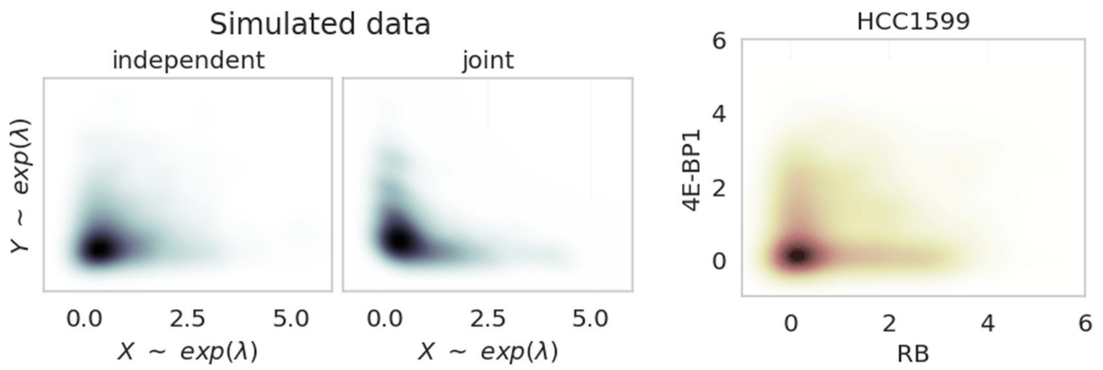


Figure 2. Density plots of simulated exponentially distributed data (left) and 2D phosphoprotein expression values (right) of breast cancer cell line HCC1599. The phosphoprotein expressions selected are the eukaryotic translation initiation factor 4E-binding protein 1 (4E-BP1) and the retinoblastoma protein (RB). Each plot is based on 1000 data points. In the simulated, dependent scenario, Gumbel's method was used as described in Section 3.3.

Table 2. Phosphoproteins and their potential role in the apoptotic process.[a]

Kinase	Effect on apoptosis	Selective reference
ERK1/2: Extracellular signal-regulated kinase 1/2	p53 Ser15 kinase	Yue and López (2020)
JNK: C-Jun N-terminal Kinase	p53 Thr81 kinase	Lee and Gu (2010)
p38: Mitogen-activated protein kinase p38	p53 Ser33 (and Ser46) kinase	Yogosawa and Yoshida (2018)
AMPK: AMP-activated protein kinase	p53 Ser46 kinase	Green *et al.* (2014)
STAT1: Signal transducer and activator of transcription 1	Cross-talk with p53	Zhang and Liu (2017)
STAT3: Signal transducer and activator of transcription 3	Interaction with Bcl-2	Grivennikov and Karin (2010)
Smad2/3: Mothers against decapentaplegic homolog 2 and 3	TGF-β-induced apoptosis (via Bim)	Ramesh *et al.* (2009)
GSK3β: Glycogen synthase kinase 3 beta	Promotion of DNA repair	Lin *et al.* (2020)
NFκB: Nuclear factor kappa B	Interaction with Bcl-2	Pires *et al.* (2018)
RB: Retinoblastoma protein	Interaction with Bcl-2	Polager and Ginsberg (2009)

[a] JNK as well as p38 have been suggested to play a role in p53 stabilization via p53 phosphorylation at Ser15 and Ser20 (Wu 2004). Other effectors for apoptosis used for analysis in Section 4 are the STAT family member protein STAT5 (Halim *et al.* 2020) and further the ratios STAT1/STAT3 as suggested by Avalle *et al.* (2012) and STAT1/STAT5. Further included is the Sarcoma proto-oncogene (Src).

substrate expressions relevant for cell fate decisions should be available. In their 2021 paper, Tognetti *et al.* (2021) claim to have generated the largest multiplex single-cell signaling dataset to date. It consists of 67 human breast cell lines, among which 62 are cancerous. 37 Phosphoproteomic marker expression define the dimensionality of the data, among which are prominent signaling pathway members as well as markers for cell fate decisions such as apoptosis and cell cycle progression. Because of its sheer magnitude, we chose this dataset for our analytic proposes. This is the technical reason. A further reason for our choice is that despite current advances in precision medicine, some cases of breast cancer are still not responsive to treatment.

The dataset is a phosphoproteomic perturbation study. Among the cell lines are all relevant breast cancer subtypes, which are luminal/hormone receptor positive (HR+), HER2/ERBB2 positive (ERBB2+) and basal/triple negative breast cancer (TNBC). While the subtype classification into luminal/basal refers to the breast cell type that gave rise to the malignancy, the classification into HR+/TNBC refers to hormonal signaling. However, in most of the cases both classification schemes are interchangeable. HR+/TNBC cell lines are further classified into an A or B type, according to the PAM50 (Prediction Analysis of Microarray 50) gene set classification, or rather the gene cluster classification as proposed by Neve *et al.* (2006). Five noncancerous breast cell lines serve as control for nonpathological signaling.

As customary for signal pathway analysis, cells were stimulated with epidermal growth factor (EGF) after a period of cell growth (48–72 h) and a night of starvation. Data provided is a time series starting from zero minutes up to one hour after EGF stimulation. While EGF is a prominent growth factor in breast cancer, it has been shown before, that a subset of breast cancer cells also respond to EGF with cell cycle arrest and apoptosis (Ali *et al.* 2018). With help of the time series provided, data allows a sophisticated analysis of the MAPK signal progression, as demonstrated in Tognetti *et al.* (2021), with respect to kinase expression levels. It has been shown in Uda *et al.* (2013) that information propagation seems to be more robust than phosphoprotein expression levels. Early applications of BannMI on the time series data share this observation (data not shown).

Among the signaling phosphoproteins, several kinases have been identified/suggested as influential for intrinsic apoptosis. In the following analysis we refer to them as "effectors for apoptosis" (EfA)s; see Table 2 for their further characterization. Throughout the paper, EfA expressions refer to their phosphorylated state. For more information about the respective phosphorylation site measured, see Tognetti *et al.* (2021). Furthermore, the dataset contains expression of cleaved Caspase-3, a molecule that is cleaved in the ongoing process of apoptosis, which we will use as marker for apoptotic cells.

We proceed as follows: First, we compare potential 1-to-1 signaling for apoptosis with the respective, potential 4-to-1 signaling in cancer cells and the control (Section 4.1). In Section 4.2, we further zoom into the data and investigate the potential signal per cell phenotype (apoptotic, proliferating or resting cells). Throughout this study, we focus on 4-to-1 signaling results that could not have been identified with a 1-to-1 analysis. In the Discussion, we provide a biological interpretation of our findings.

4.1 MI and channel capacity in programmed cell death

Since the dataset provides only one dose of EGF stimulation and discrete time points, we approximated CC by choosing the time point $t = 40$, as it maximized MI for most cell line settings. We first computed 1D CC toward cleaved Caspase-3 for the 15 EfAs as described in Table 2 (see Fig. 3). Significant CC differences (one-sided t-test: Welch, all cancer lines versus control) are marked with stars. To systematically unravel the dependency between EfAs and the apoptotic marker, we next applied the same approach of channel capacity to all $\binom{15}{4} = 1365$ 4D combinations of EfAs [see Fig. 4 (left)]. We observe that the tightly orchestrated structure, in which intrinsic apoptosis unfolds in healthy cells, is absent in cancer. Significant differences in channel capacity for control compared to cancer hold for all combinations (maximum p-value: 0.02).

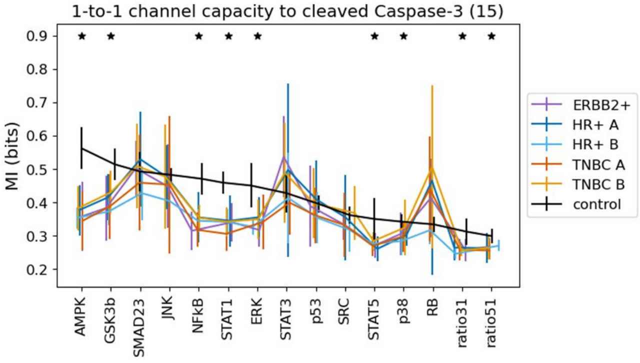


Figure 3. CC estimates between each phos. EfA (*x*-axis) and cleaved Caspase-3. Per cell line (downsampled, n=1000 each) a CC mean was derived from 15 computations. Then, results were grouped into control or cancer subtype. Standard deviations show variation among the groups. Stars indicate significant differences between cancer and control.

To extract the impact of a single EfA within the 4D setting, we ranked all combinations with respect to CC, then selected the 25 top performing quadruples and count occurrences of each EfA [see Fig. 4 (right)]. It is of interest that the top reoccurring EfAs of that ranking procedure are not identical with the EfAs of the highest 1D CC scores. In particular, glycogen synthase kinase-3 β (GSK3β), which is associated with the DNA damage response (DRR) of a cell, is ranked second in the 1D scenario. However, the kinase is ranked only eighths in the 4D scenario with only 4 out of 25 occurrences. This might indicate that its signal is not cooperative in the considered framework. In contrast, in our analysis AMPK, Smad2/3 as well as JNK are tightly linked within the signal of apoptosis and therefore might act with "joint forces". We repeated the analysis for 3D and 5D EfA combinations and received stable results with respect to significant control-cancer differences and EfA counts.

We further investigated the combinatorial effect of EfAs on apoptosis by training 4D random forests. In doing so, we found EfA combinations that are sufficient but not necessary to predict apoptosis in the control cells, see the Supplementary Material.

4.2 MI key players for apoptosis in apoptotic, proliferating, and resting cells

So far, we used MI to identify an overall significantly lowered potential apoptotic signal in cancer cells when compared to noncancerous ones. However, apoptotic, as well as proliferating cells form a minority in the control cell data (5 cell lines). To zoom into the structural dependencies of those phenotypes, we subsampled this control data for phenotypes by thresholds: cells with 5-Iodo-2'-deoxyuridine (IdU) expression (a marker for S-phase in the cell cycle) above 3.5 are selected as proliferating, cells with cleaved Caspase-3 expression above 4.5 are selected as apoptotic, and cells that do not meet any of both criteria are defined as resting [see Fig. 5 (top left)]. We added a mixed sample of the three phentotypes and refer to it as "all".

Analogue to Section 4.1, we computed 4D CC with respect to cleaved Caspase-3 and ranked EfA occurrences of the 25 highest CC-scores. Figure 5 (center) displays the computational results. Here, we want to raise attention to the ranking similarities of the proliferating and the resting cell phenotype:

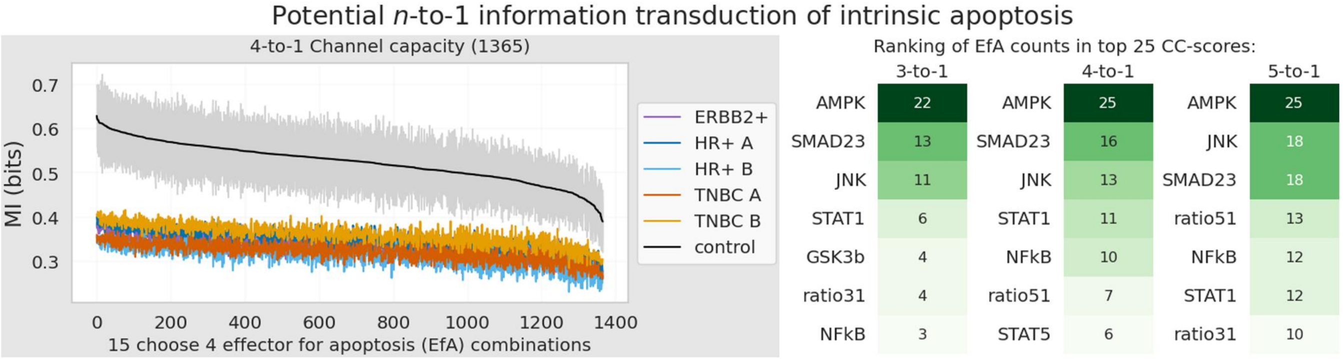
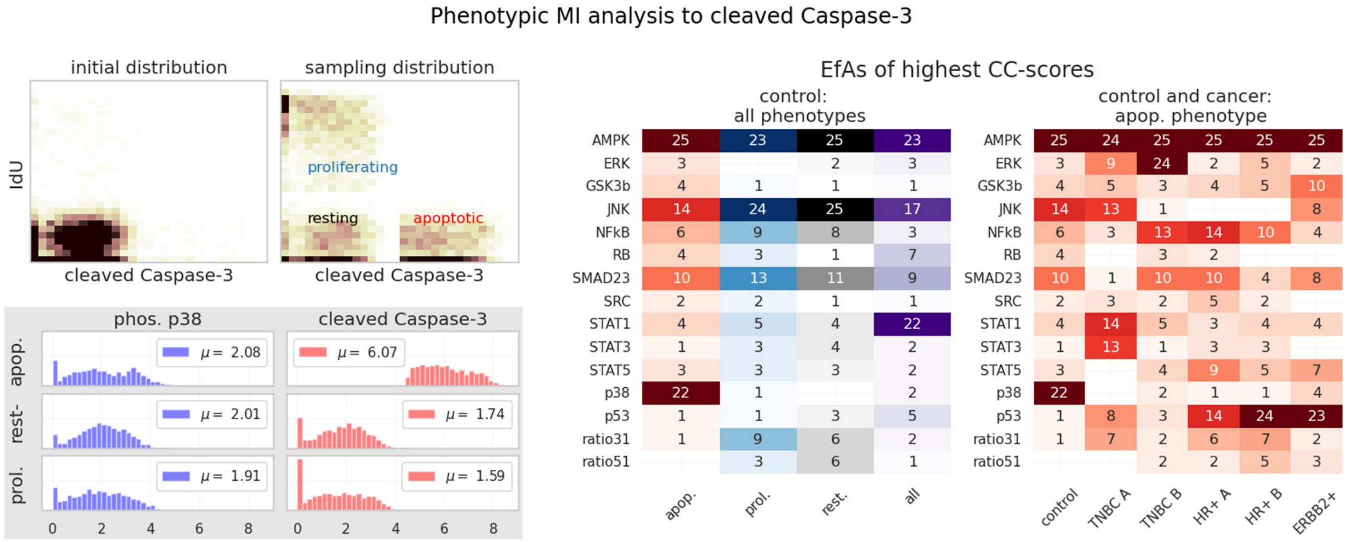


Figure 4. Left: mean CC (15 computations per cell line) for all 4D EfA combinations with respect to the apoptotic marker cleaved Caspase-3. Results were grouped into control and cancer subtype, then sorted with respect to decreasing CC value of the control cell lines. Shaded area marks standard deviations for the control cell line results. Right: EfA-quadruples of the 25 highest 4D CC scores were selected. Next, occurrences of the single EfAs in the 25 quadruples were ranked according to frequency (4-to-1). Analysis was repeated for a 3D and 5D signal.



	apop.	prol.	rest.	all
AMPK	25	23	25	23
ERK	3		2	3
GSK3b	4	1	1	1
JNK	14	24	25	17
NFkB	6	9	8	3
RB	4	3	1	7
SMAD23	10	13	11	9
SRC	2	2	1	1
STAT1	4	5	4	22
STAT3	1	3	4	2
STAT5	3	3	3	2
p38	22	1		2
p53	1	1	3	5
ratio31	1	9	6	2
ratio51		3	6	1

	control	TNBC A	TNBC B	HR+ A	HR+ B	ERBB2+
AMPK	25	24	25	25	25	25
ERK	3	9	24	2	5	2
GSK3b	4	5	3	4	5	10
JNK	14	13	1			8
NFkB	6	3	13	14	10	4
RB	4		3	2		
SMAD23	10	1	10	10	4	8
SRC	2	3	2	5	2	
STAT1	4	14	5	3	4	4
STAT3	1	13	1	3	3	
STAT5	3		4	9	5	7
p38	22		2	1	1	4
p53	1	8	3	14	24	23
ratio31	1	7	2	6	7	2
ratio51			2	2	5	3

Figure 5. Image top left: Original phenotypic distribution in control cell lines with respect to both threshold markers IdU and cleaved Caspase-3; the bordering image shows the phenotypic distribution after the downsampling procedure. Center: EfA occurrence rankings with respect to CC per phenotype for the control cell lines. Right: EfA occurrence rankings for the control cell lines and the five breast cancer subtypes with respect to the apoptotic phenotype. CC values applied for the rankings were means of 15 computations with BannMI. Sample size for all phenotype experiments was n=1000. Bottom left: the phosphoprotein expression of p38 and cleaved Caspase-3 for all phenotypes (control). While cleaved Caspase-3 expression is strongly amplified in apoptotic cells (3-fold increase), p38 expression remains almost unchanged, as can be seen in the mean expression values μ.

as in the previous subsection, a robust dependency of the identified potential apoptotic keyplayers AMPK, JNK and partially the Smad2/3 proteins Smad2 and Smad3 can be observed. Interestingly, this pattern reoccurs in the apoptotic cells, but in addition a robust dependency on MAPK pathway member p38 is observable.

We underline that this role of p38 could not have been found via a method based on differential mean value expression, as phosphorylated p38 is almost identically expressed in all cellular phenotypes, as can be seen in Fig. 5 (bottom left). Further, the Pearson correlation between phos. p38 and cleaved Caspase-3 is lowest in apoptotic cells (0.26), followed by 0.39 (proliferating) and 0.42 (resting).

So far, we restricted our phenotypic analysis to control cells to firstly understand "healthy" dependencies, as they might pave the way to understand causative apoptotic signaling. Next, we investigated the potential signaling roles of our EfAs in cancerous, apoptotic cells. We find that the consistent joint signal of phosphorylated AMPK, JNK, Smad2/3 with the switching role of p38 is absent in cancer cells [see Fig. 5 (right)]. Instead, the heterogeneity of cancer is demonstrated once again. Furthermore, these results indicate the clinical challenge to reconstruct an apoptotic signal in cancer cells. We hope that a joint signal analysis can help in this process of reconstruction.

Finally, we used BannMI as KLD estimator to investigate multivariate differential expression (DE) of the phosphorylated EfAs. We ranked EfA occurrences of 4D DE between the phenotypes in the control cells, as well as between control and cancer cells. See the results in the Supplementary Material.

5 Discussion

5.1 Biological interpretation

P53 is the central figure of intrinsic apoptosis. The transcription factor harbors a large number of phosphorylation sites that alter its functionality. According to Lavin and Gueven (2006), there are 11 serine activation sites (Ser6, Ser9, Ser15, Ser20, Ser33, Ser37, Ser46, Ser366, Ser376, Ser378, Ser392) and four further threonine activation sites (Thr18, Thr81, Thr377 and Thr387). Phosphorylation at Ser15, which was measured in the dataset by Tognetti *et al.* (2021), is mediated via the ataxia telangiectasia mutated (ATM) kinase, a central kinase regulating DNA damage response, and is essential for the transcriptional activation of p53 as follows. P53 is a short-lived protein that is degraded via ubiquitination via its main inhibitor, the mouse double minute 2 (Mdm2) protein in unstressed cells, such that only low expression levels occur in healthy cells. Phosphorylation of Ser15 (together with Ser20 and Thr81; Buschmann *et al.*, 2001) interrupts p53-Mdm2 interaction such that p53 accumulates and translocates into the nucleus. Therefore, phosphorylation of p53 at Ser15 is an indication for p53 activation and potential apoptosis induction in cancer cells. Other important p53 kinases are AMPK and JNK. Both phosphoproteins, together with the main TGF-β effectors Smad2/3 are the top occurring EfAs in the 4D CC analysis of Sections 4.1 and 4.2. In the latter, also phosphorylated p38 is among the top occurring EfAs, but only in the apoptotic phenotype. Therefore, our results indicate firstly that p53 phosphorylation is cooperative via AMPK, JNK and p38.

Secondly, the strong cooperation of Smad2/3 in the joint signal might provide further evidence for a control mechanism of apoptosis which is guarded by the signal of several pathways. Smad2/3 are the main kinases of the TGF$-\beta$ pathway. The pathway is, among others, linked to apoptosis. Cordenonsi *et al.* (2007) show that phosphorylated p53 (at Ser6 and Ser9) physically interact with Smad2/3 and thus jointly regulates the transcription of several TGF-β target genes. Among those substrates is the Bcl-2-interacting mediator of cell death protein (Bim), a pro-apoptotic member of the Bcl-2 family. Its major isoform is BimEL (where the "EL" stands for "extra long"). Interestingly, it has been shown that BimEL ubiquitination and degradation is controlled by Erk1/2 phosphorylation (Ramesh *et al.* 2009). Those are only two examples of cooperative signaling between the MAPK and the TGF-β pathway. Recently, Wang *et al.* (2022) discuss a dual control model for intrinsic apoptosis and provide an alternative model based on *in vitro* experiments with cell lines of three cancer types. In those models, apoptosis "happens" in the onset of Bcl-2 protein interaction chains with a special focus on Bcl-2 homology 3 (BH3)-only proteins, such as Bim. The authors show with help of p53 deficient cells that Bim expression is independent of p53 in their experiments. However, apoptosis is mainly observed in p53 wild type cell lines.

5.2 Future applications of BannMI

We propose BannMI, an important tool to identify cooperative signaling and to disentangle individual roles in *n*-to-1 signaling based on NN MI estimation. Results of our case study indicate that some phosphoproteomic dependencies are only observable if more than one protein expressions/or their phosphorylations, are considered at once. This new perspective allows quantification of the complex interplay of several signaling pathways in cell fate decisions along the example of apoptosis in breast cell lines/breast cancer cell lines. It is of interest for future research if our findings can be extended to other cellular types besides the human breast.

The generality of BannMI allows to zoom into any scenario of *n*-to-1 communication with moderate dimensionality (benchmarking covered dimensionality up to $d = 10$), which could be any kind of omics data.

Acknowledgements

The authors would like to thank Ulf Leser for his constructive feedback and fruitful discussions. The employed datasets were obtained as part of the Single Cell Signaling in Breast Cancer Challenge through Synapse ID `syn20564743` and contributed by Attila Gabor and Marco Tognetti.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by the German research foundation (DFG) [RTG 2424 to B.S.]. N.K. gratefully acknowledges support by the German research foundation (DFG) through the Emmy Noether [KL 3037/1–1].

Data availability

Data are available on Synapse and Mendeley Data. For the latter, check https://data.mendeley.com/datasets/gvh2vtg86r/1.

References

Ali R, Brown W, Purdy SC *et al.* Biased signaling downstream of epidermal growth factor receptor regulates proliferative versus apoptotic response to ligand. *Cell Death Dis* 2018;**9**:976.

Avalle L, Pensa S, Regis G *et al.* Stat1 and stat3 in tumorigenesis: a matter of balance. *JAKSTAT* 2012;**1**:65–72.

Azzalini A, Valle AD. The multivariate skew-normal distribution. *Biometrika* 1996;**83**:715–26.

Buschmann T, Potapova O, Bar-Shira A *et al.* Jun NH2-terminal kinase phosphorylation of p53 on thr-81 is important for p53 stabilization and transcriptional activities in response to stress. *Mol Cell Biol* 2001;**21**:2743–54.

Cordenonsi M, Montagner M, Adorno M *et al.* Integration of tgf-ß and ras/mapk signaling through p53 phosphorylation. *Science* 2007; **315**:840–3.

Dorel M, Klinger B, Gross T *et al.* Modelling signalling networks from perturbation data. *Bioinformatics* 2018;**34**:4079–86.

Erman B. Mutual information analysis of mutation, nonlinearity, and triple interactions in proteins. *Proteins Struct Funct Bioinform* 2023;**91**:121–33.

Gelman A, Carlin JB, Stern HS *et al.* *Bayesian Data Analysis*, 3rd edn. New York: CRC Press, 2015.

Green DR, Galluzzi L, Kroemer G *et al.* Metabolic control of cell death. *Science* 2014;**345**:1250256.

Grivennikov SI, Karin M. Dangerous liaisons: STAT3 and NF-κb collaboration and crosstalk in cancer. *Cytokine Growth Factor Rev* 2010;**21**:11–9.

Gumbel EJ. Bivariate exponential distributions. *J Am Stat Assoc* 1960; **55**:698–707.

Halim CE, Deng S, Ong MS *et al.* Involvement of STAT5 in oncogenesis. *Biomedicines* 2020;**8**:316.

Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 2014; **15**:1593–623.

Kamimoto K, Stringa B, Hoffmann CM *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 2023; **614**:742–51.

Karolak A, Branciamore S, McCune JS *et al.* Concepts and applications of information theory to immuno-oncology. *Trends Cancer* 2021; **7**:335–46.

Kozachenko L, Leonenko N. Sample estimate of the entropy of a random vector. *Problems Inf Transmission* 1987;**23**:95–101.

Kraskov A, Stögbauer H, Grassberger P *et al.* Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004;**69**:066138.

Lavin MF, Gueven N. The complexity of p53 stabilization and activation. *Cell Death Differ* 2006;**13**:941–50.

Lee J, Gu W. The multiple levels of regulation by p53 ubiquitination. *Cell Death Differ* 2010;**17**:86–92.

Liebl MC, Hofmann TG. Cell fate regulation upon dna damage: p53 serine 46 kinases pave the cell death road. *Bioessays* 2019; **41**:e1900127.

Lin J, Song T, Li C *et al.* Gsk-3β in dna repair, apoptosis, and resistance of chemotherapy, radiotherapy of cancer. *Biochim Biophys Acta (BBA) Mol Cell Res* 2020;**1867**:118659.

Neve RM, Chin K, Fridlyand J *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 2006;**10**:515–27.

Noshad M, Moon KR, Sekeh SY *et al.* 2017. Direct estimation of information divergence using nearest neighbor ratios. In: *2017 IEEE International Symposium on Information Theory (ISIT 2017).* Aachen, Germany: Institute of Electrical and Electronics Engineers (IEEE), 903–7.

Pires BRB, Silva RCMC, Ferreira GM *et al.* NF-kappab: two sides of the same coin. *Genes (Basel)* 2018;**9**:24.

Polager S, Ginsberg D. p53 and E2f: partners in life and death. *Nat Rev Cancer* 2009;**9**:738–48.

Ramesh S, Wildey GM, Howe PH *et al.* Transforming growth factor β (tgfβ)-induced apoptosis: the rise and fall of bim. *Cell Cycle* 2009; **8**:11–7.

Tognetti M, Gabor A, Yang M *et al.* Deciphering the signal network of breast cancer improves drug sensitivity prediction. *Cell Syst* 2021; **12**:401–18.e12.

Uda S. Application of information theory in systems biology. *Biophys Rev* 2020;**12**:377–84.

Uda S, Saito TH, Kudo T *et al*. Robustness and compensation of information transmission of signaling pathways. *Science* 2013;**341**:558–61.

Wada T, Hironaka K-I, Kuroda S *et al*. Cell-to-cell variability serves as information not noise. *Curr Opin Syst Biol* 2021;**27**:100339.

Wang Q, Kulkarni SR, Verdu S *et al*. Divergence estimation for multidimensional densities via *k*-Nearest-Neighbor distances. *IEEE Trans Inform Theory* 2009;**55**:2392–405.

Wang Y-C, Wang L-T, Hung TI *et al*. Severe cellular stress drives apoptosis through a dual control mechanism independently of p53. *Cell Death Discov* 2022;**8**:282.

Wu GS. The functional interactions between the mapk and p53 signaling pathways. *Cancer Biol Ther* 2004;**3**:156–61.

Yogosawa S, Yoshida K. Tumor suppressive role for kinases phosphorylating p53 in dna damage-induced apoptosis. *Cancer Sci* 2018;**109**:3376–82.

Yue J, López JM. Understanding mapk signaling pathways in apoptosis. *Int J Mol Sci* 2020;**21**:2346.

Zhang Y, Liu Z. STAT1 in cancer: friend or foe? *Discov Med* 2017; **24**:19–29.

Zielińska K, Katanaev V. Information theory: new look at oncogenic signaling pathways. *Trends Cell Biol* 2019;**29**:862–75.