Information-theoretic analysis of multivariate single - cell signaling responses using SLEMI

Supplemental Information

Tomasz Jetka[1], Tomasz Winarski[1], Karol Nienałtowski[1], Sławomir Błoński, & Michał Komorowski[1]

[1]Institute of Fundamental Technological Research,
Polish Academy of Sciences,
Warsaw, Poland

Contents

1 Background methods

Within information theory, a signaling system is considered as a probability distribution $P(Y|X = x)$ that for a given level of input, x, elicits an output, Y. In a basic example, the input is the concentration of a ligand that activates a receptor, and the output is the activity of a signaling effector, which might be the nuclear concentration of an activated transcription factor. In more complex examples, the input x can be a vector representing concentrations of several ligands, whereas the output Y a vector with entries representing temporally resolved concentrations of several activated transcription factors. Regardless of the specific details, the output, Y, carries information about the values of the input, x. Different values of the input, x, can occur with different frequencies, which is represented by the random variable X following a probability distribution $P(X)$.
Randomness of the output at a given input prevents the system from resolving a precise value of the input. However, uncertainty regarding input values cannot be higher than that of the overall variability of the input, as represented by the input distribution $P(X)$. Within information theory, uncertainty of the input distribution variable can be quantified as entropy

$$H(X) = -\int_{\mathcal{X}} \log_2(P(x))P(x)dx, \tag{1}$$

where $\mathcal{X}$ is the space of possible values of the signal, X. Observation of the output has a potential to reduce uncertainty regarding input value. Via the Bayes formula, plausible inputs that generated a specific output value, y, are represented as the probability distribution

$$P(X|Y = y) = \frac{P(Y = y|X)P(X)}{P(Y = y)}. \tag{2}$$

Uncertainty regarding input value can be then quantified by the entropy of the distribution $P(X|Y = y)$

$$H(X|Y = y) = -\int_{\mathcal{X}} \log_2(P(x|Y = y))P(x|Y = y)dx. \tag{3}$$

As the output is random, averaging $H(X|Y = y)$ over all possible outputs quantifies the average uncertainty regarding the input, given the output, $H(X|Y)$,

$$H(X|Y) = \int_{\mathcal{Y}} H(X|Y = y)P(y)dy, \tag{4}$$

where $\mathcal{Y}$ is the space of possible values of the output, Y. The difference between $H(X)$ and $H(X|Y)$ measures the average reduction in uncertainty regarding the input resulting from observing an output and is referred to as mutual information, $MI(X, Y)$, between the input and the output

$$MI(X, Y) = H(X) - H(X|Y). \tag{5}$$

The mutual information depends on the distribution of inputs, $P(X)$. For instance, consider two possible sets of input values. One set of inputs generates similar and/or irreproducible outputs (low sensitivity and/or high noise), while the other generates distinct and reproducible outputs. If a pathway encounters signals from the first set more frequently than from the second one, its information flow will be lower. Therefore, to quantify how much information can be transmitted through a system the maximal mutual information, with respect to all input distributions, termed information capacity, C^*, is considered

$$C^* = \max_{P(X)} MI(X, Y). \tag{6}$$

The distribution for which the maximum of mutual information is achieved is called the optimal input distribution and denoted as $P^*(X)$. The information capacity, C^* is expressed in bits, and 2^{C^*} can be interpreted, within the Shannon's coding theorem [5, 18, 6], as the number of input values that the system can effectively resolve based on the information contained in the output [6]. For instance, if $C^* = 2$, there exist four concentrations that can be distinguished with, on average, negligible error.

1.1 Experiments to quantify information capacity of signalling pathways

In a typical experiment aimed at quantification of information capacity of a signalling pathway, clonal cells are exposed to a finite number of stimulus levels $x_1, \ldots, x_m$. Then, responses to each stimulus value, x_i, are quantified in a large number of individual cells and represented as vectors y^i_j, where j varies from 1 to the number of observed cells, n_i. If the input, x, is one-dimensional, the stimuli levels are usually ordered, i.e. $x_1 \leq \ldots \leq x_m$, and distributed to cover a range from 0 to saturation. If the input x is more than one dimensional, the stimuli values $x_1, \ldots, x_m$ constitute a representative sample of physiologically relevant range. Typically however, x_i is a one-dimensional, e.g., concentration of a cytokine, and y^i_j are vectors that describe temporally resolved responses of individual cells, e.g. nuclear levels of activated transcription factors. Responses are assumed to follow a probability distribution $P(Y|X = x)$ and, hence, the collected dataset can be conceptually represented as

$$P(Y|X = x_i) \quad \sim \quad \left(y^i_1,\ y^i_2,\ \ldots,\ y^i_{n_i}\right), \tag{7}$$

where $y^i_j \in \mathbb{R}^d$ is the response of j-th cell stimulated by the stimuli level x_i.
In the typical experimental setting described above the input takes a finite number of values, m, i.e. is discreet, as opposed to continuous input values of Eq. 1 - 6. Hence, a discreet input distribution is considered, i.e., $P(x_1), \ldots, P(x_m)$. Using the discreet version of entropies (1) and (3), in which integration with respect to x is replaced by summation, the mutual information can be written as

$$MI(X,Y) = \sum_{i=1}^{m} P(x_i) \int_{\mathbb{R}^d} P(y|X = x_i) \log_2 \frac{P(y|X = x_i)}{P(y)} dy, \tag{8}$$

and is maximized over the discreet probability distributions to obtain capacity

$$C^* = \max_{P(x_1),\ldots,P(x_m)} MI(X,Y). \tag{9}$$

1.2 Existing methods to estimate mutual information and information capacity for single - cell signaling data

The conventional solution to the problem of estimating the channel capacity, Eq. 9, is provided by the Blahut-Arimoto (BA) algorithm [3, 2] and its extensions [24]. In its original version BA algorithm can be used to calculate the information capacity for a scenario, in which both the output, Y, and the input, x, take discrete values, i.e. a finite set of values. It can also be applied to systems with continuous outputs, through output discretization, i.e., conversion into a variable that takes finite set values. A systematic way to convert the output into a discreet variable within the BA framework was proposed in [4]. The discretization of the output is an efficient solution if Y is one-dimensional. Therefore, BA algorithm cannot efficiently estimate the capacity in scenarios with continuous multidimensional output, Y.

A method to calculate the capacity, Eq. 9, for signaling systems with continuous output of several dimensions have been recently proposed in [17], similarly to the work of [23] and earlier theoretical approaches of [13]. The method involves approximation of the output probability densities $P(Y|x_i)$, for $i = 1, ..., m$ with the k-nearest neighbour approach (KNN) and numerical optimisation. Broadly speaking, the KNN method utilizes the distance to the kth nearest neighbor to approximate the probability density. The method is however problematic as the estimates depend on the selection of k and require a large sample size to provide accurate results. Moreover, it involves numerical gradient optimization that is susceptible to finding local maxima, especially if the number of input values is large.

A more technical argument against KNN based estimation of the information capacity can also be made. The method proposed in [17] uses KNN directly without additional bias correction (compare with [13]) and for that reason is prone to inaccurate estimation of the mutual information for small sample size and high dimensionality of data. As a result, accurate estimation of information capacity requires accurate

estimation of density. KNN density estimators have been shown to work well for large data size, i.e., to be consistent, for k increasing with the data size. Precisely, it is required that $k = k_N \to \infty$ and $k_N/N \to 0$ as $N \to \infty$ [14], where N is the number of data points used in estimation. The rate of the estimators' convergence, i.e., the number of data points needed to obtain the correct value of the estimated density, depends on the dimension d, of the random variable, Y. Therefore, for larger d, larger N and k may be required to get accurate estimates of the density $P(Y|x_i)$ and consequently of the capacity, C^*. Moreover, the optimal k does not follow a simple scaling rule with d and N. In [1], authors suggests to choose k, that is proportional to $n^{4/(d+4)}$, however for lower dimensions such conditions lead to highly biased entropy estimators. A more detailed discussion of those difficulties in a specific case of information theoretic measures can be found in [8, 16, 20]. In summary, the above arguments suggest that the selection of k, which leads to an accurate estimation of the information capacity, may be problematic.

2 Algorithm

Below, we describe in detail the algorithm to estimate the information capacity introduced in Section 2.1 of the main paper. Thereafter, technical lemmas used to construct the algorithm are presented.

2.1 Algorithm description

The practical difficulty in calculating C^*, Eq. 9, results largely from the following two problems. Firstly, the calculation involves maximization of a nonlinear function (mutual information) over the input probability distribution, which, in general, is computationally intense and prone to finding local minima [7]. Secondly, methods for accurate estimation of multivariate density functions $P(y|X = x_i)$ and $P(y)$, required to evaluate mutual information Eq. 8, are missing [19]. In the introduced algorithm, we propose to overcome the above two issues by combining the so-called, alternate optimisation method with statistical learning tools. Below, we first describe the alternate maximization (AM) technique and then show how it can be merged with a statistical learning approaches to ensure efficient computation of the information capacity, especially for systems with high dimensional output, Y.
The AM method to calculate the information capacity, C^*, has been first employed within the Blahut and Arimoto algorithm. It is based on an alternative formulation of the information capacity problem. Precisely, Blahut and Arimoto have shown (see [3], Theorem 3) that the capacity C^*, Eq. 9, can be written as

$$C^* = \max_{P(X)} MI(X,Y) = \max_{P(X)} \max_{Q(X,Y)} J(P(X), Q(X,Y)), \tag{10}$$

where $Q(X,Y)$ is a real-valued function that for a given $Y = y$ is a discrete probability distribution with respect to X, whereas the auxiliary function J is defined as

$$J(P(X), Q(X,Y)) = \sum_{i=1}^{m} P(x_i) \int_{\mathbb{R}^d} P(y|X = x_i) \log_2 \frac{Q(x_i, y)}{P(x_i)} dy. \tag{11}$$

The auxiliary function J is a reminiscence of the mutual information. Simplistically, it was constructed from the mutual information by replacing the distribution $P(X|Y)$ with $Q(X,Y)$ to ensure efficient solution of the maximization problem in Eq. 10. Precisely, single maximisation problem, with respect to $P(X)$, Eq. 9, is replaced with the double maximisation, with respect to $P(X)$ and $Q(X,Y)$, which under certain conditions (see Lemma 1 in the subsequent section), can be solved using an iterative algorithm. Double maximisation has the advantage that the optimal solutions of the individual optimisation problems

i) $P^*(X; Q(X,Y)) = \arg\max_{P(X)} J(P(X), Q(X,Y))$,

ii) $Q^*(X,Y; P(X)) = \arg\max_{Q(X,Y)} J(P(X), Q(X,Y))$

can be found explicitly. Precisely, using Lagrange multipliers, it can be shown that the optimal value of

$$\max_{P(X)} J(P(X), Q(X,Y)) \tag{12}$$

is achieved by

$$P^*(x_i; Q(X,Y)) = \frac{\exp(D_i(Q))}{\sum_{l=1}^{m} \exp(D_l(Q))}, \tag{13}$$

where

$$D_i(Q) = \int_{\mathbb{R}^d} P(y|X = x_i) \log Q(x_i, y) dy. \tag{14}$$

In Lemma 2 in the subsequent section, we provide the precise derivation. In contrast to similar derivation in [3], we account for continuous output Y.

The second of the two individual maximisation problems

$$\max_{Q(X,Y)} J(P(X), Q(X,Y)) \tag{15}$$

has the explicit solution

$$Q^*(x_i, y; P(X)) = \frac{P(x_i)P(y|X=x_i)}{\sum_{l=1}^{m} P(x_l)P(y|X=x_l)}. \tag{16}$$

which is a well established result proved first as Theorem 1 in [3]. It is used in the unchanged form within our approach, as the input distribution in our approach and in BA algorithm is discreet.

The above solutions of the individual maximizations are subsequently combined in an iterative procedure that delivers the solution of the joint maximization. Precisely, in an initial step, arbitrary $P(X)$ and $Q(X,Y)$, denoted as $P^{(0)}(x_i), Q^{(0)}(x_i, y)$, respectively, are assumed. Thereafter, at each step, indexed by k, new $P(X)$ and $Q(X,Y)$ denoted as $P^{(k)}(x_i), Q^{(k)}(x_i, y)$, are found using the solution of the individual maximization problems based on the solution in the previous step

$$\begin{aligned} P^{(k)}(x_i) &= P^*(x_i; Q^{(k-1)}(x_i, y)) \quad \text{for } k > 0, \\ Q^{(k)}(x_i, y) &= Q^*(x_i, y; P^{(k)}(x_i)) \quad \text{for } k > 0. \end{aligned}$$

As shown in [2] the iterative scheme converges to the solution of the joint maximisation problem, Eq. 10, precisely $P^{(k)}(x_i) \xrightarrow{k\to\infty} P^*(x_i)$ and $J(P^{(k)}(X), Q^{(k)}(X,Y)) \xrightarrow{k\to\infty} C^*$.

The above optimisation scheme is efficient as it is based on the explicit solution of the individual optimisation problems. Therefore, it overcomes one of the two difficulties that prohibit efficient computations of the information capacity. Unfortunately, it requires the evaluation of the output density, $P(y|X=x_i)$, to perform integration with respect to y in Eqs. 11 and 14 as well as to calculate $Q^{(k)}(x_i, y)$ at each step of the iteration. Therefore, given difficulties in estimating $P(y|X=x_i)$, its use in estimation of the capacity for systems with multidimensional output, Y, is not straightforward and, to our knowledge, has not been explored. Below, we show that explicit knowledge of $P(y|X=x_i)$ is not necessary neither to evaluate $Q^{(k)}(x_i, y)$ nor to perform integration in Eqs. 11 and 14. Therefore, the above efficient optimisation can also be utilized for systems with multidimensional output, Y. Primarily, we show that integration with respect to y, in Eqs. 11 and 14 can be performed without knowledge of $P(y|X=x_i)$. Thereafter, we demonstrate how $Q^{(k)}(x_i, y)$ can be computed.

Eqs. 11 and 14 involve the integrals

$$\int_{\mathbb{R}^d} P(y|X=x_i) \log_2 \frac{Q(x_i, y)}{P(x_i)} dy$$

and

$$\int_{\mathbb{R}^d} P(y|X=x_i) \log_2 Q(x_i, y) dy,$$

respectively. Those integrations denote expectations with respect to the distribution $P(Y|X=x_i)$, i.e. $\mathbb{E}_{P(Y|X=x_i)}(\cdot)$. The law of large numbers implies that expectations can be approximated solely based on a sample from the distribution $P(Y|X=x_i)$. Indeed, if only the number of observations in experimental data is large enough, the average of the sample from data approximates the expectation, leading to

$$\int_{\mathbb{R}^d} P(y|X=x_i) \log_2 \frac{Q(x_i, y)}{P(x_i)} dy = \mathbb{E}_{P(Y|X=x_i)} \left(\log_2 \frac{Q(x_i, Y)}{P(x_i)} \right) \approx \frac{1}{n_i} \sum_{j=1}^{n_i} \log_2 \frac{Q(x_i, y_j^i)}{P(x_i)} \tag{17}$$

$$\int_{\mathbb{R}^d} P(y|X=x_i) \log_2 Q(x_i, y) dy = \mathbb{E}_{P(Y|X=x_i)} \log_2 Q(x_i, Y) \approx \frac{1}{n_i} \sum_{j=1}^{n_i} \log_2 Q(x_i, y_j^i). \tag{18}$$

The above shows, that the integration of Eq. 11 and 14, indeed does not require explicit knowledge of $P(y|X = x_i)$. Below, we show that computing $Q^*(x_i, y; P(X))$ does not require evaluation of $P(y|X = x_i)$ as well. Specifically, using the Bayes rule we can express $Q^*(x_i, y; P(X))$, Eq. 19, as

$$Q^*(x_i, y; P(X)) = \frac{P(x_i)P(y|X = x_i)}{\sum_{l=1}^{m} P(x_l)P(y|X = x_l)} = P(x_i|Y = y). \tag{19}$$

The above proves that calculation of $Q^*(x_i, y; P(X))$, does not require $P(y|X = x_i)$ when $P(x_i|Y = y)$ is available. Finding an approximation to $P(x_i|Y = y)$ is a classification problem in statistical learning theory [9]. A multitude of methods have been established to estimate $P(x_i|Y = y)$ from data. Therefore, we propose to approximate $Q^*(x_i, y; P(X))$ in the above formula using a statistical classifier. In the following algorithm, we present a solution using a classifier based on logistic regression. It is computationally efficient and is known to provide reliable estimates in a variety of scenarios [9]. In principle, however, any other Bayesian classifier can be utilized here. Precisely, we propose to use logistic regression model to find Q^* , i.e. to replace Eq. 13 with

$$Q^*(x_i, y; P(X)) = \hat{P}_{lr}(x_i|Y = y; P(X)), \tag{20}$$

where $\hat{P}_{lr}(x_i|Y = y; P(X))$ denotes a logistic regression model for classifying x_i based on observation of y and the prior distribution $P(X)$. Specifically, the logistic regression model of $P(x_i|Y = y)$ is based on the following equations

$$\begin{aligned}
\hat{P}_{lr}(x_1|Y = y; P(X)) &= \frac{1}{1 + \sum_{l=2}^{m} \exp\left(\alpha_l + \beta_l^T y\right)}, \\
&\vdots \\
\hat{P}_{lr}(x_i|Y = y; P(X)) &= \frac{\exp\left(\alpha_i + \beta_i^T y\right)}{1 + \sum_{l=2}^{m} \exp\left(\alpha_l + \beta_l^T y\right)}, \\
&\vdots \\
\hat{P}_{lr}(x_m|Y = y; P(X)) &= \frac{\exp\left(\alpha_m + \beta_m^T y\right)}{1 + \sum_{l=2}^{m} \exp\left(\alpha_l + \beta_l^T y\right)},
\end{aligned} \tag{21}$$

where $y \in \mathbb{R}^d$ is a vector of measurements, $\alpha_i \in \mathbb{R}$ are intercepts for the i-th class and $\beta_i \in \mathbb{R}^d$ are vectors of parameters for the i-th class. In total, the above procedure requires fitting $(m-1)\cdot(d+1)$ parameters. Practically, Eq. 20 can be obtained from experimental data, which can be done with high efficiency [11].

The logistic regression model, $\hat{P}_{lr}(x_i|y; P(X))$, depends on the input distribution, $P(X)$. The iterative optimisation procedures involves different input distributions, $P^{(k)}(x_i)$, at each step, k. Conveniently, however, even though the input distributions are different at each step, the logistic regression model does not need to be fitted to experimental data at each step of the algorithm. This advantage results from the possibility to update coefficients of logistic regression model in response to a change of the input distribution. In Lemmas 4 and 5, we show how parameters of logistic regression change when $P(X)$ changes. Results of Lemmas 4 and 5 lead to a formula relating the odds ratio of $Q^{(k)}(x_i, y; P^{(k)}(X))$ and $Q^{(k-1)}(x_i, y; P^{(k-1)}(X))$ (for $k > 0$)

$$\frac{Q^{(k)}(x_i, y)}{Q^{(k)}(x_1, y)} \approx \frac{Q^{(k-1)}(x_i, y)}{Q^{(k-1)}(x_1, y)} \cdot \frac{P^{(k-1)}(x_1)}{P^{(k-1)}(x_i)} \cdot \frac{P^{(k)}(x_i)}{P^{(k)}(x_1)}. \tag{22}$$

, which in turn enables efficient computation of $Q^{(k)}(x_i, y)$.
Combining the AM with the above method to calculate $Q^{(k)}(x_i, y)$ leads to efficient algorithm to calculate C^* that is summarised as a pseudo-code in Box 1.

Box 1 Algorithm to calculate channel capacity using statistical learning

1: Set maximum number of iterations `MAXIT` and tolerance level `tol`
2: Set $k = 0$
3: Set initial distribution of $P^{(0)}(X)$: $P^{(0)}(x_i) = \frac{n_i}{\sum_{l=1}^{m} n_l}$
4: Estimate $Q^{(0)}(x_i, y; P^{(0)}(X)) = \hat{P}_{lr}(x_i|Y = y; P^{(0)}(X))$, i.e. by logistic regression model
5: **while** k $\leq$ `MAXIT` AND $|C^{(k+1)} - C^{(k)}| >$`tol` **do**
6: Calculate $D_i(Q^{(k-1)})$ by Monte Carlo integration (see Lemma 3)

$$D_i(Q^{(k-1)}) = \mathbb{E}_{P(Y|X=x_i)}\left(\log Q^{(k-1)}(x_i, y)\right) \approx \frac{1}{n_i}\sum_{j=1}^{n_i} \log Q^{(k-1)}(x_i, y_j^i),$$

7: Optimize $\max_{P(X)} J(P, Q^{(k-1)})$ (see Lemma 2)

$$P^{(k)}(x_i) = P^*(x_i; Q^{(k-1)}(x_i, y)) = \frac{\exp(D_i(Q^{(k-1)}))}{\sum_{l=1}^{m} \exp(D_l(Q^{(k-1)}))}$$

8: Optimize $\max_{Q(X,Y)} J(P^{(k)}, Q)$ by

$$Q^{(k)}(x_i, y) = Q^*(x_i, y; P^{(k)}(x_i)) = \hat{P}_{lr}(x_i|Y = y; P^{(k)}(X))$$

which can be calculated from $Q^{(k-1)}(x_i, y)$ according to Eq. 22 (see Lemmas 4 and 5).
9: Get, $C^{(k)}$, an estimate of C^* (See Lemma 3)

$$C^{(k)} = \sum_{i=1}^{m} P^{(k)}(x_i)\left(D_i(Q^{(k)}) - \log P^{(k)}(x_i)\right)$$

10: $k = k + 1$
11: **end while**
12: **return** $C^* = C^{(k)}$, $P^*(X) = P^{(k)}(X)$.

2.2 Technical Lemmas

Below, we present major technical aspects of the derivation presented in the previous section. Firstly, we give exact conditions needed for the AM procedure to converge (Lemma 1). Secondly, we present the solution of Eq. 13 (Lemma 2) and a method to evaluate Eq. 14 (Lemma 3). Finally, we explain how parameters of the logistic regression equations are updated in the iterative optimisation (Lemmas 4 and 5).
Lemma 1 has been previously established in [25], while Lemma 2 is directly adapted from [3] with minor changes to account for the case of continuous-output channel. Lemmas 3, 4 and 5 are new contributions developed to provide the theoretical basis for the proposed algorithm.

2.2.1 Alternating maximisation (AM) procedure

One of the main problems in calculating channel capacity is performing the optimisation with respect to $P(X)$ (Eq. 9). We propose to use AM for this purpose (Eq. 10). The general case of this procedure is summarised in the following lemma

Lemma 1. *Let $f(u_1, u_2)$ be a real-valued function and consider the optimisation problem*

$$f^* = \max_{u_1 \in \mathcal{U}_1, u_2 \in \mathcal{U}_2} f(u_1, u_2).$$

In addition, denote by $u_2^*(u_1)$ the solution of $\arg\max_{u_2 \in \mathcal{U}_2} f(u_1, u_2)$ and by $u_1^*(u_2)$ the solution of $\arg\max_{u_1 \in \mathcal{U}_1} f(u_1, u_2)$. A sequence $\left(u_1^k, u_2^k\right)$ is constructed as follows:

1. *(u_1^0, u_2^0) is an arbitrary starting point from set $\in \mathcal{U}_1 \times \in \mathcal{U}_2$*
2. *for $k > 0$*

$$\begin{cases} u_1^k &= u_1^*\left(u_2^{k\text{-}1}\right) \quad \text{for } k > 0. \\ u_2^k &= u_2^*\left(u_1^k\right), \quad \text{for } k > 0. \end{cases}$$

Then, the sequence $\left(u_1^k, u_2^k\right)$ converges to the optimal solution, i.e. $f(u_1^k, u_2^k) \xrightarrow{k \to \infty} f^*$, if

- *function f is continuous, concave and bounded from above;*
- *partial derivatives of f are continuous;*
- *sets A_1 and A_2 are convex.*

The proof of the Lemma 1 is given in [25], while the use of AM in the case of mutual information and channel capacity is justified by [2, 3].

Optimising prior probabilities

The application of AM procedure to calculate channel capacity requires optimisation of Eq. 10, in turns, a) with respect to $P(X)$ and b) with respect to $Q(X, Y)$. Below we show the exact derivation of the solution to the former.

Lemma 2. *The optimisation problem*

$$\max_{P(x_1),\ldots,P(x_m)} \quad \sum_{i=1}^{m} \int_{\mathbb{R}^d} P(x_i) P(y|X = x_i) \log_2 \frac{Q(x_i, y)}{P(x_i)} dy, \tag{23}$$

$$w.r.t. \quad \sum_{i=1}^{m} P(x_i) = 1, \tag{24}$$

$$0 \leq P(x_i) \leq 1 \text{ for each } i \tag{25}$$

with $P(y|X = x_i)$ being a probability density function for each x_i and $Q(x_i, y)$ being a probability function for each y is solved by

$$P^*(x_i) = \frac{\exp(D_i)}{\sum_{l=1}^{m} \exp(D_l)},$$

where $D_i = \int_{\mathbb{R}^d} P(y|X = x_i) \log Q(x_i, y) dy$

Proof Let's reformulate the objective function

$$\sum_{i=1}^{m} \int_{\mathbb{R}^d} P(x_i)P(y|X = x_i) \log_2 \frac{Q(x_i, y)}{P(x_i)} dy =$$

$$= \sum_{i=1}^{m} \int_{\mathbb{R}^d} P(x_i)P(y|X = x_i) \log_2 Q(x_i, y) dy - \sum_{i=1}^{m} \int_{\mathbb{R}^d} P(x_i)P(y|X = x_i) \log_2 P(x_i) dy =$$

$$= \sum_{i=1}^{m} P(x_i) \int_{\mathbb{R}^d} P(y|X = x_i) \log_2 Q(x_i, y) dy - \sum_{i=1}^{m} P(x_i) \log_2 P(x_i) \int_{\mathbb{R}^d} P(y|X = x_i) dy =$$

$$= \sum_{i=1}^{m} P(x_i) D_i - \sum_{i=1}^{m} P(x_i) \log_2 P(x_i) = \sum_{i=1}^{m} P(x_i) \left(D_i - \log_2 P(x_i)\right),$$

where $D_i = \int_{\mathbb{R}^d} P(y|X = x_i) \log Q(x_i, y) dy$ does not depend on $P(x_i)$ and can be obtained by Monte Carlo integration (see Lemma 3). Therefore the initial optimisation problem can be reformulated as

$$\max_{P(X)} \quad \sum_{i=1}^{m} P(x_i) \left(D_i - \log_2 P(x_i)\right) \tag{26}$$

$$\text{w.r.t.} \quad \sum_{i=1}^{m} P(x_i) = 1 \tag{27}$$

$$0 \leq P(x_i) \leq 1 \text{ for each } i \tag{28}$$

which can be easily solved by Lagrange multipliers method that yields system of equations

$$\begin{array}{rcl} 0 &=& D_1 - \log P(x_1) - 1 - \lambda \quad \Rightarrow \quad \lambda = D_1 - \log P(x_1) - 1 \\ &\vdots& \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \\ 0 &=& D_m - \log P(x_m) - 1 - \lambda \quad \Rightarrow \quad \lambda = D_m - \log P(x_m) - 1 \\ 0 &=& \left(\sum_{i=1}^{m} P(x_i)\right) - 1 \end{array}$$

It follows that for any i

$$\frac{D_i - \log P(x_i) - 1}{D_1 - \log P(x_1) - 1} = 1 \Rightarrow \frac{P(x_i)}{P(x_1)} = \frac{\exp D_i}{\exp D_1}$$

and using the fact that $\sum_{i=1}^{m} P(x_i) = 1$, we arrive at the formula

$$P^*(x_i) = \frac{\exp(D_i)}{\sum_{l=1}^{m} \exp(D_l)}.$$

The next lemma shows how D_i from Lemma 2 can be estimated from experimental data.

Lemma 3. *Let for each x_i, $P(y|X = x_i)$ be a probability density function. Assuming that a sample $\left(y_1^i,\ y_2^i,\ \ldots,\ y_{n_i}^i\right)$ from distribution $P(y|X = x_i)$ is available (for each x_i), the value of*

$$D_i = \int P(y|X = x_i) \log Q(x_i, y) dy \tag{29}$$

can be calculated using Monte Carlo integration.
Proof
Equation 29 can be rewritten as

$$D_i = \int P(y|X = x_i) \log Q(x_i, y) dy = \mathbb{E}_{P(Y|X=x_i)} \log Q(x_i, Y),$$

where $\mathbb{E}_{P(Y|X=x_i)}$ *is the expectation with respect to distribution* $P(Y|X = x_i)$*. Then, by the law of large numbers, the expectation can be approximated by averaging over the sample from experimental data. Specifically,*

$$D_i = \mathbb{E}_{P(Y|X=x_i)} \left(\log Q(x_i, Y)\right) \approx \frac{1}{n_i} \sum_{j=1}^{n_i} \log Q(x_i, y_j^i).$$

Evidently, this is achievable in practice, if the function $Q(x_i, y)$ *is known and can be evaluated for any* x_i *and* y*.*

Updating logistic regression

In the algorithm, we propose to avoid estimations of logistic regression models in each algorithm's iteration and instead to update the parameters of logistic model according to changes in distribution $P(X)$. Specifically, our algorithm assumes that in its each step next elements of sequence $(P^{(k)}, Q^{(k)})$ are calculated based on their previous values. Especially, $Q^{(k)}(x_i, y) = \hat{P}_{lr}(x_i|Y = y; P^{(k)}(X))$. Theoretically, it would mean that in each step of the iteration, a full estimation of logistic regression model needs to be performed. However, in practice $Q^{(k)}(x_i, y)$ in step k can be analytically calculated from $Q^{(k\text{-}1)}(x_i, y)$. This result is summarised in the Eq. 22. Below we show the exact derivation of this relation in Lemmas 4 and 5.

Lemma 4. *Let* X *be a discrete random variable with values form set* $\{x_1, \ldots, x_m\}$*, while* $P^I(x_i)$ *and* $P^{II}(x_i)$ *are two different probability distributions defined over this set. Additionally, assume that distribution* $P(Y|X = x_i)$ *is fixed for each* x_i *and* $\hat{P}^I_{lr}(X|Y = y)$ *is the logistic regression model obtained from* $P(Y|X)$ *and* $P^I(x_i)$*. Then, the relationship for another logistic model,* $\hat{P}^{II}_{lr}(X|Y = y)$*, based on* $P(Y|X)$ *and* $P^{II}(x_i)$ *is given by formulas*

$$\frac{\hat{P}^{II}_{lr}(x_i|Y = y)}{\hat{P}^{II}_{lr}(x_1|Y = y)} \approx \frac{\hat{P}^{I}_{lr}(x_i|Y = y)}{\hat{P}^{I}_{lr}(x_1|Y = y)} \cdot \frac{P^I(x_1)}{P^I(x_i)} \cdot \frac{P^{II}(x_i)}{P^{II}(x_1)}. \quad \textit{for each } i \tag{30}$$

Proof
Logistic regression is used to predict labels of X from measurements of Y. Then, multinomial logistic regression estimates following equations (label $X = x_1$ is the reference)

$$\hat{P}^I_{lr}(x_1|Y = y) = \frac{1}{1 + \sum_{r=2}^{m} \exp\left(\alpha_r + \beta_r^T y\right)}$$

$$\vdots$$

$$\hat{P}^I_{lr}(x_i|Y = y) = \frac{\exp\left(\alpha_i + \beta_i^T y\right)}{1 + \sum_{r=2}^{m} \exp\left(\alpha_r + \beta_r^T y\right)},$$

$$\vdots$$

$$\hat{P}^I_{lr}(x_m|Y = y) = \frac{\exp\left(\alpha_m + \beta_m^T y\right)}{1 + \sum_{r=2}^{m} \exp\left(\alpha_r + \beta_r^T y\right)},$$

where $y \in \mathbb{R}^d$ is an observed vector of measurements, $\alpha_i \in \mathbb{R}$ are intercepts for i-th class and $\beta_i \in \mathbb{R}^d$ are vector of parameters for i-th class.

The logistic regression model $\hat{P}_{lr}(x_i|Y=y)$ is meant to approximate the true distribution $P(x_i|Y=y)$. On the other hand, using the Bayes formula and definitions of conditional probabilities

$$\frac{P(x_i|Y=y)}{P(x_1|Y=y)} = \frac{\frac{P(y|X=x_i)P(x_i)}{P(y)}}{\frac{P(y|X=x_1)P(x_1)}{P(y)}} = \frac{P(y|X=x_i)}{P(y|X=x_1)} \cdot \frac{P(x_i)}{P(x_1)}$$

and hence

$$\frac{P(y|X=x_i)}{P(y|X=x_1)} = \frac{P(x_i|Y=y)}{P(x_1|Y=y)} \cdot \frac{P(x_1)}{P(x_i)}.$$

If the logistic regression model is a good approximation of the distribution $P(x_i|Y=y)$, then

$$\frac{P(y|X=x_i)}{P(y|X=x_1)} = \frac{P(x_i|Y=y)}{P(x_1|Y=y)} \cdot \frac{P(x_1)}{P(x_i)} \approx \frac{\hat{P}_{lr}(x_i|Y=y;P(X))}{\hat{P}_{lr}(x_1|Y=y;P(X))} \cdot \frac{P(x_1)}{P(x_i)}. \tag{31}$$

Notice that in our scenario, the left-hand side of last equation is assumed to be fixed. Therefore, any changes in priors of right-hand side must be accounted by changes in logistic regression estimates. Eventually, using Eq. (31) twice for distributions $P^I(x_i)$ and $P^{II}(x_i)$, respectively, we get

$$\frac{\hat{P}^I_{lr}(x_i|Y=y)}{\hat{P}^I_{lr}(x_1|Y=y)} \cdot \frac{P^I(x_1)}{P^I(x_i)} \approx \frac{P(y|X=x_i)}{P(y|X=x_1)} \approx \frac{\hat{P}^{II}(x_i|Y=y)}{\hat{P}^{II}(x_1|Y=y)} \cdot \frac{P^{II}(x_1)}{P^{II}(x_i)}. \tag{32}$$

In consequence,

$$\frac{\hat{P}^{II}_{lr}(x_i|y)}{\hat{P}^{II}_{lr}(x_1|y)} \approx \frac{\hat{P}^I_{lr}(x_i|y)}{\hat{P}^I_{lr}(x_1|y)} \cdot \frac{P^I(x_1)}{P^I(x_i)} \cdot \frac{P^{II}(x_i)}{P^{II}(x_1)},$$

which gives a straightforward formula for updating $\hat{P}_{lr}(x_i|y)$.
We can also express the relation (30) in more details, using parameters of the logistic model. Indeed, if we substitute formulas of logistic regression into Eq. 30 we get (for any i)

$$\exp\left(\alpha_i^{II} + \left(\beta_i^{II}\right)^T y\right) = \exp\left(\alpha_i^{I} + \left(\beta_i^{I}\right)^T y\right) \frac{\frac{P^I(x_1)}{P^I(x_i)}}{\frac{P^{II}(x_1)}{P^{II}(x_i)}}.$$

This implies that for each $i \in \{1,\ldots,m\}$ and for any $k > 0$

$$\alpha_i^{II} = \alpha_i^{I} + \log\frac{P^I(x_1)}{P^{II}(x_1)} - \log\frac{P^I(x_i)}{P^{II}(x_i)},$$

$$\beta_i^{II} = \beta_i^{I},$$

what means that only intercepts of the logistic regression model changes with the change of the underlying probabilities $P(X)$.

Lemma 5. *Consider $(k-1)$ and k-th steps of the algorithm presented in Box 1. Then,*

$$Q^{(k-1)}(x_i,y) = \hat{P}_{lr}(x_i|Y=y;P^{(k-1)}(X))$$

is a logistic regression model estimated with the assumption that X follows the distribution $P^{(k-1)}(X)$. Let $P^{(k)}(x_i)$ be an updated input distribution. In order to complete the iteration step, i.e. to find parameters of

$$Q^{(k)}(x_i,y) = \hat{P}_{lr}(x_i|Y=y;P^{(k)}(X))$$

one can use the following formula

$$\frac{Q^{(k)}(x_i,y)}{Q^{(k)}(x_1,y)} \approx \frac{Q^{(k-1)}(x_i,y)}{Q^{(k-1)}(x_1,y)} \cdot \frac{P^{(k-1)}(x_1)}{P^{(k-1)}(x_i)} \cdot \frac{P^{(k)}(x_i)}{P^{(k)}(x_1)}.$$

which relates odds ratio of $Q^{(k-1)}(x_i,y)$ and $Q^{(k)}(x_i,y)$.
Proof
To obtain the above relation, for any $k > 0$, it is enough to use following substitutions in the Lemma 4: $P^I = P^{(k-1)}(x_i)$, $P^{II} = P^{(k)}(x_i)$ and $\hat{P}^I_{lr} = Q^{(k-1)}(x_i,y)$, $\hat{P}^{II}_{lr} = Q^{(k)}(x_i,y)$.

3 Comparison with existing approaches and numerical validation

Below, we employ three test scenarios to examine the performance of the proposed algorithm. In the first two scenarios, we conduct a comparison with the method proposed in [17], which is based on k-nearest neighbor (KNN) density estimation. It is virtually the only method that have been applied to study multivariate signalling responses. We demonstrate that our method outcompetes the KNN-based method of [17] in terms of estimation accuracy, computational time, and robustness to algorithm's settings. In the third scenario, we show that our approach provides accurate estimates regardless of the functional form of the output distribution.

3.1 Test scenario 1: the proposed approach ensures accurate, efficient and robust estimation of information capacity

As discussed in Section 1.2, the KNN-based method is virtually the only available method that enables computation of the information capacity for systems with multidimensional outputs, Y. To show that our method indeed provides a significant advantage over the KNN based estimation of information capacity, we replicate a test model introduced by the authors of KNN based method [17].
The test model considers a channel with two possible input values, $X \in \{x_1, x_2\}$, and output $Y|X = x_i$ given by a d-dimensional Gaussian distributions with identical, diagonal covariance matrices, and mean vectors that are the same except the first dimension. Precisely,

$$Y|x_i \sim \mathcal{N}(\mu_i, \Sigma); \quad Y \in \mathbb{R}^d, \text{ for } i = 1, 2;$$

$$\mu_1 = (0, 0, \ldots, 0), \quad \mu_2 = (2, 0, \ldots, 0);$$

$$\Sigma = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

For the above model, we used our approach and the KNN based method to calculate capacities for different dimensions, d, of the output Y. Besides, we considered different sample sizes, N, i.e., the number of observations corresponding to each input value. The KNN based method requires specification of the parameter k that determines the number of neighboring data points that are used to estimate the probability density of the output. Following [17], we initially set $k = 10$, and, thereafter, analysed dependence of the estimates on k.

Fig S1A shows capacities estimated using our approach (blue) and the KNN based method (red) for d ranging from 2 up to 30, for three sample sizes, N, 500, 2000, and 4000. For reference, true capacity calculated using numerical integration and optimization of the exact model is also plotted. Clearly, our approach outcompetes the KNN method in terms of bias, i.e. difference between the mean estimate and the true value, for all dimensions and sample sizes. The bias of the KNN based estimates is most striking for smallest considered sample size ($N = 500$). Moreover, Fig S1B shows that estimates obtained with our method converge quickly, with increasing N, to the true value of capacity, whereas estimates of the KNN method remain biased even for large sample size. Computation times corresponding to panels A and B are plotted in panels C and D, respectively. In this example, our method is approximately two-fold faster than the KNN approach. Relatively minor gain results from the model having only two input values. Further, in test scenario 2, we show that the gain can 100-fold for > 10 input values.

Importantly, in contrast to the KNN based method, our approach is robust to arbitrary assumptions regarding the parameters of the algorithm. Precisely, the KNN method involves setting the parameter k that determines the number of neighboring data points that are used to estimate the probability density of the output. As a result, the obtained capacity estimates are not robust. Our approach is free from such assumptions and, hence, guarantees robust estimation. For the KNN based estimates presented in

FigS1A and B, we have used $k = 10$, as suggested by the authors of [17]. However, the estimate values are influenced by choice of the parameter k. Therefore, the comparison of FigS1A and B present a scenario that is optimistic for the KNN based approach. In Fig S1E, we show that the choice of k can introduce even stronger bias of the KNN based information capacity estimates, which further demonstrates advantage of our method.

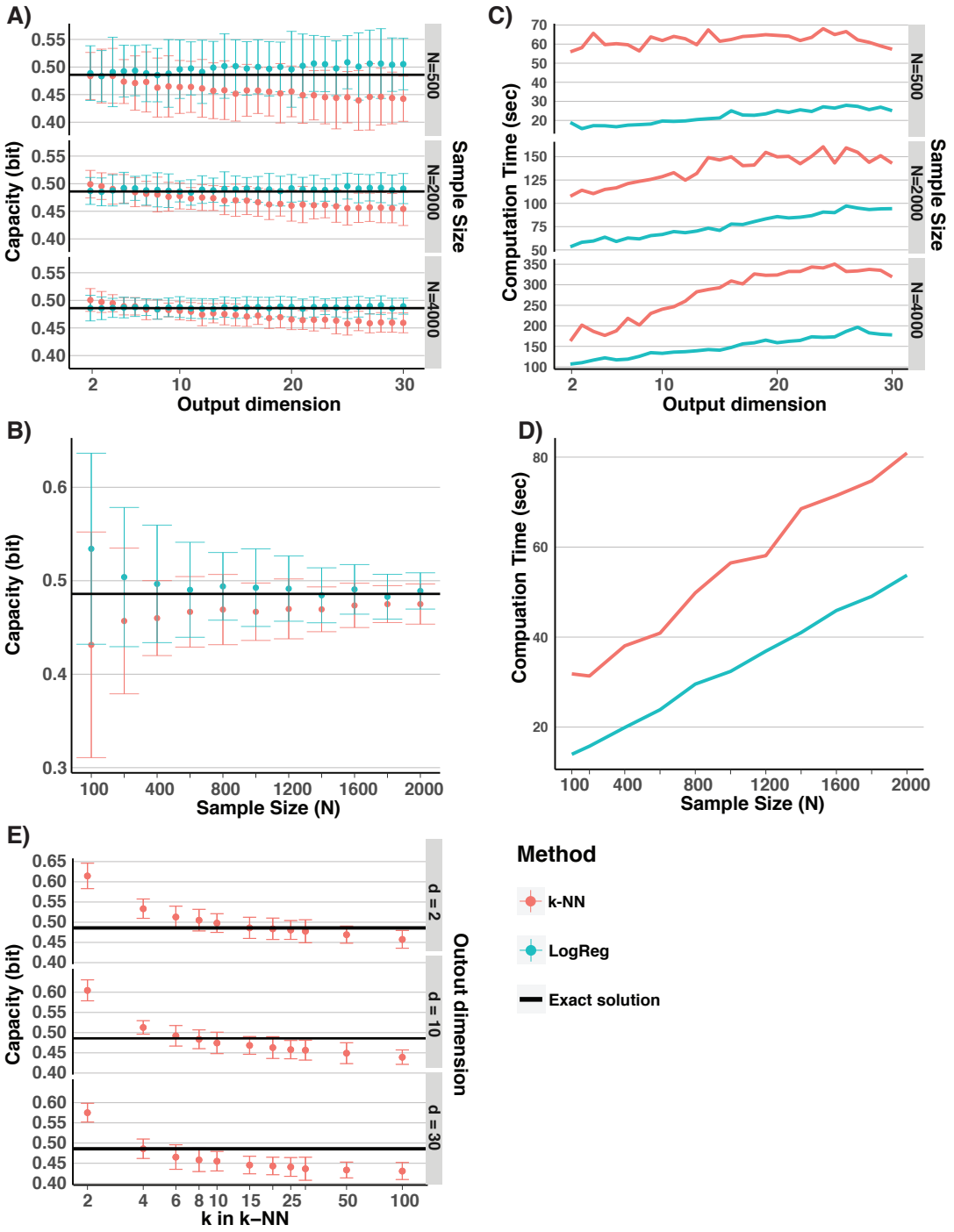

Figure S1: Test scenario 1. **(A)** Information capacity estimates as a function of the dimension d for three different sample sizes N. Blue and red dots correspond to our method and the KNN based method, respectively. Black bold line marks the true value of capacity. **(B)** Information capacity estimates as a function of the sample size N for $d = 10$. Lines are coloured as in (A). **(C)** Computation time needed to obtain a single estimate in (A). **(D)** Computation time needed to obtain a single estimate in (B). **(E)** Information capacity estimates of the KNN based method as a function of k for three different dimension d and $N = 2000$. The error-bars in all panels (A),(B),and (E) show standard deviation of capacity estimates from 40 repeated samplings. In panels (A-D), k=10 was assumed. The times reported in panels (C) and (D) correspond to computations performed by a single core on a workstation with Intel® Xeon® E5-1650 3.50 GHz processor and 32 GB RAM.

3.2 Test scenario 2: use of the KNN based method may lead to significantly biased estimates

The test scenario 1, indicates that, in contrast to the KNN method, our framework provides more accurate estimates of the capacity as well as is computationally more efficient. Importantly, it guarantees robust estimation as estimate values do not depend on parameters of the algorithm. Nevertheless, the differences in accuracies and computational time are not large. The test example was, however, replicated from [17] that introduced the KNN-based method. Below, we provide a simple example that demonstrates that for small sample size KNN-based estimation may lead to substantially biased estimates even for one-dimensional distributions. Moreover, we demonstrate that computational advantage of our method is significant.

The example aims to reflect an experimental setup, in which one-dimensional responses of individual cells to a range of stimuli are quantified. Precisely, the test model considers a channel with the log-normally distributed output, Y. The mean, $\mu(x)$ and variance σ^2 of the log-output are assumed to be the sigmoid function, and a constant, respectively. Precisely,

$$Y|x_i \sim \exp\left(N(\mu(x), \sigma^2)\right),$$

for $\mu(x) = 10 \cdot \frac{x}{1+x}$, $\sigma^2 = 1$.
For the model, we generated samples of the output Y for 11 different input values x_i ranging from 0 to 100 (Fig. S2A). Thereafter, we examined how accuracy of estimation of the information capacity depends on sample size, N. We have assumed values of N starting from 50 up to 2000. For sample sizes typical for biological experiments, i.e. tens or hundreds of measured cells, our method provides substantially more accurate estimates (Fig. S2B). The accuracy of estimation of the KNN method further deteriorates for different values of k, (Fig. S2C).
We also report computational times of both methods as a function of the number of considered input values, x_i's, (Fig. S2D) and of the sample size, N (Fig. S2E). Our method is orders of magnitude faster for the number of input values larger than 6. Moreover, it has much better scaling properties with respect to both the number of input values and the sample size.

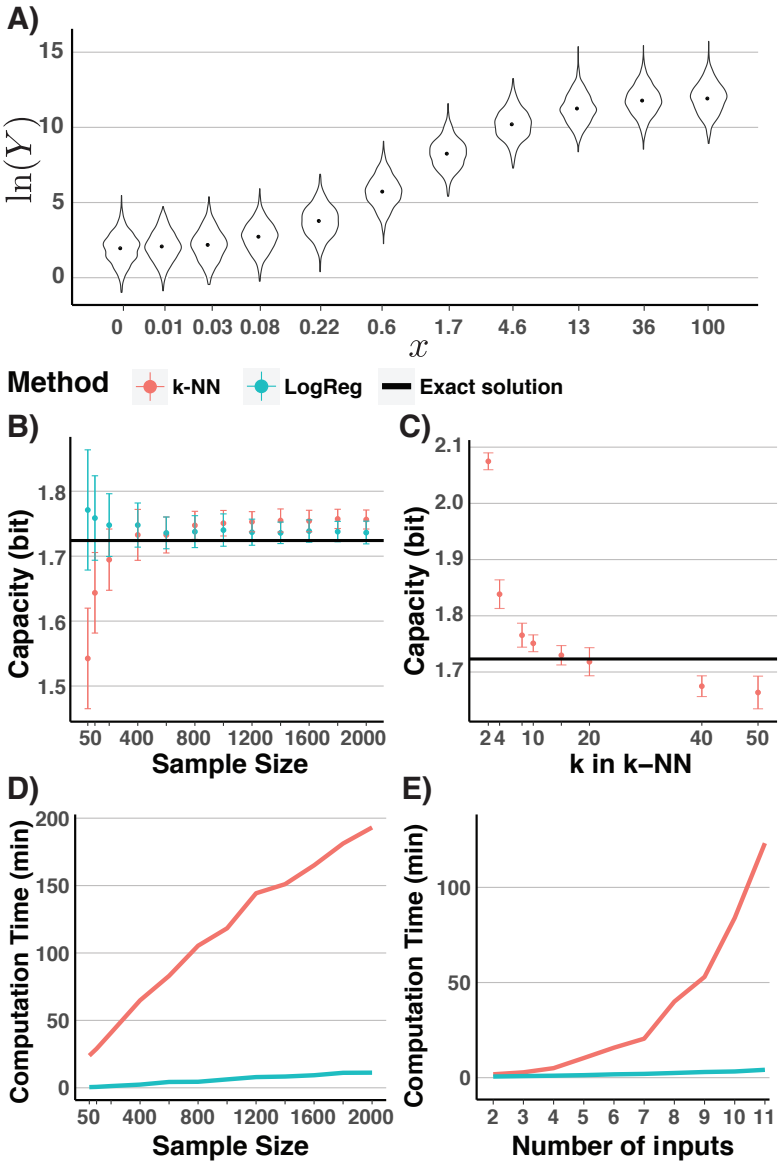

Figure S2: Test scenario 2. **(A)** A violin plot representation of the conditional output distribution $Y|x_i$ for 11 different inputs. **(B)** Information capacity estimates as a function of the sample size N Blue and red lines correspond to our method and the KNN based method, respectively. The bold black line marks the true value of capacity. For the KNN estimation, $k = 10$ was assumed. **(C)** Information capacity estimates of the KNN method as a function of k compared with the true value (bold black line). The error-bars in B and C show the standard deviation of capacity estimates from 40 repeated samplings. $N = 1000$ was assumed. **(D)** Computation time of our method (blue) and the KNN method as a function of the sample size N. **(E)** Computation time of our method (blue) and the KNN method as a function of the number of considered input values. Input values were subsequently added starting with x_1 and x_2, only, and ending up with all 11 considered input values. The times reported in panels (D) and (E) correspond to computations performed by a single core on a workstation with Intel® Xeon® E5-1650 3.50 GHz processor and 32 GB RAM.

3.3 Test scenario 3: the proposed approach provides accurate estimates regardless of the specific form of the output distribution

The theory described in Section 2 guarantees the correctness of our approach if the classifier based on logistic regression, $\hat{P}_{lr}(x_i|y)$, is a good approximation of the conditional input probabilities, $P(x_i|y)$. The approximation may not be accurate if the conditional input probabilities cannot be represented by the functional form assumed by logistic regression, Eq. 21. The conditional input probabilities, $P(x_i|y)$, are dependent, via Bayes formula, on the conditional output probabilities, $P(y|x_i)$. Therefore, to test the robustness of our approach against different forms of the conditional input distributions, $P(x_i|y)$, we have assumed various output distributions, $P(y|x_i)$. Our test strategy is also implied by practical applications in which, input distributions, $P(x_i|y)$, are functions of experimentally measured output distributions, $P(y|x_i)$.
Specifically, we have examined four different forms of univariate conditional output probability densities $P(y|x_i)$, with five possible input values, $X \in \{x_1, \ldots, x_5\}$ (Fig. S3A). Precisely, we considered

Scenario 1: $Y|x_i \sim$ Exponential

Scenario 2: $Y|x_i \sim$ Gamma

Scenario 3: $Y|x_i \sim$ Normal

Scenario 4: $Y|x_i \sim$ Lognormal.

For each of the above distributions, we have assumed that

- the difference between expected values of subsequent output's conditional distributions is fixed at 1, i.e.,
$$\mathbb{E}(Y|X = x_{i+1}) - \mathbb{E}(Y|X = x_i) = 1$$
- for each i, samples of data are of the same size, N, i.e.
$$Y|X = x_i \quad \sim \quad \left(y_1^i,\ y_2^i,\ \ldots,\ y_N^i\right)$$
- the variance, σ^2 is varied between 0.01^2, 10^2 yielding 14 variants in total,
- samples size, N, equals to either $100, 500$ or 1000.

Moreover, for Gamma, Normal and Lognormal distributions we assumed that for each input i, the variance of the distribution $Y|X = x_i$ is the same, i.e.
$$\mathrm{Var}(Y|X = x_i) = \sigma^2$$
Combinations of different forms of the input distribution, values of N and σ resulted in 168 different settings. For each setting, we have calculated capacity using our method. For a benchmark, we computed true channel capacity by numerical integration of Eq. 6 and extensive numerical optimization. Deviations of the capacity calculated with our method from the true capacity as a function of the standard deviation σ are presented in Fig. S3B-E. In the considered 168 different settings the absolute error of our method did not exceed 0.1 bit and typically was lower than 0.03 bits. We report absolute errors rather than actual capacities compared to the true values as for each level of the standard deviation the capacity may vary according to other factors considered.

The above test example allows us to conclude that our method is robust to the shape of the output distribution, at least as long distributions typical for cellular signaling systems are considered.

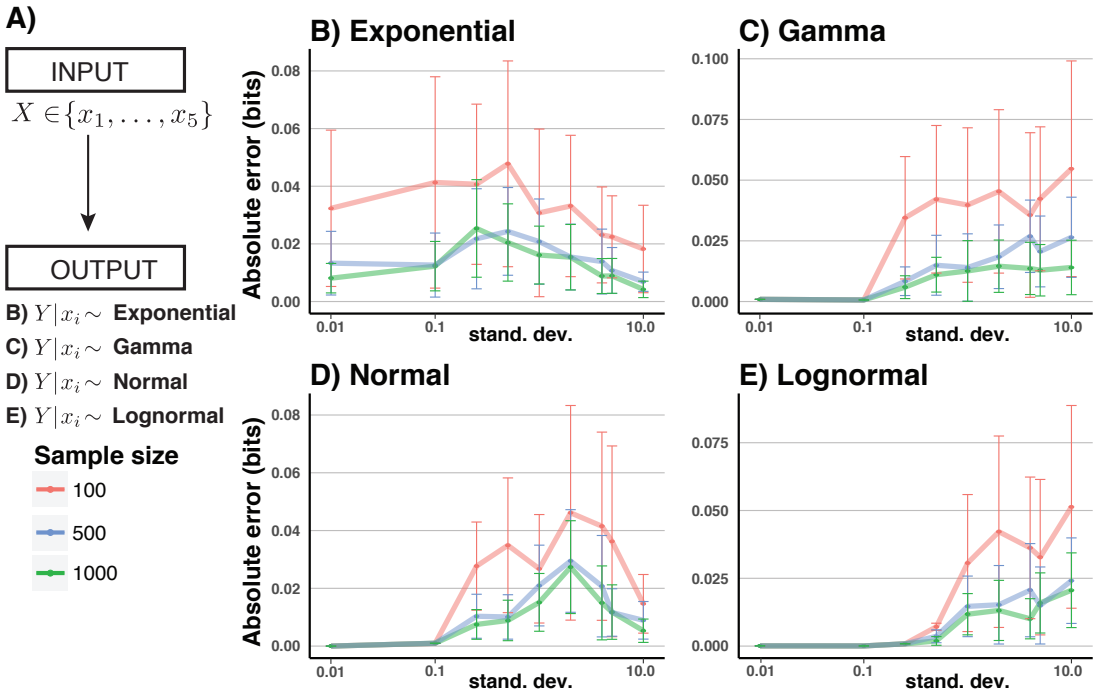

Figure S3: Test scenario 3. **(A)** Schematic representation of the test model. **(B-E)** The absolute error of the capacity calculated using our approach with respect to the capacity calculated exactly, i.e., using numerical integration and optimisation. Absolute errors are plotted as a function of the standard deviations of the output, distributions, σ. Each panel corresponds to different type of output distribution, as indicated. Individual lines correspond to different sample size N. The error-bars show standard deviations of capacity estimates based on 40 repeated samplings.

4 Analysis of the Nf-κB responses to TNF-α

The NF-κB signaling is one of the key pathways involved in the control of the immune system and one of the first cellular signalling systems studied within the framework of information theory. So far, several papers quantified its information capacity [17, 4, 12, 26]. The main points of interest include whether single-cell responses are switch-like or encode more information about the quantity of stimulus. [4, 17, 26, 21]. In addition, stimuli have been shown to regulate temporal profiles of signaling effectors? activities and response dynamics turned out to augment information transmission compared to non-dynamic responses [17, 26]. Given the broad interest to study NF-κB responses within the framework of information theory we have selected the pathway to demonstrate that our framework can indeed provide a novel biological insight.
Broadly speaking, in unstimulated cells the NF-κB protein complex resides in the cytoplasm. Upon activation of the TNF-α receptor, a cascade of events leads to nuclear translocation of NF-κB where it acts as a transcription factor. Activated genes participate in a feedback loop that results in subsequent export of the NF-κB out of the nucleus. The nuclear level of NF-κB is often considered as an immediate output of the system. To examine TNF-α induced NF-κB responses, we have performed a set of experiments in which we stimulated cells with a range of doses of TNF-α. In the examined cell line relA protein, which is a part of the NF-κB complex, was stably fused with a fluorescent dsRed protein. Therefore, life imaging could be used to quantify cytoplasmic and nuclear dsRed fluorescence over time in individual cells. Fluorescence levels were than used as a proxy of cytoplasmic and nuclear levels of NF-κB. The normalised data obtained in the experiments are shown in Fig. S4C-D. The date were used to perform analysis presented in the Fig. 1 of the main paper. Below we discuss details related to experimental methods, image acquisition and data analysis.

4.1 Experimental methods

We used immortalised murine embryonic fibroblasts cell line (3T3) expressing fluorescent fusion proteins relA-dsRed as wells H2B-GFP for nuclei identification. The cell line was kindly provided by prof. S. Tay and was previously used in several studies, including [21, 12, 22, 15]. Cells were grown and maintained in a conditioned incubator at 37°C, 5 % CO2 in transparent DMEM medium (Life Technologies). As preparation, 48 hours before the experiment, 3T3 cells was resuspended in 3ml of transparent DMEM medium on a confocal dish and keep separately in incubator. The temperature in the microscope's chamber was set to 37°C and CO2 influx to 5% and 10L. Prior to TNF-α stimulation and imaging. Confocal dish with cultured cells was put in the chamber for an initial phase for about 20 minutes. For NF-α stimulation medium was sucked out from plate and replaced with a solution of TNF-α and DMEM, which is considered as the start of the experiment (t=0 in the presentation of results). Overall cells were stimulated with 11 different doses of 5-minute pulses of TNF-α (Sigma-Aldrich), ranging from 0 ng/ml to 100 ng/ml. Live imaging was performed using confocal microscope, Leica TCS SP5 X with environmental chamber. Channels of dsRed and GFP were used for NF-κB levels and nuclear area identification, respectively. A single imaging experiment lasted 3 hours, during which an image has been captured every 3 minute both in green (GFP) and red (dsRed) channels, simultaneously at 9 different position at the plate. Experiment with each concentration of stimulation has been repeated at least four times to test reproducibility and to allow for a sufficient number of observations.

4.2 Image analysis

Images captured every three minutes have been saved as LIF files and then exported to TIFF format using Leica Application Suite 4. TIFF images were segmented to identify nuclear and cytoplasmic areas of individual cells. Individual cells at subsequent images have been tracked to reconstruct cytoplasmic and nuclear fluorescence in individual cells over time. For each stimulation level we have quantified fluorescence in over 700 hundred cells yielding over 10 000 cells in total. This has been achieved using a customised CellProfiler [10] pipeline as well as manual quality control. Specifically, the developed

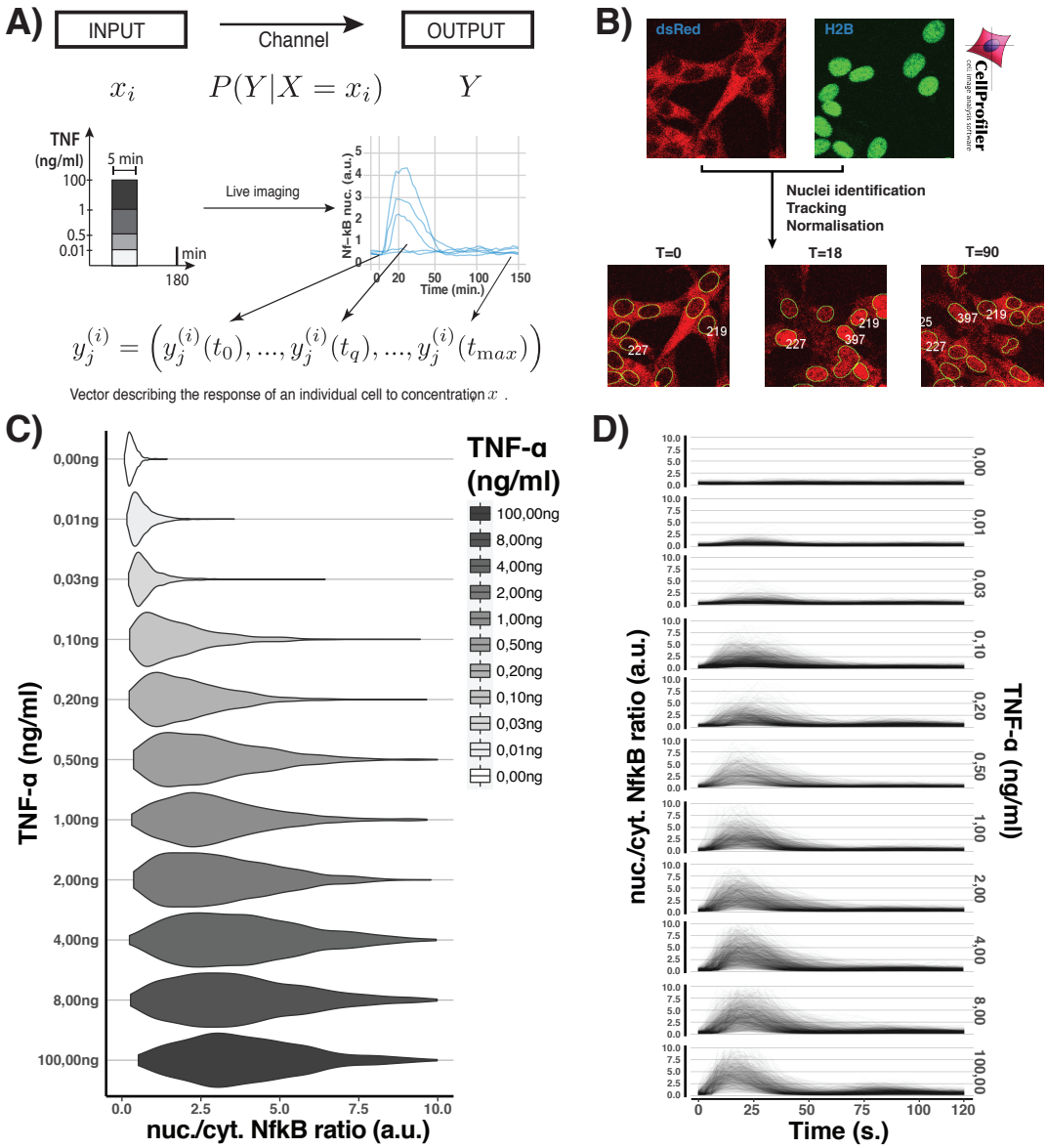

Figure S4: Analysis of the Nf-κB responses to TNF-α. (**A**) Nf-κB responses can be seen as a communication channel. Concentration of TNF-α, x, is considered as input that induces nuclear translocation of the Nf-κB complex. Temporal profile of nuclear Nf-κB level is then considered as output,and the input-output relationship is represented as the probability distribution $P(Y|X = x)$. In experiments, we performed 5-minute pulses of TNF-α stimulation and imaged single cells using confocal microscopy. Images were taken every 3 minutes for 3 hours. (**B**) Acquired images were processed using a customised CellProfiler pipeine. Images were acquired in two channels, corresponding to dsRed and GFP fluorescent proteins. In the examined cell line relA protein, which is a part of the NF-κB complex, was stably fused with a fluorescent dsRed protein. Therefore, life imaging could be used to quantify cytoplasmic and nuclear dsRed fluorescence over time in individual cells. Fluorescence levels were than used as a proxy of cytoplasmic and nuclear levels of NF-κB. In addition,in the studied cell line, the histone protein H2B was fused to the GFP , therefore the GFP fluorescence for nuclear image segmentation. Nuclei were tracked at from 0 to >120 minute with standard algorithm available in CellProfiler. The manipulation of images and data analysis were perforemd in Python, ImageMagick and R. (**C**) Violin plots of NF-κB responses at 21 min. after stimulation with different doses of TNF-α . (**D**) Temporally resolved responses of single cells (1 line = 1 identified cell) to different doses of TNF-α .

CellProfiler pipeline (available from authors upon request) allowed us to : 1) segment images to identify nuclear area of individual cells (based on GFP fluorescence emitted by histone fluorescent fusion protein H2B-GFP in our cell line); 2) segment cytoplasmic area of individual cells (based on both dsRed and GFP fluorescence); 3) track nuclei between subsequent time frames by simple overlap approach; 4) quantify dsRed fluorescence of the nucleus, $\mathrm{NfKB}_{nuc.}$, in individual cells; 5) quantify dsRed background fluorescence (defined as area not occupied by cells), $\mathrm{NfKB}_{back.}$;
After automated analysis using the above pipeline, we carried out a manual quality control for detection of segmentation and tracking errors. Cytoplasmic Nf-κB , $\mathrm{NfKB}_{cyt.}$ was estimated by enlarging each nuclei by 3 pixels and calculating mean dsRed fluorescence within the ring between enlarged and original nuclei. Next, we substracted background fluorescence, $\mathrm{NfKB}_{back.}$, both from $\mathrm{NfKB}_{nuc.}$ and $\mathrm{NfKB}_{cyt.}$. To account for photo-bleaching, a simple exponential model was fitted to the time series of the total fluorescence of experimental images and the resulting bleaching factor has been used to scale $\mathrm{NfKB}_{nuc.}$ and $\mathrm{NfKB}_{cyt.}$, respectively. Finally, a ratio $\frac{\mathrm{NfKB}_{nuc.}}{\mathrm{NfKB}_{cyt.}}$ is used as a measure of Nf-κB activation in cells. As a last step, due to small discrepancies in image acquisition times, measurements were interpolated into an unified time interval (from 0 to 120, every 3 minutes) using a simple linear interpolation method. All corrections and normalisation have been conducted in R using custom code and standard packages (available from authors upon request).

4.3 Experimental data set

The data obtained in the experiments can be conceptually represented as

$$y_j^{(i)} \quad \sim \quad P(Y|X = x_i), \tag{33}$$

where x_i is the concentration of the 5 minutes pulse TNF-α stimulation, and $y_j^{(i)}$ is a vector that describes the response of the i-th cell to stimulation level x_i. Precisely,

$$y_j^{(i)} = \left(y_j^{(i)}(t_0), ..., y_j^{(i)}(t_{\mathrm{max}}) \right), \tag{34}$$

where $y_j^{(i)}(t_q)$ is the nuclear to cytoplasmic ratio of the i-th cell at time t_q. The data contain responses for 11 different concentrations, i.e. $x_i \in \{0, 0.01, 0.03, 0.1, 0.2, 0.5, 1, 2, 4, 8, 100\}$ ng/ml. The responses were measured every three minutes for two hours. Precisely, $t_0 = 0$, $t_q - t_{q-1} = 3$ and $t_{\mathrm{max}} = 120$. Measured responses are presented in Fig. S4C and Fig. S4D.

4.4 Quantification of information capacity

Capacities of time - point responses (Fig. 1B in MP) were computed by assuming that the output is composed of individual timepoints

$$y_j^{(i)}(t_{\mathrm{q}}), \tag{35}$$

for t_{q} from t_0 to t_{max}. On the other hand, capacities of time - series responses (Fig. 1C in MP) were calculated for complete responses truncated at time t_{tr}

$$\left(y_j^{(i)}(t_0), \ldots, y_j^{(i)}(t_{tr}) \right), \tag{36}$$

and calculated capacities for t_{tr} from t_0 to t_{max}.

Uncertainties of the information capacities have been estimated using bootstrap re-sampling. Precisely for each capacity estimate presented in Fig. 1, we have computed standard deviation of the capacity calculated 50 times using random 80% of the initial dataset. Grey ribbons denote mean capacity $\pm$ standard deviation.

4.5 Probabilities of correct discrimination between input values

The information capacity, C^*, can be seen as a measure of an overall signaling fidelity and 2^{C^*} can be interpreted, within the Shannon's coding theorem [5, 18, 6], as the number of input values that the system can effectively resolve [6]. Nevertheless, information capacity does not give a direct answer on how to determine these distinguishable states. Similarly, it does not indicate what an error of discrimination between different input values is.

Conveniently, our framework enables calculation of probabilities of correct discrimination between different input values directly from experimental data. Precisely, consider a task of guessing the input value x_i, for i ranging from 1 to m, based on observing the specific output y. It can be shown [6] that the strategy that maximizes the probability of correct guessing is the maximum *a posteriori* rule that selects the input value that has highest *a posteriori* probability. Denoting the most likely input value that led to the specific output y as $\hat{x}(y)$, the maximum *a posteriori* rule can be formally written as

$$\hat{x}(y) = \arg \max_{x_1,\ldots,x_m} \hat{P}\left(x|Y=y\right). \tag{37}$$

As described in Section 2, our algorithm to calculate the information capacity, C^* is based on the approximation of the conditional input probabilities, $P(x_i|Y=y)$, which is the same as *a posteriori* probabilities, with the logistic regression estimates, $\hat{P}_{lr}(x_i|Y=y)$. Therefore, for any output y, the *a posteriori* probability, and the most likely input value can be determined. These can then be used to calculate the probabilities of correct discrimination by counting the average number of correctly assigned observations, y_j^i. Precisely, consider data of a typical experiment aimed to quantify information capacity, as described for the NF-κN system and generally in Section 1.1. Then, the probabilities of correct discrimination can be calculated with logistic regression, $\hat{P}_{lr}(x_i|Y=y)$, as

$$P(\text{correct discrimination}) = \frac{1}{n_1 + \ldots + n_m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{1}(\hat{x}(y_j^i) = x^i), \tag{38}$$

where

$$\hat{x}(y_j^i) = \arg \max_{x_1,\ldots,x_m} \hat{P}_{lr}\left(x|y_j^i\right), \tag{39}$$

and $\mathbf{1}(\cdot)$ is the indicator function that takes the values of 1 for $\hat{x}(y_j^i)$ equal to x^i and 0 in the opposite case. In the Figure 1 D and E of the main paper, we presented probabilities of correct discrimination of the NF-κB system. Precisely, we calculated probabilities of correct discrimination between pairs of input values. These can be formally written as

$$P(\text{correct discrimination between } x_v \text{ and } x_u) = \frac{1}{n_v + n_u} \sum_{i \in \{u,v\}} \sum_{j=1}^{n_i} \mathbf{1}(\hat{x}(y_j^i) = x^i). \tag{40}$$

For calculations, we assumed equally probable input signals. We performed 50 bootstrap re-sampling, and in the figures, we presented means of the obtained probabilities of correct discrimination. It is worth noting that probability of correct discrimination for pairs of signals is ≥ 0.5 as mere random guessing guarantees 0.5 chance of being correct.

References

[1] Ian S Abramson. Adaptive density flattening – a metric distortion principle for combating bias in nearest neighbor methods. *The Annals of Statistics*, pages 880–886, 1984.

[2] Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

[3] Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

[4] R Cheong, A Rhee, C J Wang, I Nemenman, and A Levchenko. Information Transduction Capacity of Noisy Biochemical Signaling Networks. *Science*, 334(6054):354–358, 2011.

[5] Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994.

[6] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[7] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[8] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial Intelligence and Statistics*, pages 277–286, 2015.

[9] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2016.

[10] Lee Kamentsky, Thouis R Jones, Adam Fraser, Mark-Anthony Bray, David J Logan, Katherine L Madden, Vebjorn Ljosa, Curtis Rueden, Kevin W Eliceiri, and Anne E Carpenter. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics*, 27(8):1179–1180, 2011.

[11] S Sathiya Keerthi, K B Duan, Shirish Krishnaj Shevade, and Aun Neow Poo. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61(1):151–165, 2005.

[12] Zbigniew Korwek, Karolina Tudelska, Paweł Nałęcz-Jawecki, Maciej Czerkies, Wiktor Prus, Joanna Markiewicz, Marek Kochańczyk, and Tomasz Lipniacki. Importins promote high-frequency NF-κB oscillations increasing information channel capacity. *Biology Direct*, 11(1):61, 2016.

[13] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[14] Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, pages 1049–1051, 1965.

[15] Jakub Pękalski, Paweł J Żuk, Marek Kochańczyk, Michael Junkin, Ryan Kellogg, Savaş Tay, and Tomasz Lipniacki. Spontaneous NF-κB activation by autocrine TNF-α signaling: a computational analysis. *PLOS ONE*, 8(11):e78887, 2013.

[16] Barnabás Póczos and Jeff Schneider. On the estimation of α-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617, 2011.

[17] Jangir Selimkhanov, Brooks Taylor, Jason Yao, Anna Pilko, John Albeck, Alexander Hoffmann, Lev Tsimring, and Roy Wollman. Accurate information transmission through dynamic biochemical signaling networks. *Science*, 346(6215):1370–1373, 2014.

[18] Claude E Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27:623–656, 1948.

[19] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

[20] Kumar Sricharan, Raviv Raich, and Alfred O Hero. K-nearest neighbor estimation of entropies with confidence. In *Proceedings of 2011 IEEE International Symposium on Information Theory*, pages 1205–1209, 2011.

[21] Savaş Tay, Jacob J Hughey, Timothy K Lee, Tomasz Lipniacki, Stephen R Quake, and Markus W Covert. Single-cell NF-κB dynamics reveal digital activation and analog information processing in cells. *Nature*, 466(7303):267, 2010.

[22] Savaş Tay, Jake Hughey, Tim K Lee, Markus Covert, and Stephen R Quake. Cells respond digitally to variation in signal intensity via stochastic activation of NF-κB. *Biophysical Journal*, 98(3):429a–430a, 2010.

[23] Margaritis Voliotis, Rebecca M Perrett, Chris McWilliams, Craig A McArdle, and Clive G Bowsher. Information transfer by leaky, heterogeneous, protein kinase signaling systems. *Proceedings of the National Academy of Sciences*, 111(3):E326–E333, 2014.

[24] Pascal O Vontobel, Aleksandar Kavcic, Dieter M Arnold, and Hans-Andrea Loeliger. A generalization of the Blahut-Arimoto algorithm to finite-state channels. *IEEE Transactions on Information Theory*, 54(5):1887–1918, 2008.

[25] Raymond W Yeung. *Information Theory and Network Coding*. Springer Science & Business Media, 2008.

[26] Qiuhong Zhang, Sanjana Gupta, David L Schipper, Gabriel J Kowalczyk, Allison E Mancini, James R Faeder, and Robin EC Lee. NF-κB dynamics discriminate between TNF-α doses in single cells. *Cell systems*, 5(6):638–645, 2017.