RESEARCH ARTICLE

Information-theoretic analysis of multivariate single-cell signaling responses

Tomasz Jetka[ID], Karol Nienałtowski[ID], Tomasz Winarski, Sławomir Błoński, Michał Komorowski[ID]*

Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

* m.komorowski@sysbiosig.org

Abstract

Mathematical methods of information theory appear to provide a useful language to describe how stimuli are encoded in activities of signaling effectors. Exploring the information-theoretic perspective, however, remains conceptually, experimentally and computationally challenging. Specifically, existing computational tools enable efficient analysis of relatively simple systems, usually with one input and output only. Moreover, their robust and readily applicable implementations are missing. Here, we propose a novel algorithm, SLEMI —statistical learning based estimation of mutual information, to analyze signaling systems with high-dimensional outputs and a large number of input values. Our approach is efficient in terms of computational time as well as sample size needed for accurate estimation. Analysis of the NF-κB single—cell signaling responses to TNF-α reveals that NF-κB signaling dynamics improves discrimination of high concentrations of TNF-α with a relatively modest impact on discrimination of low concentrations. Provided R-package allows the approach to be used by computational biologists with only elementary knowledge of information theory.

Author summary

In light of single-cell, live-imaging experiments understanding of how cells transmit information about identity and quantity of stimuli is incomplete. When exposed to the same stimulus individual cells exhibit substantial cell-to-cell heterogeneity. Besides, stimuli have been shown to regulate temporal profiles of signaling effectors. Therefore, it is, for instance, not entirely clear whether single-cell responses are binary or contain more information about the quantity of stimuli. The above questions resulted in a considerable interest to study cellular signaling within the framework of information theory. Unfortunately, the utilization of the information-theoretic perspective is handicapped in part by the lack of suitable methods that account for multivariate signaling data. Here, we propose a novel algorithm that breaks a considerable computational barrier by allowing the effective information-theoretic analysis of highly-dimensional single-cell measurements. Our approach is computationally efficient, robust and straightforward to use. Moreover, we provide a simple R-package implementation.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Biochemical descriptions of cellular signaling appear to require quantitative support to explain how complex stimuli (inputs) are translated and encoded in activities of pathway's effectors (outputs) [1–5]. An attractive approach seems to be offered by probabilistic modeling and information theory [6–10], which provide a mathematical language to describe input-output relationships of complex and stochastic cellular processes. Specifically, the unique perspective of information theory holds a promise of gaining new insights into functional aspects of signaling, as opposed to biochemical and mechanistic descriptions [1, 9, 11–16]. So far, quantification of an overall signaling fidelity and analysis of factors by which it is determined have proven to be useful applications of information theory in studies of cellular signaling [5, 17–19]. Nevertheless, exploring the information-theoretic approach remains conceptually and technically challenging [7, 9, 10]. In particular, for systems with multiple inputs and outputs existing theoretical tools are computationally inefficient and require a large sample size for accurate analysis.

Within information theory, regardless of specific details of a signaling pathway, a signaling system can be considered as an input-output device that measures an input signal, x, by eliciting a stochastic output, Y [7, 9, 10]. In a typical example, the input, x, is the concentration of a ligand, e.g., cytokine, that activates a receptor. The output, Y, is an activity of one or more signaling effectors, e.g., of transcription factors quantified over time. As cellular signaling systems are inherently stochastic, the information about the input contained in the output is imprecise and only a limited number of input values can be resolved [6, 15, 20, 21]. To date, a number of studies have experimentally examined fidelity of various signaling systems, e.g., [13, 15, 19, 22–24]. In a typical experiment aimed to quantify fidelity, input values, $x_1 \leq x_2 \ldots \leq x_m$, ranging from 0 to saturation are considered. In some scenarios, utilization of physiologically arising input concentrations, e.g., morphogen gradients, is also possible [5, 16, 25]. Then, for each input level, x_i, cellular responses are quantified in a large number, say n_i, of individual cells. Single-cell responses are typically represented as vectors, y_l^i, that contain entries with quantified activities of signaling effectors, where l varies from 1 to the number of measured cells, n_i. Further, cell-to-cell heterogeneity of responses is simplistically used as a proxy of how reproducible signaling output of an individual cell is. Formally, responses corresponding to each of the inputs, x_i, are assumed to follow a probability distribution

$$y_l^i \sim P(Y|X = x_i), \tag{1}$$

which is reconstructed from the data and serves as a model of the single-cell response to the input x_i.

Given the above, information theory offers to quantify the overall fidelity of signaling in terms of how many input values, x_i, can be resolved based on information contained in the responses. The key factor that determines how many inputs can be resolved is the degree of the overlap between the distributions corresponding to different inputs. For instance, two inputs x_1 and x_2 can be easily resolved if the corresponding output distributions $P(Y|X = x_1)$ and $P(Y|X = x_2)$ are completely distinct, non-overlapping. Then, a given response, y, can be assigned without error to the only input for which it can occur. In the scenarios in which $P(Y|X = x_1)$ and $P(Y|X = x_2)$ are overlapping the inputs x_1 and x_2 cannot be resolved as a given response y is equally likely to arise for both inputs. The second factor that is essential for formal quantification of the overall fidelity is how frequently the different inputs occur. For

illustration, consider a system with three input values, x_1, x_2 and x_3. Suppose, x_1 and x_2 induce very similar output distributions, i.e., $P(Y|X = x_1) \approx P(Y|X = x_2)$, whereas the input x_3 induces distribution $P(Y|X = x_3)$ that is distinct, non-overlapping, with the first two. How many inputs are on average resolved in this system depends on how frequently different inputs occur. If for instance, x_1 and x_3 occur frequently, e.g., $P(x_1) = P(x_3) \approx 1/2$, whereas, x_2 has a negligible incidence, i.e., $P(x_2) \approx 0$, the fidelity of signaling allows to resolve two inputs on average. This is because the two inputs that can be resolved occur with high probability. On the other hand, if inputs x_1 and x_2 occur with probability $\approx 1/2$, and x_3 has probability close to 0, then only one state can be resolved on average, i.e., concatenation of x_1 and x_2. The frequencies of inputs are taken into account in the form of the input distribution $P(X) = (P(x_1), \ldots, P(x_m))$.

The degree of the overlap between the output distributions as well as frequencies of inputs are used to evaluate the overall signaling fidelity in the form of the mutual information

$$MI(X, Y) = \sum_{i=1}^{m} P(x_i) \int_{\mathbb{R}^d} P(y|X = x_i) \, \log_2 \frac{P(y|X = x_i)}{P(y)} \, dy, \tag{2}$$

where $P(y)$ is the overall distribution of the output implied by a given distribution of the input, i.e., $P(y) = \sum_{i=1}^{m} P(y|X = x_i)P(x_i)$. MI is expressed in bits and 2^{MI} can be interpreted as the number of inputs that the system can resolve on average. Definition of mutual information stems from axioms proposed by C. Shannon [26] and can also be written in a more intuitive form of entropy differences, see supporting S1 Text. Selection of the input distribution that is suitable for quantification of information transfer in a specific application can be problematic and provides a degree of arbitrariness. The uniform distribution, i.e., one that gives the same weight to all inputs appear to be a suitable choice in some applications [15]. Alternatively, a most favorable input distribution, $P(X)$, i.e., the one that maximizes information transfer, can be found. The maximization of mutual information with respect to the input distribution defines the information capacity, C^*. Formally,

$$C^* = \max_{P(X)} MI(X, Y). \tag{3}$$

Information capacity is expressed in bits and 2^{C^*} can be interpreted as the maximal number of inputs that the system can effectively resolve.

In summary, the overall fidelity of a signaling system depends on the degree of the overlap between distributions corresponding to different inputs. The degree of the overlap is translated into the logarithm of the number of resolvable inputs by mutual information, which takes into account how frequently different inputs are transmitted. On the other hand, the information capacity quantifies the logarithm of the number of resolvable inputs under input frequencies that maximize the information transfer. Information capacity, as opposed to mutual information, does not depend on input distribution and therefore may provide a less arbitrary quantification of the overall fidelity.

Existing tools to compute the mutual information and the information capacity [16, 19, 22, 27–29] utilise the data, y_l^i, to construct approximations, $\hat{P}(Y|X = x_i)$, of the output distributions, $P(Y|X = x_i)$, for each input, x_i. Thereafter, the approximations, $\hat{P}(Y|X = x_i)$, rather than the exact probabilities, $P(Y|X = x_i)$, are used for evaluation of the mutual information and information capacity, according to Eqs 2 and 3. The available algorithms differ in the way, in which, the approximations of the output distributions, $\hat{P}(y|X = x_i)$, are constructed. Specifically, Blahut—Arimoto (BA) algorithm [22, 27–29] uses a discrete approximation. All possible values of responses are divided into a finite set of intervals and frequencies of responses falling into the same interval as y_l^i are used as the approximation of $P(y_l^i|X = x_i)$. On the other hand,

methods based on the small noise approximation assume Gaussian output with a limited variance [16, 25, 30, 31]. Finally, the approach of [19], following the earlier work [32], uses the k-nearest neighbors (KNN) method, in which continuous approximations of $P(y_i^j|X = x_i)$ are constructed based on the distance of y_i^j to the k-th most similar response. Each of the above approaches is practically limited by the dimensionality of the output, Y. The BA algorithm can be essentially applied to systems with the one-dimensional output only. On the other hand, for multidimensional outputs, an accurate estimation of $P(Y|X = x_i)$ using KNN requires a relatively large sample size [33]. Moreover, KNN demands arbitrary specification of the parameter k, which for insufficient data size does not generally guarantee unbiased estimation [32–35], and yields estimation sensitive to algorithm's settings, i.e., is not parameter-free. Also, KNN based approaches, when used to compute capacity, often require solving computationally expensive optimization problems.

In summary, for multidimensional outputs, the practical difficulty in calculating mutual information, Eq 2, results largely from the lack of methods for accurate estimation of multivariate probability distributions, $P(y|X = x_i)$. In addition, calculation of C^*, Eq 3, can be problematic, as it requires maximization of the nonlinear function, i.e., *MI*, over the input probability distribution, which can be computationally intense. In Section 1 in S1 Text, we provide more background on information theory and existing computational tools. Here further, we propose an alternative framework, statistical learning estimation of mutual information (SLEMI) that appears to significantly simplify the calculation of the mutual information and the information capacity, especially for systems with high-dimensional outputs and a large number of input values. Moreover, our framework enables simple quantification of the extent to which different inputs can be discriminated.

Results

The framework to calculate the mutual information and the information capacity proposed here is based on an estimation of the conditional input distribution, $P(X|Y = y)$, as opposed to the output distributions $P(Y|X = x_i)$ in the existing approaches. Therefore, it bypasses the estimation of, possibly high dimensional, probability densities. Moreover, we show that the use of the conditional input distribution, $P(X|Y = y)$, can be combined with an efficient iterative optimization scheme to avoid, potentially problematic, numerical optimization. As a result, the introduced algorithm is computationally and statistically efficient and ensures robust, parameter-free estimation. Besides, we demonstrate that estimation of the conditional input distribution, $P(X|Y = y)$, provides a simple way to compute probabilities of correct discrimination between different inputs, which augments insight given by the information capacity. The critical advantage of SLEMI over available approaches is that it allows for efficient and robust analysis of systems with multidimensional outputs and a large number of inputs, which we demonstrate here using numerical test models. Further, we deploy SLEMI to examine single-cell signaling responses of the NF-κB pathway to TNF-α stimulation. Our analysis reveals that the NF-κB signaling dynamics improves discrimination of high concentrations of TNF-α with a modest impact on discrimination of low concentrations. A robust implementation of the proposed computational tool is provided as the R-package that can be used by computational biologists with only elementary knowledge of information theory.

Efficient estimation of the mutual information and the information capacity

In contrast to existing approaches, instead of estimating, possibly highly dimensional, conditional output distributions $P(Y|X = x_i)$, we propose to estimate the discrete, conditional input

distribution, $P(x_i|Y = y)$, which is known to be a simpler problem [36, 37]. Estimation of the *MI* using estimates of $P(x_i|Y = y)$, denoted here as $\hat{P}(x_i|Y = y)$, is possible as the *MI*, Eq 2, can be alternatively written as [38]

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

(4)

Although $P(Y|X = x_i)$ is still present in the above sum, it represents averaging of the term $\log_2 \frac{P(x_i|Y=y)}{P(x_i)}$ with respect to $P(Y|X = x_i)$. The experimental data, however, constitutes a sample from the distribution $P(Y|X = x_i)$. The average with respect to distribution $P(Y|X = x_i)$ can be, therefore, approximated by the average with respect to data, which is justified by the law of large numbers. Precisely, for a given $P(X)$ and $\hat{P}(x_i|Y = y)$, *MI* can be approximated with the following formula

$$MI(X, Y) \approx \sum_{i=1}^{m} P(x_i) \sum_{l=1}^{n_i} \frac{1}{n_i} \log_2 \frac{\hat{P}(x_i|Y = y_l^i)}{P(x_i)}.$$

(5)

An estimator $\hat{P}(x_i|Y = y)$, can be built using a variety of Bayesian statistical learning methods. For simplicity and efficiency, here we propose to use logistic regression, which is known to work well in a range of applications [39–43]. In principle, however, other classifiers could also be considered. The logistic regression estimators of $P(x_i|Y = y)$ arise from a simplifying assumption that log-ratio of probabilities, $P(x_i|Y = y)$ and $P(x_m|Y = y)$ is linear. Precisely,

$$\log \left(\frac{P(x_i|Y = y)}{P(x_m|Y = y)} \right) \approx \alpha_i + \beta_i^T y.$$

The above formulation allows fitting the logistic regression equations to experimental data, i.e., finding values of the parameters, α_i and β_i that best represent the data. Once logistic regression parameters are known, estimates $\hat{P}(x_i|Y = y)$ can be constructed. Estimates, $\hat{P}(x_i|Y = y)$, allow, in turn, to calculate mutual information using Eq 5. The use of logistic regression, therefore, constitutes a simple way to evaluate mutual information for multidimensional data without knowledge of $P(Y|X = x_i)$. Moreover, fitting the logistic regression equations to experimental data is efficient and available in most statistical software packages. Even though the assumed linear relationship may seem to be an oversimplification, the logistic regression approach has been shown to work exceptionally well in a variety of scenarios and gained broad applicability [36]. *Methods* contain more details on the form and estimation of the logistic regression model.

In addition to the possibility of effective evaluation of the mutual information for models with the multivariate output, Y, the use of the logistic regression enables to overcome the potentially problematic numerical, typically gradient, maximization of the mutual information with respect to the input distribution, $P(X)$, in computations of the information capacity. Precisely, the numerical optimization can be bypassed, by dividing the maximization with respect to the input distribution, $P(X)$, into two simpler maximization problems, for which explicit solutions exist. Thereafter, a solution of the joint maximization can be obtained from the two explicit solutions in an iterative procedure known as alternate maximization. Compared to gradient optimisation, alternate maximization is superior in terms of computational efficiency, and much less prone to numerical complications. A complete description of the maximization procedure is technically demanding and therefore is provided in *Methods* section. In particular, the algorithm's pseudo-code is presented in Box 1 therein.

In summary, the use of logistic regression described above allows computing *MI* without estimation of the possibly highly dimensional output distributions $P(Y|X = x_i)$. Moreover, it allows for efficient maximization of *MI* without gradient-based methods. In *Methods* and Section 3 in S1 Text, we perform several numerical tests to show how the above design of the algorithm leads to practical benefits in terms of the accuracy of estimation and computational efficiency. Specifically, we demonstrate that SLEMI: (i) provides more accurate estimates than the KNN method, especially for small sample size and highly dimensional output; (ii) delivers robust estimates, insensitive to algorithm's settings; and (iii) has desired properties in terms of computational cost, especially scales well with the number of input values.

Probabilities of correct discrimination

The information capacity, C^*, tells us how many inputs can be discriminated on average. What it does not tell us directly is which inputs, and to what extent, can be discriminated. Specifically, the same information capacity can result from different patterns of discriminability between input signals. For illustration, consider three inputs, x_1, x_2 and x_3. Assume that inputs x_1 and x_2 induce very similar output distributions, i.e., $P(Y|X = x_1) \approx P(Y|X = x_2)$, as opposed to x_3 that induces a distinct distribution, $P(Y|X = x_3)$. In such a scenario, the information capacity is approximately 1 bit, as two inputs can be discriminated, i.e., x_3 can be resolved from either of the other two. The capacity of 1 bit would also result from a scenario, in which roles of input values are swapped, say, $P(Y|X = x_2)$ is distinct from the overlapping $P(Y|X = x_1)$ and $P(Y|X = x_3)$. Therefore, it appears that an insight regarding which inputs, and to what extent, can be discriminated can usefully augment computation of the information capacity.

Here, we argue that the extent to which different inputs can be discriminated can be conveniently quantified and visualized using the probability of correct discrimination (PCD) between input pairs. We define the PCD between a pair of input values, x_i and x_j, as the fraction of cellular responses that can be assigned correctly to one of the two inputs based on the information contained in the signaling response, *Y*. If the distributions $P(Y|X = x_i)$ and $P(Y|X = x_j)$ are entirely distinct, knowing the response, *y*, allows assigning each cellular response to the correct input without error. PCD between x_i and x_j is then equal to 1. If, on the other hand, these two distributions are completely overlapping knowing the response, *y*, does not provide any information to assign a cell with a given response to the correct input. In such a case, the discrimination is close to random, yielding half of the cells being assigned correctly, i.e., PCD equals 0.5. Depending on the degree of the overlap, the PCD varies between 0.5 and 1. Further, calculation of PCDs for all input pairs can provide insight regarding which inputs, and to what extent, can be discriminated.

The above intuitions can be mathematically formalized, Fig 1. For formal quantification, in order to treat both inputs equally, we assume that both have the same frequency, $P(X) = (1/2, 1/2)$, or equivalently that half of the considered cells is stimulated with either of the input values, Fig 1A. How many cells can be assigned correctly depends on the overlap between the distributions $P(Y|X = x_i)$ and $P(Y|X = x_j)$, Fig 1B. The conditional input distribution, $P(x_i|Y = y)$ expresses the probability that a given response *y* was generated by the stimulation level x_i, Fig 1C. Equivalently, $P(x_j|Y = y)$ is the frequency, at which the given response, *y*, is generated by the stimulation level x_j. The probabilities $P(x_i|Y = y)$ and $P(x_j|Y = y)$ tell us, therefore, how often assignment of the observation *y* to the input x_i and x_j, respectively, is correct. To maximize the probability of correct assignment, the response *y* should be assigned to the input for which it is most likely, i.e. to x_i if $P(x_i|Y = y) \geq P(x_j|Y = y)$, or to x_j, otherwise. Therefore, the observation *y* can be assigned correctly with the probability equal to the maximum of $P(x_i|Y = y)$ and $P(x_j|Y = y)$. Precisely, the probability of correct discrimination between input x_i and x_j
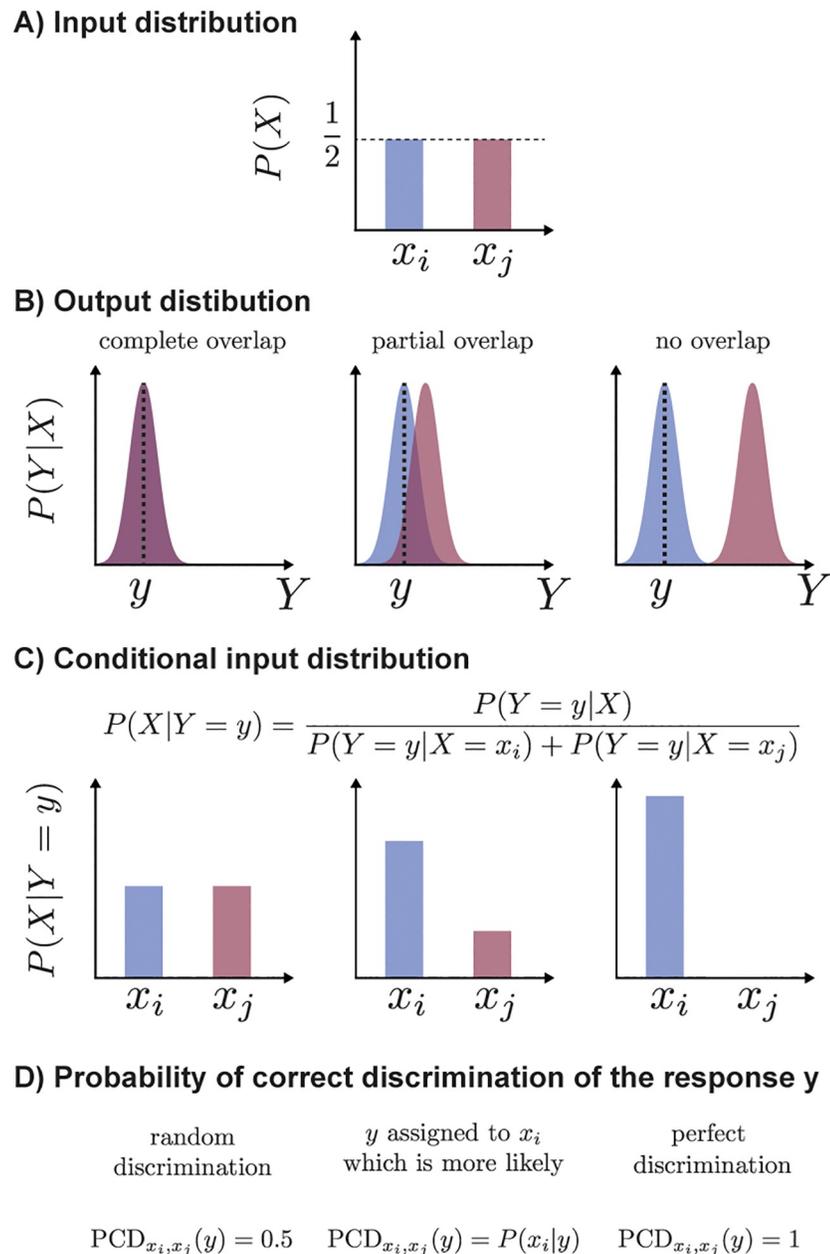
A) Input distribution

B) Output distibution

complete overlap   partial overlap   no overlap

C) Conditional input distribution

$$P(X|Y=y) = \frac{P(Y=y|X)}{P(Y=y|X=x_i) + P(Y=y|X=x_j)}$$

D) Probability of correct discrimination of the response y

random y assigned to x_i perfect
discrimination   which is more likely   discrimination

$$\text{PCD}_{x_i,x_j}(y) = 0.5 \qquad \text{PCD}_{x_i,x_j}(y) = P(x_i|y) \qquad \text{PCD}_{x_i,x_j}(y) = 1$$

Fig 1. Probabilities of correct discrimination between two inputs, x_i and x_j. The input distribution, $P(X) = (1/2, 1/2)$, visualized in (**A**), and the conditional output probabilities $P(Y|X)$, presented in (**B**), can be translated, via the Bayes formula, into conditional input distributions, $P(X|Y)$, visualized in (**C**). The conditional input distribution, $P(X|Y)$, serves to calculate the probability of correct discrimination of the observation y, as shown in (**D**). Precisely, for any fixed output, y, vertical line in (B), the conditional input probability, in (C), $P(X|Y = y)$, quantifies how likely it is that y was generated by each of the inputs. The probability of correct discrimination of the observation, y is given as the maximum of $P(x_i|Y = y)$ and $P(x_j|Y = y)$. Completely overlapping conditional output probabilities $P(Y|X)$, left column, yield random discrimination as opposed to non-overlapping distributions yielding perfect discrimination, right column. The use of the conditional input distribution, $P(X|Y)$, enables quantification of intermediate scenarios, middle column.

for the response y, denoted as $PCD_{x_i,x_j}(y)$, is calculated as

$$\text{PCD}_{x_i,x_j}(y) = \max\{P(x_i|Y=y), P(x_j|Y=y)\}, \tag{6}$$

which is visualised in Fig 1D. The average of the above probabilities over cellular responses, y_l^i, corresponding to the input x_i is equal to $\frac{1}{n_i}\sum_{l=1}^{n_i} PCD_{x_i,x_j}(y_l^i)$ and quantifies the average probability of correct discrimination of responses induced by the input x_i. Then, the overall probability of correct discrimination between x_i and x_j is given as

$$\text{PCD}_{x_i,x_j} = \frac{1}{2}\frac{1}{n_i}\sum_{l=1}^{n_i} PCD_{x_i,x_j}(y_l^i) + \frac{1}{2}\frac{1}{n_j}\sum_{l=1}^{n_j} PCD_{x_i,x_j}(y_l^j). \tag{7}$$

In summary, the probability of correct discrimination between inputs x_i and x_j, PCD_{x_i,x_j}, quantifies the fraction of cellular responses that can be correctly assigned to either of the inputs based on the output, Y. Therefore, the calculation of PCDs for all pairs of input values reveals the extent to which different inputs are discriminated.

From the computational perspective, PCDs are defined in terms of the conditional input probabilities, $P(x_i|Y=y)$, Therefore, similarly to the mutual information, these can be calculated using logistic regression. In *Methods* we provide practical details on how to compute PCDs. In the analysis of the NF-κB signaling data, presented below, we show how quantification of PCDs along with computation of the information capacity C^* helps to provide insight regarding how signaling dynamics increases overall signaling fidelity.

Signaling dynamics of the NF-κB system strongly improves discrimination of only high TNF-α concentrations

NF-κB pathway is one of the key biochemical circuits involved in the control of the immune system [44–46]. It is also one of the first cellular signaling systems studied within the framework of information theory [22]. So far, several papers examined its dose dependency, e.g., [12, 44, 47] and quantified its information capacity, e.g., [13, 19, 22]. Interestingly, response dynamics have been shown to have greater signaling capacity compared to time-point, non-dynamic, responses [13, 19]. To demonstrate what benefits result from efficient calculation of the information capacity and of the probabilities of correct discrimination, we have measured NF-κB responses (y_l^i's in our notation) to a range of 5 minutes pulses of TNF-α concentrations (x_i's), in single—cells, using life confocal imaging. Experimental methods are described in Sections 4.1—4.3 in S1 Text. Fig 2A shows temporally resolved responses to representative four concentrations, whereas Fig. IV, in S1 Text, to all ten considered concentrations. Further, we used the data to calculate the information capacity between TNF-α concentration and cellular response for two different scenarios: time-point and time-series. Precisely, for the time-point scenario, we considered single-cell measurements at each time-point separately. In this case, the signaling output of an individual cell at a given time, y, is represented by a single number, which is different for different time-points. On the other hand, for the time-series scenario, we considered single-cell measurements from the beginning of the experiment until an indicated time. In this case signaling output of an individual cell, y, is a vector corresponding to a time window from 0 till a given time. Fig 2B and 2C show information capacity as the function of time for the time-point and time-series scenario, respectively.

For the time-point scenario, the capacity increases at early times and reaches the maximum of ≈ 1 bit at ≈ 20 minutes, which coincides with the time of maximum response of trajectories shown in Fig 2A. Interestingly, the second peak of information transfer, of ≈ 0.3 bits, appears
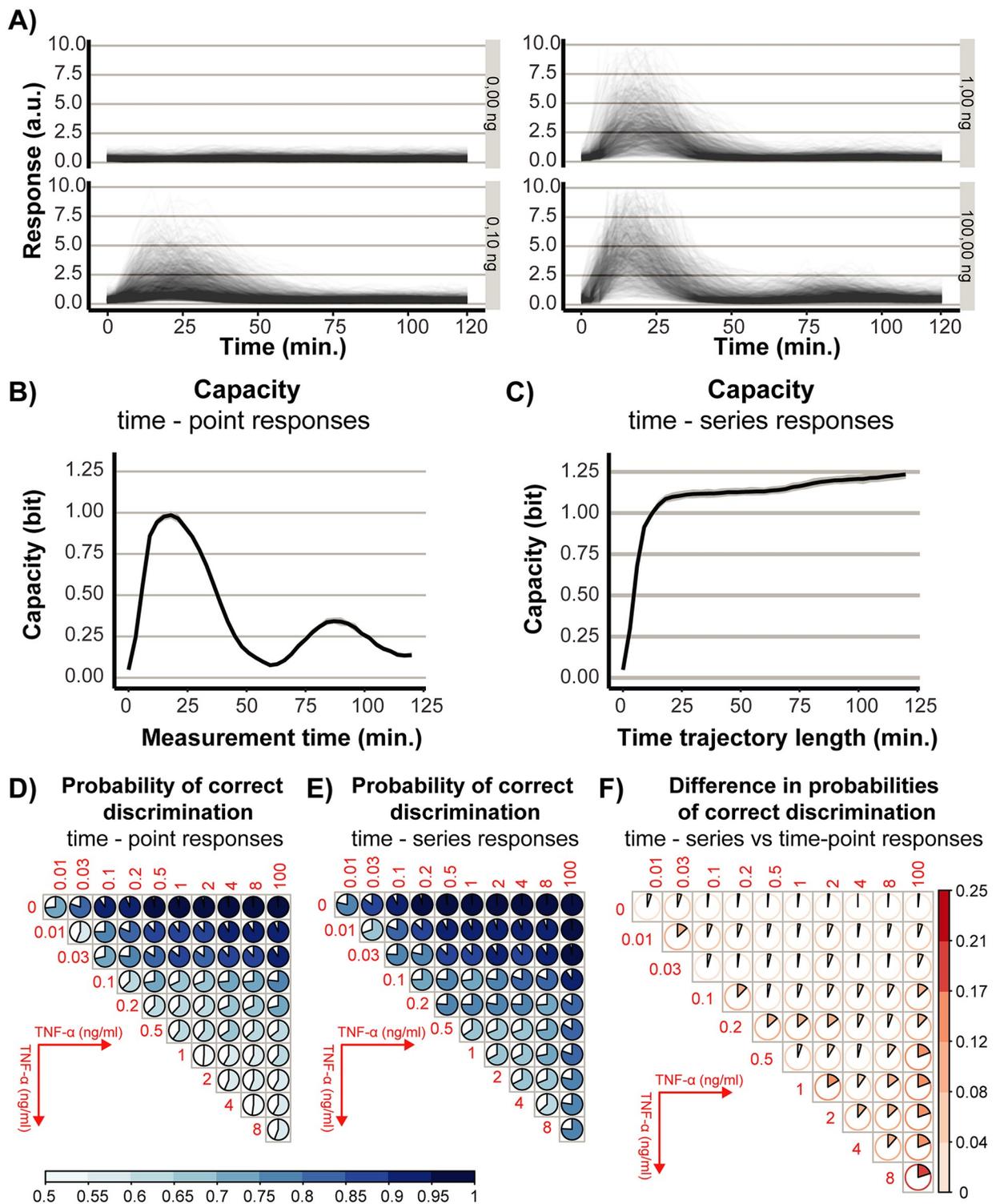
Fig 2. Information-theoretic analysis of the NF-κB responses to TNF-α stimulation. (A) Temporally resolved responses of individual cells to selected concentrations of TNF-α. The panel corresponds to Fig. IV in S1 Text. **(B)** The information capacity as a function of time for time-point responses. **(C)** As in (B) but for time-series responses. **(D)** Probabilities of the correct pairwise discrimination between TNF-α concentrations for time-point responses at 21 minutes. The color filled fraction of the circle marks the probability of correct discrimination. **(E)** The same as in (D) but for time-series responses. **(F)** Differences between probabilities in (D) and (E). Modeling details: Uncertainties of estimates (grey ribbons in B and C) were obtained by bootstrapping 80% of data (repeated 100 times). Probabilities in (D) and (E) present mean of 50 bootstrap re-sampling.

https://doi.org/10.1371/journal.pcbi.1007132.g002

at ≈ 90 minutes. Inspection of the response trajectories, Fig 2A, around 90 minutes allows for an interpretation of the second peak. Comparison of the response trajectories corresponding to 1 and 100 ng/ml of TNF-α, Fig 2A, indicates the emergence of a fraction of cells that exhibit a second peak in response to the highest considered concentration. The second peak is reminiscent of the oscillatory behavior that is typical for the NF-κB pathway when exposed to continuous, as opposed to 5 minutes, stimulation [44, 48–50]. The second peak in response trajectories carries some information about TNF-α and, therefore, contributes to the second peak of information transfer.

For the time-series scenario the capacity also rapidly increases at early times to reach 1 bit at 20 minutes. For later times, the capacity continues to increase but at a much slower rate, with a modest acceleration around 70 minutes, to reach ≈ 1.3 bits at the end of the experiment. As the information contained in shorter time-series is contained in longer time-series, the capacity does not decrease. An increase in the capacity in a given time interval indicates that new information is arriving. Analogously, a time interval with a plateau demonstrates the lack of new information being transmitted at times of that interval. Therefore, our analysis demonstrates that most of the information, ≈ 1 bit, is transferred relatively early, i.e., within the first 20 minutes after stimulation. Later times provide ≈ 0.3 bits of new information. In Section 4.5 in S1 Text we use the method proposed in [51] to further examine the redundancy of information contained in responses at individual time-points.

The higher information content of the time-series poses the question: what type of information is contained in the time-series responses that is not encapsulated in the time-point responses? The information capacity *per se*, being an overall measure of signaling fidelity, does not tell us to what extent specific inputs can be discriminated. Therefore, the information capacity alone cannot reveal which inputs gain discriminability due to signaling dynamics. In order to address the above question in detail, we have calculated the probabilities of correct discrimination, PCDs, for each pair of concentrations, x_i, x_j. Analogously as in the computation of the capacity, we have considered the time-point and time-series scenarios. PCDs for the two scenarios are shown in Fig 2D and 2E. The filled fractions of the pies mark the PCDs between the corresponding pairs of concentrations, i.e., the fractions of cells that can be assigned to the correct input. The plots, primarily reveal how the time-point and time-series capacities of 1 and 1.3 bits, respectively, translate into discrimination between pairs of inputs. They show that concentrations far apart can be easily discriminated in both scenarios. For instance, PCDs between 0 and 100 ng/ml are close to 1. The discrimination of closer concentrations is more difficult. For instance, in both scenarios, PCDs between 0 and 0.01 ng/ml are approximately 0.75, which corresponds to 0.25 probability of incorrect assignment.

Interestingly, for time-point responses, PCDs between pairs of concentrations ≥ 1 ng/ml are close to 0.5, which implies lack of discriminability. This is, however, not the case for time-series responses where PCDs are considerably greater than 0.5. Comparison of PCDs between the two scenarios reveals, therefore, which states gain discriminability due to signaling dynamics. Fig 2F presents differences between PCDs for time-series and time-points. The increase in discriminability resulting from signaling dynamics is particularly striking for high concentrations, say > 1 ng/ml. These concentrations are only weakly discriminated for the time-point scenario. Certain low concentrations also gain discriminability, e.g., 0.01 and 0.03. However, overall, the increase in discriminability is not so significant for low concentrations as these are relatively well discriminated in the time-point scenario.

The above analysis yields similar conclusions to these presented in [19]. Here, we used 5 minutes TNF-α as opposed to continuous lipopolysacharide stimulation in [19]. Also, we used a more complete, higher dimensional, response trajectories, which allowed to plot the temporal profile of information transfer. Our analysis of PCDs complemented calculation of the

capacity by revealing discriminability between different inputs. Indeed, values far apart are discriminated virtually without error. Closer concentration gain discriminability due to dynamics of signaling, whereas the gain is particularly strong for high concentrations.

R-package

Our algorithm is available as robustly implemented R-package, SLEMI. It is designed to be used by computational biologists with a limited background in information theory. It includes functions to calculate mutual information, information capacity and probabilities of correct discrimination. These functions take as the argument a data frame, `data`, containing signaling responses stored in the following form

$$
\begin{array}{c|cccc}
\text{input} & \text{output 1} & \text{output 2} & \text{output 3} & \ldots \\
\hline
n_1 \begin{cases} x_1 \\ \vdots \\ x_1 \end{cases} & \begin{array}{c} y^1_{1,1} \\ \vdots \\ y^1_{n_1,1} \end{array} & \begin{array}{c} y^1_{1,2} \\ \vdots \\ y^1_{n_1,2} \end{array} & \begin{array}{c} y^1_{1,3} \\ \vdots \\ y^1_{n_1,3} \end{array} & \\
n_2 \begin{cases} x_2 \\ \vdots \\ x_2 \end{cases} & \begin{array}{c} y^2_{1,1} \\ \vdots \\ y^2_{n_2,1} \end{array} & \begin{array}{c} y^2_{1,2} \\ \vdots \\ y^2_{n_2,2} \end{array} & \begin{array}{c} y^2_{1,3} \\ \vdots \\ y^2_{n_2,3} \end{array} & \ldots \\
\vdots & \vdots & \vdots & \vdots \\
n_m \begin{cases} x_m \\ \vdots \\ x_m \end{cases} & \begin{array}{c} y^m_{1,1} \\ \vdots \\ y^m_{n_m,1} \end{array} & \begin{array}{c} y^m_{1,2} \\ \vdots \\ y^m_{n_m,2} \end{array} & \begin{array}{c} y^m_{1,3} \\ \vdots \\ y^m_{n_m,3} \end{array} &
\end{array} \tag{8}
$$

Each row, l, represents a single cell. The first column contains stimulation levels, x_i. Further columns contain entries corresponding to measurements of cellular output, $y^i_{l,d}$, e.g., subsequent elements of a time-series.

Upon download with the function `install_github()` of the 'devtools' package
`install_github("sysbiosig/SLEMI")`
the considered information-theoretic measures can be calculated by running
`mi_logreg_main(data)`
for *MI*, with uniformly distributed inputs;
`capacity_logreg_main(data)`
for information capacity C^*; and
`prob_discr_pairwise(data)`
for probabilities of correct discrimination.

More details on installation and applicability are provided in the package's *User Manual*. A step-by-step *Testing Procedures* file is also available to assist with running essential functions. The package includes the NF-κB dataset as well as scripts to reproduce Fig 2. Computations needed to plot each panel of the figure, without bootstrap, do not exceed several minutes on a regular laptop.

Discussion

Building upon existing approaches, our framework considerably simplifies information-theoretic analysis of multivariate single-cell signaling data. It benefits from a novel algorithm,

which is based on the estimation of the discrete input distribution as opposed to the estimation of continuous output distributions. Conveniently, the algorithm does not involve numerical gradient optimization. These factors result not only in short computational times but, also, in relatively low sample sizes needed to obtain accurate estimates. Therefore, our framework is particularly suitable to study systems with high dimensional outputs and a large number of input values. Also, the approach relates the information capacity to the probability of discrimination between different input values.

Information theory seems to offer useful tools to provide a better understanding of how cells transmit information about identity and quantity of stimuli, and further how signaling systems enable cells to perform complex functions. Such tools might have a fair potential to successfully augment more traditional approaches. The latter provided a relatively good understooding of the overall molecular and biochemical mechanisms how individual cells transmit signals to effectors [1]. However, we lack understanding of how the stimuli are translated into distinct responses and, hence, how to effectively control cellular processes and decisions [1, 3, 15]. Specifically, the induction of distinct responses in individual cells by means of biochemical interventions in non-trivial settings is most often problematic [52–54]. Results of our work appear to contribute a relevant tool to apply information-theoretic analysis to more complex, particularly highly multivariate, data sets on signaling systems than achievable with available approaches. The multivariate aspect seems to be particularly relevant given the complexity of cellular processes. Hopefully, current and future development of multivariate single-cell measurement techniques, accompanied by computational tools, will enable utilization of the information-theoretic perspective in more complex scenarios. These in turn appear to have a potential to provide a comprehensive insight into how cells can derive a variety of distinct outputs from complex inputs using noisy, cross-wired and dynamic signaling pathways.

Methods

Logistic regression

The logistic regression model is the state-of-the-art statistical method to estimate the probability of a given observation, i.e., data vector, y, belonging to one of the m considered classes. In the setting of the paper, classes correspond to input values and observations to signaling responses. The method assumes that the overall frequencies of observations belonging to each class are described by a probability distribution. In the paper's setting, these probabilities correspond to the input distribution $P(X) = (P(x_1), \ldots, P(x_m))$.

The method is based on the assumption that for a given $P(X)$, the ratio of the probability of the observation y belonging to class i to the same probability for the class m is linear with respect to y. Precisely, denoting the logistic regression estimate of the probability of a given observation, y, belonging to the class i as $\hat{P}_{lr}(x_i | Y = y; P(X))$, the above assumption writes as follows

$$\log\left(\frac{\hat{P}_{lr}(x_1 | Y = y; P(X))}{\hat{P}_{lr}(x_m | Y = y; P(X))}\right) \approx \alpha_1 + \beta_1^T y,$$

$$\vdots$$

$$\log\left(\frac{\hat{P}_{lr}(x_i | Y = y; P(X))}{\hat{P}_{lr}(x_m | Y = y; P(X))}\right) \approx \alpha_i + \beta_i^T y, \tag{9}$$

$$\vdots$$

$$\log\left(\frac{\hat{P}_{lr}(x_{m-1} | Y = y; P(X))}{\hat{P}_{lr}(x_m | Y = y; P(X))}\right) \approx \alpha_{m-1} + \beta_{m-1}^T y$$

and $\sum_{i=1}^{m} \hat{P}_{lr}(x_i | Y = y; P(X)) = 1$. Given the linear form of the above equations, for a data set given as Eq 8, estimation of the parameters α_i and β_i can be done efficiently with state-of-the-art methods [36, 55]. Further, the above equations imply that estimates, $\hat{P}_{lr}(x_i | Y = y; P(X))$, can be explicitly written as

$$
\begin{aligned}
\hat{P}_{lr}(x_1 | Y = y; P(X)) &= \frac{\exp\left(\alpha_1 + \beta_1^T y\right)}{1 + \sum_{r=1}^{m-1} \exp\left(\alpha_r + \beta_r^T y\right)}, \\
&\;\;\vdots \\
\hat{P}_{lr}(x_i | Y = y; P(X)) &= \frac{\exp\left(\alpha_i + \beta_i^T y\right)}{1 + \sum_{r=1}^{m-1} \exp\left(\alpha_r + \beta_r^T y\right)}, \\
&\;\;\vdots \\
\hat{P}_{lr}(x_{m-1} | Y = y; P(X)) &= \frac{\exp\left(\alpha_{m-1} + \beta_{m-1}^T y\right)}{1 + \sum_{r=1}^{m-1} \exp\left(\alpha_r + \beta_r^T y\right)}, \\
\hat{P}_{lr}(x_m | Y = y; P(X)) &= \frac{1}{1 + \sum_{r=1}^{m-1} \exp\left(\alpha_r + \beta_r^T y\right)}.
\end{aligned}
\tag{10}
$$

Maximization algorithm to compute capacity

Below, we describe maximization of the mutual estimation, MI, with respect to the input distribution $P(X)$, using the so-called alternate optimization, which bypasses gradient optimization. The proposed algorithm is largely based on the original Blahut-Arimoto approach [27, 28]. It is adapted to work with logistic regression and, hence, with continuous and multidimensional output, Y. The provided Lemmas are minor modifications of the original Blahut-Arimoto results to account for continuous and multidimensional output.

The algorithm is based on the following five key components. One, the maximization of mutual information with respect to input distribution, $P(X)$, is replaced with a double maximization, i.e., maximization with respect to the input distribution, $P(X)$, and with respect to tailored auxiliary function, $Q(X|Y)$. The function $Q(X|Y)$ is introduced to dissect the effects of the input distribution, $P(X)$, and the conditional input distribution, $P(Y|X)$, on the mutual information, MI. Two, explicit solutions of the individual maximizations are found. Three, the individual maximizations are combined in an iterative procedure to provide the solution of the joint maximization. Four, integrals involved in the evaluation of the optimal solutions of individual maximizations are computed through averaging with respect to data. Five, logistic regression is used to evaluate optimal $Q(X|Y)$ at each step of the iterative procedure. Each of the above five elements is described in detail below, and the complete algorithm is summarized in Box 1.

Information capacity as a double maximization problem. The first element of the proposed approach involves replacing the maximization in the capacity definition, Eq 3, with the double maximization. Precisely, it can be shown (Lemma 1 in S1 Text) that in the setting of Eq 1, the capacity, C^*, Eq 3, can be written as

$$
C^* = \max_{P(X)} \max_{Q(X|Y)} J(P(X), Q(X|Y)),
\tag{11}
$$

where $Q(X|Y)$, for a given $Y = y$, is a discrete probability distribution with respect to X, whereas

J is an auxiliary function defined as

$$J(P(X), Q(X|Y)) = \sum_{i=1}^{m} P(x_i) \int_{\mathbb{R}^d} P(y|X = x_i) \ \log_2 \frac{Q(x_i|y)}{P(x_i)} dy. \tag{12}$$

The function J is introduced for technical convenience and was obtained from the mutual information, Eq 4, by replacing $P(X|Y)$ with $Q(X|Y)$. To prove Lemma 1 in S1 Text, we also shows that for $Q(X|Y) = P(X|Y)$, the function J is maximised and equal to the mutual information. It is a generalisation of Theorem 1a of [27] to work for systems with continuous and multidimensional output, Y.

Individual maxima can be found explicitly.   Compared to the single maximization problem with respect to $P(X)$ in Eq 3, the double maximization in Eq 11 with respect to $P(X)$ and $Q(X|Y)$ has the advantage that the optimal solutions of the individual optimisation problems can be found analytically.

Precisely, the individual maximizations

$$P^*(X; Q(X|Y)) = \arg \max_{P(X)} J(P(X), Q(X|Y)), \tag{13}$$

$$Q^*(X|Y; P(X)) = \arg \max_{Q(X|Y)} J(P(X), Q(X|Y)) \tag{14}$$

have explicit solutions. The solution of the maximization in Eq 13 can be found using Lagrange multipliers. Lemma 2 in S1 Text shows that for a given $Q(X|Y)$ the optimal value of

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

is achieved by

$$P^*(x_i; Q(X|Y)) = \frac{\exp(D_i(Q))}{\sum_{r=1}^{m} \exp(D_r(Q))}, \tag{16}$$

where

$$D_i(Q) = \int_{\mathbb{R}^d} P(y|X = x_i) \ \log_2 Q(x_i|y) dy. \tag{17}$$

The above is similar to the solution presented in the conventional BA algorithm [27] but accounts for continuous output Y.

Solution of the maximization in Eq 14 is a well established result provided in [27] as Theorem 1, which is used here in an unchanged form. Precisely, for a given $P(X)$, Eq 14 has the explicit solution

$$Q^*(x_i|y; P(X)) = \frac{P(x_i)P(y|X = x_i)}{\sum_{r=1}^{m} P(x_r)P(y|X = x_r)}. \tag{18}$$

Alternate maximization.   Further, similarly as in the BA approach, we propose to combine the above solutions, i.e., Eqs 16 and 18, in the so called alternate maximization (AM) iterative procedure to deliver the solution of the joint maximization of Eq 11. Precisely, in an initial step, arbitrary $P(X)$ and $Q(X|Y)$, denoted here 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. Precisely, $P^{(k)}(x_i)$ and $Q^{(k)}(x_i|y)$ are set to optimal solutions of the

individual maximization problems of Eqs 13 and 14 given previous values $P^{(k-1)}(x_i)$, $Q^{(k-1)}(x_i|y)$, i.e.,

$$P^{(k)}(x_i) = P^*(x_i; Q^{(k-1)}(x_i|y)), \tag{19}$$

$$Q^{(k)}(x_i|y) = Q^*(x_i|y; P^{(k)}(x_i)). \tag{20}$$

By Lemma 2 the capacity, C^*, is then approximated as

$$C^{(k)} = J(P^{(k)}(X), Q^{(k)}(X|Y)) = \sum_{i=1}^{m} P^{(k)}(x_i)(D_i(Q^{(k)}) - \log_2 P^{(k)}(x_i)). \tag{21}$$

In Lemma 4, we generalise the result of [28], and we show that for systems with continuous output Y, the iterative scheme converges to the solution of the joint maximization problem, Eq 11, precisely

$$P^{(k)}(x_i) \xrightarrow{k\to\infty} P^*(x_i)$$

and

$$C^{(k)} \xrightarrow{k\to\infty} C^*.$$

Integration through averaging. For a practical implementation of the above AM method, solutions of the individual maximizations problems need to be found for a given data set. Precisely, Eqs 12, 16 and 18 need to be numerically evaluated, which is seemingly problematic, as these equations depend on the unknown conditional probabilities $P(Y|X)$. Therefore, we propose a strategy to evaluate Eqs 12, 16 and 18 without knowledge of $P(Y|X)$. Primarily, we show below that integration with respect to y, in Eqs 12 and 17, can be performed without knowledge of $P(y|X = x_i)$. Thereafter, we demonstrate how $Q^{(k)}(x_i|y)$ can be computed.

Eqs 12 and 17 involve 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, that 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 computed based on the sample approximates the expectation, leading to

$$\int_{\mathbb{R}^d} P(y|X = x_i) \, \log_2 \frac{Q(x_i|y)}{P(x_i)} \, dy \quad = \quad \mathbb{E}_{P(Y|X=x_i)}\left(\log_2 \frac{Q(x_i|Y)}{P(x_i)}\right) \approx \frac{1}{n_i} \sum_{l=1}^{n_i} \log_2 \frac{Q(x_i|y_l^i)}{P(x_i)}, \tag{22}$$

$$\int_{\mathbb{R}^d} P(y|X = x_i) \, \log_2 Q(x_i|y) dy \quad = \quad \mathbb{E}_{P(Y|X=x_i)} \log_2 Q(x_i|Y) \approx \frac{1}{n_i} \sum_{l=1}^{n_i} \log_2 Q(x_i|y_l^i). \tag{23}$$

The above shows, that the integration of Eqs 12 and 17, indeed does not require explicit

knowledge of $P(y|X = x_i)$. Below, we show that computation of $Q^*(x_i|y; P(X))$ in Eq 18 does not require evaluation of $P(y|X = x_i)$ either.

Incorporation of logistic regression. The Bayes conditional probability formula

$$P(x_i|Y = y) = \frac{P(x_i)P(y|X = x_i)}{\sum_{r=1}^{m} P(x_r)P(y|X = x_r)}$$

implies that, for a given $P(X)$, $Q^*(x_i|y; P(X))$ defined by Eq 18 is equivalent to $P(x_i|Y = y)$, i.e.,

$$Q^*(x_i|y; P(X)) = P(x_i|Y = y). \tag{24}$$

Therefore, for a given $P(X)$ finding $Q^*(x_i, y; P(X))$ is equivalent to finding $P(x_i|Y = y)$ and, therefore, $P(y|X = x_i)$ is not required. As discussed earlier, approximation of $P(x_i|Y = y)$ is a classification problem in the statistical learning theory [36] and here we propose to approximate $P(x_i|Y = y)$ and, hence, also $Q^*(x_i|y; P(X))$ using logistic regression. Precisely,

$$Q^*(x_i|y; P(X)) \approx \hat{P}_{lr}(x_i|Y = y; P(X)), \tag{25}$$

where $\hat{P}_{lr}(x_i|Y = y; P(X))$ denotes the logistic regression model, Eq 10, for classifying x_i based on responses, y, and the input distribution, $P(X)$. Moreover, in Lemma 6 in S1 Text we show that the parameters of logistic regression have to be estimated only once. Precisely, change of $P(X)$ from $P^{(k-1)}(X)$ to $P^{(k)}(X)$ at each step of the iterative procedure requires only an update of the intercept parameters α_i, for i from 1 to $m - 1$, according to the following formula

$$\alpha_i^{(k)} = \alpha_i^{(k-1)} + \log\left(\frac{P^{(k-1)}(x_m)}{P^{(k-1)}(x_i)}\right) - \log\left(\frac{P^{(k)}(x_m)}{P^{(k)}(x_i)}\right), \tag{26}$$

where $\alpha_i^{(k)}$ is the intercept parameter of the logistic regression model, Eq 10, at k-th iteration of the alternate maximization procedure. Parameters, β_i, for $i = 1$ to $m - 1$, remain the same for all iterations.

Pseudo-code. The above five elements are combined into the algorithm to compute information capacity, C^* in the following way. Primarily, the initial values of the input distribution, denoted as $P^{(0)}(X)$, are set, e.g., to be equal to relative frequencies of measurements available for each input, $P^{(0)}(x_i) = \frac{n_i}{\sum_{r=1}^{m} n_r}$. Thereafter, logistic regression model is constructed to obtain initial value of the function $Q()$, denoted as $Q^{(0)}$, i.e., $Q^{(0)}(x_i|y; P^{(0)}(X)) = \hat{P}_{lr}(x_i|Y = y; P^{(0)}(X))$. Thereafter, at each iteration, k, distribution $P(X)$, and function $Q(X|Y)$ are updated, as well as the approximation of the capacity provided by the function $J()$ is evaluated according to Eq 12. Iterations are repeated until convergence is reached. The algorithm is summarized as a pseudo-code in Box 1.

Practical calculation of the probabilities of correct discrimination

Probabilities of correct discrimination, PCDs defined in Eqs 6 and 7, are expressed in terms of probabilities $P(x_i|Y = y)$. Logistic regression model, Eq 10, on the other hand, provides estimates of these probabilities. Therefore, logistic regression estimates, $\hat{P}_{lr}(x_i|Y = y; P(X))$, can be used to estimate probabilities of correct discrimination.

In order to estimate PCDs, for a given pair of input values x_i and x_j, the logistic regression model needs to be fitted using response data corresponding to the two considered inputs, i.e. y_l^r, for $r \in \{i, j\}$ and l ranging from 1 to n_r. To ensure that both inputs have equal contribution to the calculated discriminability, equal probabilities should be assigned, $P(X) = (P(x_i), P(x_j)) = (1/2, 1/2)$. Once the regression model is fitted, probability of assigning a given cellular

Box 1 Algorithm to calculate channel capacity using statistical learning

1: Set maximum number of iterations `MAXIT` and tolerance level `tol`

2: Set $k = 1$

3: Initialise $C^{(k)}$, e.g. $C^{(-1)} = -\infty$, $C^{(0)} = 0$

4: Set initial distribution of $P^{(0)}(X)$: $P^{(0)}(x_i) = \dfrac{n_i}{\sum_{r=1}^{m} n_r}$

5: 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

6: **while** $k \leq$ `MAXIT` AND $|C^{(k-1)} - C^{(k-2)}| >$ `tol` **do**

7:    Calculate $D_i(Q^{(k-1)})$ by Monte Carlo integration

$$D_i(Q^{(k-1)}) = \mathbb{E}_{P(Y|X=x_i)}\left(\log_2 Q^{(k-1)}(x_i|y)\right) \approx \frac{1}{n_i}\sum_{l=1}^{n_i}\log_2 Q^{(k-1)}(x_i|y_l^i),$$

8:    Optimize $\max_{P(X)} J(P, Q^{(k-1)})$

$$P^{(k)}(x_i) = P^*(x_i; Q^{(k-1)}(x_i|y)) = \frac{\exp(D_i(Q^{(k-1)}))}{\sum_{r=1}^{m}\exp(D_r(Q^{(k-1)}))}$$

9:    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 26

10:    Get, $C^{(k)}$, an estimate of C^*

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

11:    $k = k + 1$

12: **end while**

13: **return** $C^* = C^{(k)}$, $P^*(X) = P^{(k)}(X)$.

response, y, to the correct input value is estimated as

$$\max\{\hat{P}_{lr}(x_i|Y = y; P(X)), \hat{P}_{lr}(x_j|Y = y; P(X))\}.$$

Note that $P(x_j|Y = y) = 1 - P(x_i|Y = y)$ as well as $\hat{P}_{lr}(x_j|Y = y; P(X)) = 1 - \hat{P}_{lr}(x_i|Y = y; P(X))$. Averaging the above over all responses corresponding to input values x_i and x_j, i.e., with respect to the distribution $P(y) = \frac{1}{2}P(y|X = x_i) + \frac{1}{2}P(y|X = x_j)$, yields $PCD(x_i, x_j)$

$$
\begin{aligned}
PCD(x_i, x_j) \quad &\approx \quad \frac{1}{2}\frac{1}{n_i}\sum_{l=1}^{n_i}\max\{\hat{P}_{lr}(x_i|Y = y_l^i; P(X)), \hat{P}_{lr}(x_j|Y = y_l^i; P(X))\} \\
&+ \quad \frac{1}{2}\frac{1}{n_j}\sum_{l=1}^{n_j}\max\{\hat{P}_{lr}(x_i|Y = y_l^j; P(X)), \hat{P}_{lr}(x_j|Y = y_l^j; P(X))\}.
\end{aligned}
\tag{27}
$$

To account for the possibility of overfitting of the regression model, the bootstrap procedure needs to be used [36]. For instance, available data should be randomly divided into a training data set, i.e., data set used to fit logistic regression, and test data set, i.e., the set to evaluate Eq 27. The average of the PCDs from the bootstrap procedure should be used as a final estimate of the probability of correct discrimination.

Numerical validation

In order to validate the accuracy of the proposed information capacity estimators, examine the computational performance of the algorithm, and highlight advantages of the introduced approach, we have designed four test scenarios and carried out a comparison with the KNN method. We have chosen KNN for the comparison as it is virtually the only available technique that allows estimating MI and C^* for systems with multidimensional output, Y. One of the test scenarios, Scenario 1, is presented below whereas the remaining three scenarios, Scenarios 2-4, are part of S1 Text. Scenario 1 demonstrates that the proposed approach, here further referred to as SLEMI, (i) provides more accurate estimates than the KNN method; (ii) provides robust, parameter-free estimates; and (iii) has desired properties in terms of the computational cost. Scenario 2 replicates diagnostics proposed for the KNN methods in reference [19] and demonstrates that, in contrast to the KNN method, accuracy of SLEMI estimates persists for high dimensionality of output data. Scenario 3 validates the accuracy of SLEMI estimates against several different shapes of the output distribution. Finally, Scenario 4 uses a model of a transcription factor activity [20] to demonstrate that SLEMI can be used to quantify the information capacity of frequency encoded signals.

The test model used as Scenario 1 aims to reflect a simple 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,

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

for $\mu(x) = \frac{V \cdot x}{1+x}$, $V = 10$, $\sigma^2 = 1$. For the above model, we considered eleven input values, $m = 11$. Input vales range from 0 to 100, $x_i \in [0, 100]$. Sample distributions of output corresponding to considered input values are shown in Fig 3A.

To test the estimation accuracy, we computed information capacity estimates using SLEMI and KNN method for different sample size, N, i.e., the number of data points corresponding to each input value used for estimation. True information capacity was evaluated numerically. Fig 3B presents both information capacity estimates, as well as the true value, as the function of N, for N ranging from 50 up to 2000. For sample sizes typical for biological experiments, i.e., tens or hundreds of measured cells, SLEMI clearly provides more accurate estimates than KNN.

In contrast to SLEMI, KNN estimates depend on the choice of the parameter k, whereas clear rules how k should be selected are missing [56–58]. For computation of the KNN estimates in Fig 3A we used $k = 10$, i.e., 10 closest points to each observation y were used to approximate the density $P(y|X = x_i)$. To highlight the impact of the parameter k, we re-calculated the KNN estimates of Fig 3B using k ranging from 2 to 50, at fixed $N = 1000$. As show in Fig 3C, selection of k has an impact on the value of the capacity estimates that may lead to considerable bias. SLEMI estimates are free from this disadvantage.

Further, we have examined how the computational time needed to obtain estimates scales with sample size, N, and the number of input values, m. Fig 3D presents computational times
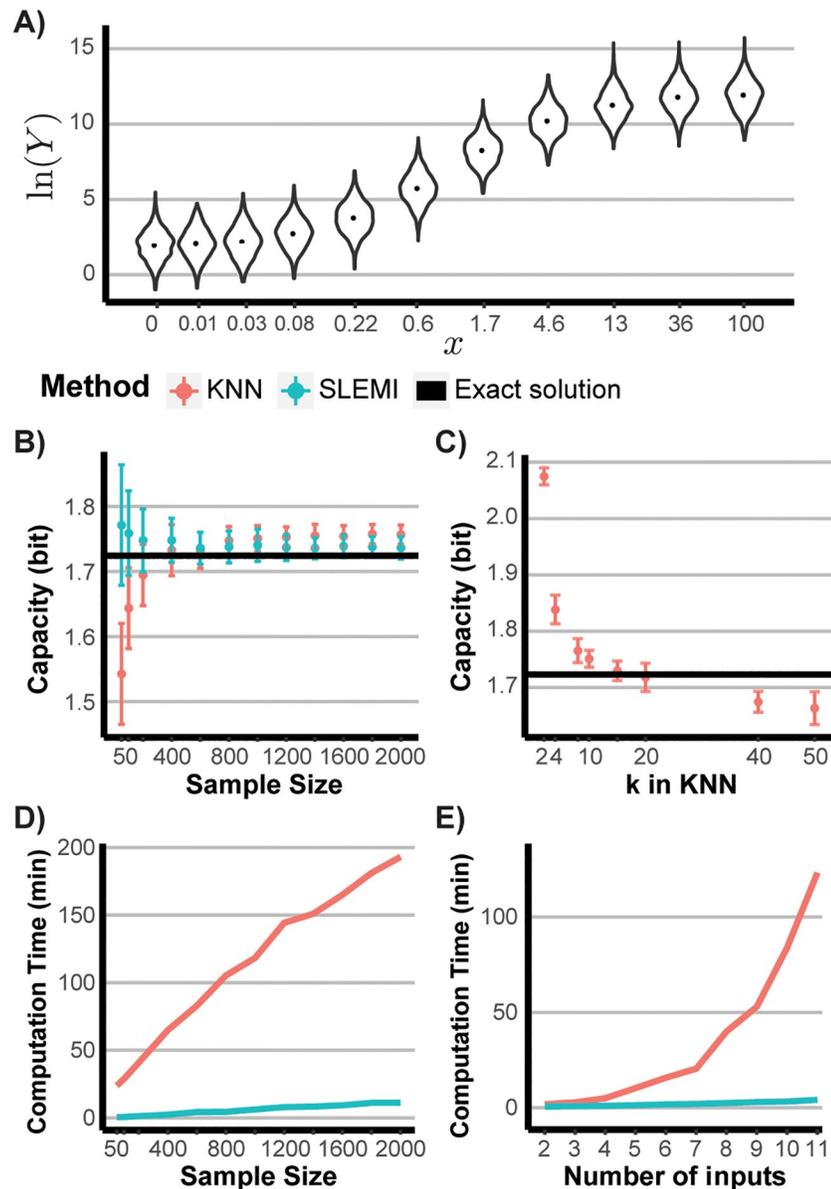
Fig 3. Test scenario 1. (A) A violin plot representation of the conditional output distribution $Y|x_i$ for 11 considered inputs. **(B)** Information capacity estimates as the function of the sample size N. Blue and red lines correspond to SLEMI and KNN estimates, respectively. The bold black line marks the true value of the 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 SLEMI and KNN method as the function of the sample size N. **(E)** Computation time of SLEMI (blue) and the KNN method as the 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.

https://doi.org/10.1371/journal.pcbi.1007132.g003

as a function on N. Both methods exhibit linear increase with N (Fig 3D). Computational time of SLEMI increases at the lower rate. Although we optimized implementations of both methods to ensure a fair comparison, the rate of increase may depend on specifics of the code used. Finally, we calculated the computational time as a function of the number of input values.

Initially, we considered two input values, $m = 2$, and increased, one by one, up to eleven input values, $m = 11$. The computational time of SLEMI increases linearly, whereas computational time corresponding to KNN method increases at least quadratically (Fig 3E). Linear scaling with respect to the number of input values is important for the method to be applicable to study complex systems with multiple inputs.

In summary, the above test scenario shows that SLEMI provides more accurate estimates of the information capacity, C^*, than the KNN method, especially for small sample size. Importantly, estimates are parameter-free, which contributes to robust estimation. Also, computational time scales linearly with respect to the number of input values. The test scenario presented above considered one-dimensional output, Y, as the demonstration of key benefits resulting from using SLEMI did not require a multidimensional complex model. The presented advantages, however, are of particular importance when systems with multivariate output, Y, are studied. For multivariate systems robust estimation of the information capacity, C^*, with KNN method can be problematic due to the choice of the parameter k and numerical optimization. Therefore, in the test Scenario 2, in S1 Text, we have confirmed that, unlike for the KNN method, high accuracy of information capacity estimates persist for multidimensional data.

Supporting information

S1 Text. Supplementary information PDF file contains expanded description of theoretical and experimental methods.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Tomasz Jetka, Karol Nienałtowski, Michał Komorowski.

Data curation: Tomasz Jetka.

Formal analysis: Tomasz Jetka, Karol Nienałtowski, Michał Komorowski.

Funding acquisition: Michał Komorowski.

Investigation: Tomasz Jetka, Karol Nienałtowski, Tomasz Winarski, Michał Komorowski.

Methodology: Tomasz Jetka, Tomasz Winarski, Sławomir Błoński, Michał Komorowski.

Project administration: Michał Komorowski.

Resources: Tomasz Jetka, Tomasz Winarski, Sławomir Błoński, Michał Komorowski.

Software: Tomasz Jetka, Karol Nienałtowski.

Supervision: Sławomir Błoński, Michał Komorowski.

Validation: Tomasz Jetka, Karol Nienałtowski.

Visualization: Tomasz Jetka, Karol Nienałtowski, Michał Komorowski.

Writing – original draft: Tomasz Jetka, Michał Komorowski.

Writing – review & editing: Tomasz Jetka, Karol Nienałtowski, Michał Komorowski.

References

1. Nurse P. Life, logic and information. Nature. 2008; 454(7203):424. https://doi.org/10.1038/454424a PMID: 18650911

2. Purvis JE, Lahav G. Encoding and decoding cellular information through signaling dynamics. Cell. 2013; 152(5):945–956. https://doi.org/10.1016/j.cell.2013.02.005 PMID: 23452846

3. Antebi YE, Nandagopal N, Elowitz MB. An operational view of intercellular signaling pathways. Current opinion in systems biology. 2017; 1:16–24. https://doi.org/10.1016/j.coisb.2016.12.003 PMID: 29104946

4. Jetka T, Nienałtowski K, Filippi S, Stumpf MP, Komorowski M. An information-theoretic framework for deciphering pleiotropic and noisy biochemical signaling. Nature communications. 2018; 9(1):4591. https://doi.org/10.1038/s41467-018-07085-1 PMID: 30389942

5. Petkova MD, Tkačik G, Bialek W, Wieschaus EF, Gregor T. Optimal Decoding of Cellular Identities in a Genetic Network. Cell. 2019; 176(4):844–855. https://doi.org/10.1016/j.cell.2019.01.007 PMID: 30712870

6. Berg HC, Purcell EM. Physics of chemoreception. Biophysical journal. 1977; 20(2):193–219. https://doi.org/10.1016/S0006-3495(77)85544-6 PMID: 911982

7. Brennan MD, Cheong R, Levchenko A. How information theory handles cell signaling and uncertainty. Science. 2012; 338(6105):334–335. https://doi.org/10.1126/science.1227946 PMID: 23087235

8. Komorowski M, Miękisz J, Stumpf MP. Decomposing noise in biochemical signaling systems highlights the role of protein degradation. Biophysical journal. 2013; 104(8):1783–1793. https://doi.org/10.1016/j.bpj.2013.02.027 PMID: 23601325

9. Waltermann C, Klipp E. Information theory based approaches to cellular signaling. Biochimica et Biophysica Acta (BBA)-General Subjects. 2011; 1810(10):924–932. https://doi.org/10.1016/j.bbagen.2011.07.009

10. Tkačik G, Walczak AM. Information transmission in genetic regulatory networks: a review. Journal of Physics: Condensed Matter. 2011; 23(15):153102. https://doi.org/10.1088/0953-8984/23/15/153102 PMID: 21460423

11. Mehta P, Goyal S, Long T, Bassler BL, Wingreen NS. Information processing and signal integration in bacterial quorum sensing. Molecular systems biology. 2009; 5(1):325. https://doi.org/10.1038/msb.2009.79 PMID: 19920810

12. Lee RE, Walker SR, Savery K, Frank DA, Gaudet S. Fold change of nuclear NF-κB determines TNF-induced transcription in single cells. Molecular cell. 2014; 53(6):867–879. PMID: 24530305

13. Zhang Q, Gupta S, Schipper DL, Kowalczyk GJ, Mancini AE, Faeder JR, et al. NF-κB Dynamics Discriminate between TNF Doses in Single Cells. Cell systems. 2017; 5(6):638–645. https://doi.org/10.1016/j.cels.2017.10.011 PMID: 29128333

14. Filippi S, Barnes CP, Kirk PD, Kudo T, Kunida K, McMahon SS, et al. Robustness of MEK-ERK dynamics and origins of cell-to-cell variability in MAPK signaling. Cell reports. 2016; 15(11):2524–2535. https://doi.org/10.1016/j.celrep.2016.05.024 PMID: 27264188

15. Suderman R, Bachman JA, Smith A, Sorger PK, Deeds EJ. Fundamental trade-offs between information flow in single cells and cellular populations. Proceedings of the National Academy of Sciences. 2017; 114(22):5755–5760. https://doi.org/10.1073/pnas.1615660114

16. Tkačik G, Callan CG, Bialek W. Information flow and optimization in transcriptional regulation. Proceedings of the National Academy of Sciences. 2008.

17. Bowsher CG, Swain PS. Identifying sources of variation and the flow of information in biochemical networks. Proceedings of the National Academy of Sciences. 2012; 109(20):E1320–E1328. https://doi.org/10.1073/pnas.1119407109

18. Levchenko A, Nemenman I. Cellular noise and information transmission. Current opinion in biotechnology. 2014; 28:156–164. https://doi.org/10.1016/j.copbio.2014.05.002 PMID: 24922112

19. Selimkhanov J, Taylor B, Yao J, Pilko A, Albeck J, Hoffmann A, et al. Accurate information transmission through dynamic biochemical signaling networks. Science. 2014; 346(6215):1370–1373. https://doi.org/10.1126/science.1254933 PMID: 25504722

20. Hansen AS, O'Shea EK. Limits on information transduction through amplitude and frequency regulation of transcription factor activity. Elife. 2015; 4:e06559. https://doi.org/10.7554/eLife.06559

21. Komorowski M, Tawfik DS. The Limited Information Capacity of Cross-Reactive Sensors Drives the Evolutionary Expansion of Signaling. Cell systems. 2019;. https://doi.org/10.1016/j.cels.2018.12.006 PMID: 30660612

22. Cheong R, Rhee A, Wang CJ, Nemenman I, Levchenko A. Information Transduction Capacity of Noisy Biochemical Signaling Networks. Science. 2011; 334(6054):354–358. https://doi.org/10.1126/science.1204553 PMID: 21921160

23. Billing U, Jetka T, Nortmann L, Wundrack N, Komorowski M, Waldherr S, et al. Robustness and Information Transfer within IL-6-induced JAK/STAT Signalling. Communications biology. 2019; 2(1):27. https://doi.org/10.1038/s42003-018-0259-4 PMID: 30675525

24. Voliotis M, Perrett RM, McWilliams C, McArdle CA, Bowsher CG. Information transfer by leaky, heterogeneous, protein kinase signaling systems. Proceedings of the National Academy of Sciences. 2014; 111(3):E326–E333. https://doi.org/10.1073/pnas.1314446111

25. Tkačik G, Dubuis JO, Petkova MD, Gregor T. Positional information, positional error, and read-out precision in morphogenesis: a mathematical framework. Genetics. 2014; p. 114.

26. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948; 27:623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

27. Blahut R. Computation of channel capacity and rate-distortion functions. IEEE transactions on Information Theory. 1972; 18(4):460–473. https://doi.org/10.1109/TIT.1972.1054855

28. Arimoto S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory. 1972; 18(1):14–20. https://doi.org/10.1109/TIT.1972.1054753

29. Vontobel PO. A generalized Blahut- Arimoto algorithm. In: Information Theory, 2003. Proceedings. IEEE International Symposium on. IEEE; 2003. p. 53.

30. Dubuis JO, Tkačik G, Wieschaus EF, Gregor T, Bialek W. Positional information, in bits. Proceedings of the National Academy of Sciences. 2013; p. 201315642. https://doi.org/10.1073/pnas.1315642110

31. Crisanti A, De Martino A, Fiorentino J. Statistics of optimal information flow in ensembles of regulatory motifs. Physical Review E. 2018; 97(2):022407. https://doi.org/10.1103/PhysRevE.97.022407 PMID: 29548237

32. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical review E. 2004; 69 (6):066138. https://doi.org/10.1103/PhysRevE.69.066138

33. Mack Y, Rosenblatt M. Multivariate k-nearest neighbor density estimates. Journal of Multivariate Analysis. 1979; 9(1):1–15. https://doi.org/10.1016/0047-259X(79)90065-4

34. Wang Q, Kulkarni SR, Verdú S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. IEEE Transactions on Information Theory. 2009; 55(5):2392–2405. https://doi.org/10.1109/TIT.2009.2016060

35. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences. 2014; p. 201309933. https://doi.org/10.1073/pnas.1309933111

36. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York, NY, USA:; 2001.

37. Silverman BW. Density Estimation for Statistics and Data Analysis. Routledge; 2018.

38. Cover TM, Thomas JA. Elements of information theory. John Wiley & Sons; 2012.

39. de Vos F, Schouten TM, Hafkemeijer A, Dopper EG, van Swieten JC, de Rooij M, et al. Combining multiple anatomical MRI measures improves Alzheimer's disease classification. Human brain mapping. 2016; 37(5):1920–1929. https://doi.org/10.1002/hbm.23147 PMID: 26915458

40. Zaidi NA, Webb GI, Carman MJ, Petitjean F, Cerquides J. ALRn: accelerated higher-order logistic regression. Machine Learning. 2016; 104(2-3):151–194. https://doi.org/10.1007/s10994-016-5574-8

41. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences. 2013; 110(15):5802–5805. https://doi.org/10.1073/pnas.1218772110

42. Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genetic epidemiology. 2011; 35(S1):S5–S11. https://doi.org/10.1002/gepi.20642 PMID: 22128059

43. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics. 2007; 9 (1):30–50. https://doi.org/10.1093/biostatistics/kxm010 PMID: 17429103

44. Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, Covert MW. Single-cell NF-κB dynamics reveal digital activation and analog information processing in cells. Nature. 2010; 466(7303):267. https://doi.org/10.1038/nature09145 PMID: 20581820

45. Oeckinghaus A, Hayden MS, Ghosh S. Crosstalk in NF-κB signaling pathways. Nature immunology. 2011; 12(8):695. https://doi.org/10.1038/ni.2065 PMID: 21772278

46. Sun SC. The non-canonical NF-κB pathway in immunity and inflammation. Nature Reviews Immunology. 2017; 17(9):545. https://doi.org/10.1038/nri.2017.52 PMID: 28580957

47. DeFelice MM, Clark HR, Hughey JJ, Maayan I, Kudo T, Gutschow MV, et al. NF-κB signaling dynamics is controlled by a dose-sensing autoregulatory loop. Sci Signal. 2019; 12(579):eaau3568. https://doi.org/10.1126/scisignal.aau3568 PMID: 31040261

48. Lipniacki T, Paszek P, Brasier AR, Luxon B, Kimmel M. Mathematical model of NF-κB regulatory module. Journal of theoretical biology. 2004; 228(2):195–215. https://doi.org/10.1016/j.jtbi.2004.01.001 PMID: 15094015

49. Ashall L, Horton CA, Nelson DE, Paszek P, Harper CV, Sillitoe K, et al. Pulsatile stimulation determines timing and specificity of NF-κB-dependent transcription. Science. 2009; 324(5924):242–246. https://doi.org/10.1126/science.1164860 PMID: 19359585

50. Nienałtowski K, Włodarczyk M, Lipniacki T, Komorowski M. Clustering reveals limits of parameter identifiability in multi-parameter models of biochemical dynamics. BMC systems biology. 2015; 9(1):65. https://doi.org/10.1186/s12918-015-0205-8 PMID: 26415494

51. Potter GD, Byrd TA, Mugler A, Sun B. Dynamic sampling and information encoding in biochemical networks. Biophysical journal. 2017; 112(4):795–804. https://doi.org/10.1016/j.bpj.2016.12.045 PMID: 28256238

52. Behar M, Barken D, Werner SL, Hoffmann A. The dynamics of signaling as a pharmacological target. Cell. 2013; 155(2):448–461. https://doi.org/10.1016/j.cell.2013.09.018 PMID: 24120141

53. Rué P, Arias AM. Cell dynamics and gene expression control in tissue homeostasis and development. Molecular systems biology. 2015; 11(2):792. https://doi.org/10.15252/msb.20145549 PMID: 25716053

54. Symmons O, Raj A. What's luck got to do with it: single cells, multiple fates, and biological nondeterminism. Molecular cell. 2016; 62(5):788–802. https://doi.org/10.1016/j.molcel.2016.05.023 PMID: 27259209

55. Keerthi SS, Duan KB, Shevade SK, Poo AN. A fast dual algorithm for kernel logistic regression. Machine Learning. 2005; 61(1):151–165. https://doi.org/10.1007/s10994-005-0768-5

56. Sricharan K, Raich R, Hero AO. K-nearest neighbor estimation of entropies with confidence. In: Proceedings of 2011 IEEE International Symposium on Information Theory; 2011. p. 1205–1209.

57. Loftsgaarden DO, Quesenberry CP. A nonparametric estimate of a multivariate density function. The Annals of Mathematical Statistics. 1965; p. 1049–1051. https://doi.org/10.1214/aoms/1177700079

58. Gao S, Ver Steeg G, Galstyan A. Efficient estimation of mutual information for strongly dependent variables. In: Artificial Intelligence and Statistics; 2015. p. 277–286.