Decomposing information into copying versus transformation

Artemy Kolchinsky[1] and Bernat Corominas-Murtra[2]*
[1] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[2] *Institute of Science and Technology Austria, Am Campus 1, A-3400, Klosterneuburg, Austria*

In many real-world systems, information can be transmitted in two qualitatively different ways: by *copying* or by *transformation*. *Copying* occurs when messages are transmitted without modification, e.g., when an offspring receives an unaltered copy of a gene from its parent. *Transformation* occurs when messages are modified systematically during transmission, e.g., when non-random mutations occur during biological reproduction. Standard information-theoretic measures do not distinguish these two modes of information transfer, although they may reflect different mechanisms and have different functional consequences. Starting from a few simple axioms, we derive a decomposition of mutual information into the information transmitted by copying and by transformation. Our decomposition applies whenever the source and destination of the channel have the same set of outcomes, so that a notion of message identity exists, although generalizations to other kinds of channels and similarity notions are explored. Furthermore, copy information can be interpreted as the minimal work needed by a physical copying process, relevant to better understand the physics of replication. We use the proposed decomposition to explore a model of amino acid substitution rates. Our results apply to any system in which the fidelity of copying, rather than simple predictability, is of critical relevance.

I. INTRODUCTION

Shannon's information theory provides a powerful set of tools for quantifying and analyzing information transmission. A particular measure of interest is *mutual information*, which is the most common way of quantifying the amount of information transmitted from a source to a destination. Mutual information has fundamental interpretations and operationalizations in a variety of domains, ranging from telecommunications [1, 2], gambling and investment [3–5], biological evolution [6], statistical physics [7, 8], and many others. Nonetheless, it has long been observed [9, 10] that mutual information does not distinguish between a situation in which the destination receives a *copy* of the source message versus one in which the destination receives some systematically *transformed* version of the source message (where by "systematic" we refer to transformations that do not arise purely from noise).

As an example of where this distinction matters, consider the transmission of genetic information during biological reproduction. When this process is modeled as a communication channel from parent to offspring, the amount of transmitted genetic information is often quantified by mutual information [11–15]. However, during reproduction, genetic information is not only copied but can also undergo systematic transformations in the form of "nonrandom mutations" [16, 17]. While mutual information does not distinguish which part of genetic information is transmitted by exact copying and which part is transmitted by nonrandom mutations, these two modes of information transmission are driven by different mechanisms and have dramatically different evolutionary and functional implications, not least because nonrandom mutations typically lead to deleterious or even fatal consequences.

The goal of this paper is to find a general decomposition of the information transmitted by a channel into contributions from copying versus from transformation. In Fig. 1, we provide a schematic that visually illustrates the problem. Essentially, we seek a decomposition of transmitted information into copy and transformation that distinguishes the example provided in (Fig. 1a), where the copy is perfect, from the one provided in (Fig. 1b), where the message has been systematically scrambled, from the one provided in (Fig. 1c), where the channel is completely noisy. Of course, we want also such a decomposition to apply in less extreme situations, where part of the information is copied and part is transformed.

The distinction between copying and transformation is important in many other domains beyond the case of biological reproduction outlined above. For example, in many models of animal communication and language evolution, agents exchange signals across noisy channels, and then use these signals to try to agree on common referents in the external world [10, 18–24]. In such models, successful communication occurs when information is transmitted by copying; if signals are systematically transformed — e.g., by scrambling — the agents will not be mutually intelligible, even though mutual information between them may be high. As another example, the distinction is relevant in the study of information flow during biological development, where recent work has investigated the ability of regulatory networks to decode development signals, such as positional information, from gene expression patterns [25]. In this scenario, information is copied when developmental signals are decoded correctly, and transformed when they are systematically decoded in an incorrect manner. Other examples are provided by Markov chain models, which are commonly used to study computation and other dynamical processes in physics [26], biology [27] or sociology [28], among other fields. Indeed, Markov chains can be seen as a communication channel in which the system state transmits information from the past into the future. In this context, copying occurs when the system maintains its state constant over time (remains in fixed points) and transformation occurs when the state undergoes systematic changes (e.g., performs some kind of non-trivial computations).

Interestingly, while the distinction between copy and transformation information seems natural, it has not been previ-

* Author for correspondence: bernat.corominas-murtra@ist.ac.at

ously considered in the information-theoretic literature. This may be partly due to the different roles that information theory has historically played: on one hand, a field of applied mathematics concerned with the engineering problem of optimizing information transmission — its original purpose — on the other a set of quantitative tools for analyzing intrinsic properties of real-world systems. Because of its origins in engineering, much of information theory, including Shannon's channel-coding theorem, which established mutual information as a fundamental measure of transmitted information [2, 29, 30], is formulated under the assumption of an external agent who can appropriately encode and decode information for a given communication channel, in this way accounting for any transformations performed by the channel. However, in many real-world systems, there is no additional external agent who codes for the channel [10, 31], and one is interested in the ability to copy information across a channel without any additional encoding or decoding. This latter property is the main subject of this paper.

A final word is required to motivate our information-theoretic approach. It is standard to characterize the ability of a channel to copy messages via the "probability of error" statistic, which we indicate as ϵ [2]. In particular, ϵ reflects the probability that the destination receives a different message than the one that was sent by the source, while $1-\epsilon$ reflects the probability that the destination receives the same message as was sent by the source. However, for our purposes, this approach is insufficient. First of all, while $1-\epsilon$ quantifies the propensity of a channel to copy information, ϵ does not measure propensity to transmit information by transformation, since it increases both in the presence of transformation and in the presence of noise (in other words, ϵ is high both in channels like Fig. 1b and channels Fig. 1c). Among other things, this means that $1-\epsilon$ and ϵ cannot be used to compute a channel's "copying efficiency" (i.e., which portion of the total information transmitted across a channel is copied). Second, and more fundamentally, ϵ and $1-\epsilon$ are not information-theoretic quantities, in the sense that they do not reflect the *amount* of information. For instance, $1-\epsilon$ is bounded between 0 and 1 for all channels, whether considering a simple binary channel or a high-speed fiber-optic line — in the language of physics, ϵ is an intensive property, rather than extensive one that scales with the size of the channel. We instead seek measures which quantify the amount of copied and transformed information, and which can grow as the capacity of the channel under consideration increases.

We present a decomposition of information that distinguishes copied from transformed information. We derive our decomposition by proposing four natural axioms that copy and transformation information should satisfy, and then identifying the unique measure that satisfy these axioms. Our resulting measure is easy to compute and can be used to decompose either the total mutual information flowing across a channel, or the specific mutual information corresponding to a given source message, or an even more general measure of acquired information called *Bayesian surprise*.

The paper is laid out as follows. We present our approach in the next section. In Section III, we demonstrate that copy information is related to, though different from, the notion of *rate-distortion* in information theory [2]. We also show that while our basic decomposition is defined for discrete-state channels where the source and destination share the same set of possible messages, so that the notion of "exact copy" is simple to define, our measures can be generalized to continuous-state channels as well as other possible definitions of copying. In Section IV, we show that our measure can be used to quantify the thermodynamic efficiency of physical copying processes, a central topic in the biological domain. In Section V, we demonstrate our measures on a real world dataset of amino acid substitution rates.

II. COPY AND TRANSFORMATION INFORMATION

We first derive our proposed measure of copy information from four simple and intuitive axioms, and then construct the decomposition of mutual information into *copy* and *transformation*. Before proceeding, we briefly present some basic concepts from information theory that will be useful for our further developments.

A. Preliminaries

We use the random variables X and Y to indicate the source and destination, respectively, of a communication channel (as defined in detail below). We assume that the source X and destination Y both take outcomes from the same countable set $\mathcal{A}$. We use Δ to indicate the set of all probability distributions whose support is equal to, or a subset of, $\mathcal{A}$. We use notation like $p_Y, q_Y, \cdots \in \Delta$ to indicate marginal distributions over Y, and $p_{Y|x}, q_{Y|x}, \cdots \in \Delta$ to indicate conditional distributions over Y, given the event $X = x$. Where clear from context, we will simply write $p(y), q(y), \ldots$ and $p(y|x), q(y|x), \ldots$, and drop the subscripts.

For any two distributions s and q over the same set of outcomes, the *Kullback-Leibler* (KL) divergence is defined as

$$D_{\mathsf{KL}}(s\|q) := \sum_x s(x) \log \frac{s(x)}{q(x)}. \tag{1}$$

In this paper we will make use of the KL between Bernoulli distributions — that is, distributions over two states of the type $(a, 1-a)$ — which is sometimes called "binary KL". We will use the notation $\mathsf{d}(a,b)$ to indicate the binary KL,

$$\mathsf{d}(a,b) := a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}. \tag{2}$$

KL is always non-negative, and equal to 0 if and only if $s(x) = q(x)$ for all x. Throughout this paper, we assume base 2 (meaning that information is measured in bits), unless otherwise noted.

In information theory, a *communication channel* specifies the conditional probability distribution of receiving different messages at a destination given messages transmitted by a source. Let $p_{Y|X}(y|x)$ indicate such a conditional probability

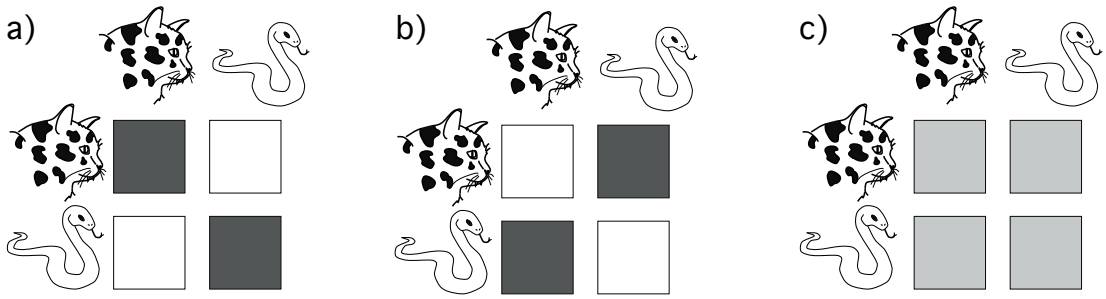

Figure 1. An illustration of the problem of copy and transformation. Consider three channels, each of which can transmit two messages, indicated by *cat* and *snake* –e.g., alarm calls in an animal communication system. In all panels, the rows indicate the message selected at the source, the columns indicate the message received at the destination, and the shade of the respective square indicates the conditional probability of the destination message given source message. For the channel in (a), all information is copied: the channel maps *cat*→*cat* and *snake*→*snake* with probability 1. For the channel in (b), all information is transformed: the channel maps *cat*→*snake* and *snake*→*cat* with probability 1. Note that for any source distribution, the mutual information between source and destination is the same in (a) and (b). The channel in (c) is completely noisy: the probability of receiving different messages at the destination do not depend on the message selected at the source, and the mutual information between source and destination is 0. Observe that transformation is different than noise, in it still involves the transmission of information.

distribution. The amount of information transferred across a communication channel, given some probability distribution of source messages $s_X(x)$, is quantified using the *mutual information* (MI) between the source and the destination [2],

$$I_p(Y:X) := \sum_x s(x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y)}, \tag{3}$$

where $p(y)$ is the marginal probability of receiving message y at the destination, defined as

$$p(y) := \sum_x s(x) p(y|x). \tag{4}$$

When writing $I_p(Y:X)$, we will omit the subscript p indicating the channel where it is clear from context. MI is a fundamental measure of information transmission, and can be operationalized in numerous ways [2]. It is non-negative, and large when (on average) the uncertainty about the message at the destination is low, given the source message.

Importantly, MI can also be written as a weighted sum of so-called *specific MI*[32] terms [33], one for each outcome of X,

$$I(Y:X) = \sum_x s(x) I(Y:X{=}x), \tag{5}$$

where the specific MI for outcome x is given by

$$I(Y:X{=}x) := \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} = D_{\mathsf{KL}}(p_{Y|x} \| p_Y). \tag{6}$$

Each $I(Y:X=x)$ indicates the contribution to MI arising from the particular source message x. We will sometimes use the term *total mutual information* (total MI) to refer to Eq. (3), so as to distinguish it from specific MI.

Specific MI also has an important Bayesian interpretation. Consider an agent who begins with a set of prior beliefs about Y, as specified by the prior distribution $p_Y(y)$. Upon observing the event $X = x$, the agent updates their beliefs, resulting in the posterior distribution $p_{Y|x}(y)$. The KL divergence between the posterior and the prior, $D_{\mathsf{KL}}(p_{Y|x} \| p_Y)$ (Eq. (6)), is called *Bayesian surprise* [34], and quantifies the amount of information acquired by the agent. It reaches its minimum value of zero, indicating that no information is acquired, if and only if the prior and posterior distributions match exactly. Bayesian surprise plays a fundamental role in Bayesian theory, including in the design of optimal experiments [35–38] and the selection of "non-informative priors" [39, 40]. Specific MI is a special case of Bayesian surprise, when the prior p_Y is the marginal distribution at the destination, as determined by a choice of source distribution s_X and channel $p_{Y|X}$ according to Eq. (4). In general, however, Bayesian surprise may be defined for any desired prior p_Y and posterior distribution $p_{Y|x}$, without necessarily making reference to a source distribution s_X and a communication channel $p_{Y|X}$.

Because Bayesian surprise it is a general measure that includes specific MI as a special case, we will formulate our analysis of copy and transformation information in terms of $D_{\mathsf{KL}}(p_{Y|x} \| p_Y)$. Note that while the notation $p_{Y|x}$ implies conditioning on the event $X = x$, formally $p_{Y|x}$ can be any distribution whatsoever. Thus, we do not technically require that there exist some full joint or conditional probability distribution over X and Y. Throughout the paper we will refer to the distributions $p_{Y|x}$ and p_Y as the "posterior" and "prior".

Proofs and derivations are contained in the appendix.

B. Axioms for copy information

To derive a measure of copy information, we propose four axioms for copy information. Our axioms are motivated by the following ideas: First, our decomposition should apply

at the level of individual source message, i.e., we wish to be able to decompose each specific mutual information term (or more generally, Bayesian surprise) into a *(specific) copy information* term and a non-negative *(specific) transformation information* term. Second, we postulate that if there are two channels with the same marginal distribution at the destination, then the channel with the larger $p(x|x)$ (probability of destination getting message x when the source transmits message x) should have larger copy information for source message x (this is, so to speak, our "central axiom"). Alternatively, under the Bayesian interpretation, this postulate can be interpreted as saying that if two Bayesian agents both start with the same prior and receive a piece of evidence x, then copy information should be greater for the agent whose posterior probability of x is larger.

Formally, we assume that each copy information term is a real-valued function of the posterior distribution, the prior distribution, and the source message x, written generically as $F(p_{Y|x}, p_Y, x)$. Given any measure of copy information F, the transformation information associated with message x is then the remainder of $D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$ beyond F,

$$F^{\mathsf{trans}}(p_{Y|x}, p_Y, x) := D_{\mathsf{KL}}(p_{Y|x}\|p_Y) - F(p_{Y|x}, p_Y, x). \tag{7}$$

We now propose a set of axioms that any measure of copy information should satisfy.

First, we postulate that copy information should be bounded between 0 and the Bayesian surprise, $D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$. Given Eq. (7), this guarantees that both F and F^{trans} are non-negative.

Axiom 1. $F(p_{Y|x}, p_Y, x) \geq 0$.

Axiom 2. $F(p_{Y|x}, p_Y, x) \leq D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$.

Then, we postulate that copy information for source message x should increase monotonically as the posterior probability of x increases, assuming the prior distribution is held fixed. (This is the "central axiom" mentioned above.)

Axiom 3. *If* $p_{Y|x}(x) \leq q_{Y|x}(x)$, *then* $F(p_{Y|x}, p_Y, x) \leq F(q_{Y|x}, p_Y, x)$.

In Appendix C, we show that any copy information that satisfies the above three axioms must obey $F(p_{Y|x}, p_Y, x) = 0$ whenever $p_{Y|x}(x) \leq p_Y(x)$. What happens in the case $p_{Y|x}(x) > p_Y(x)$, however, is not totally determined. In fact, there are some trivial functions, such as $F(p_{Y|x}, p_Y, x) = 0$ for all $p_{Y|x}$, p_Y, and x, that satisfy the above axioms. Such trivial cases are excluded by our final axiom, which states that for all prior distribution and all posterior probabilities $p_{Y|x}(x) > p_Y(x)$, there are posterior distributions that contain *only* copy information.

Axiom 4. *For any* p_Y *and* $c \in [p_Y(x), 1]$, *there exists a posterior distribution* $p_{Y|x}$ *such that* $p_{Y|x}(x) = c$ *and* $F(p_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$.

C. The measure D_x^{copy}

We now present the unique measure, which we call D_x^{copy}, that satisfies the copy information axioms proposed in the last section. Given a prior distribution p_Y, posterior distribution $p_{Y|x}$, and source message x, D_x^{copy} is defined as

$$\begin{aligned} D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) := \min_{r_Y \in \Delta} \; & D_{\mathsf{KL}}(r_Y\|p_Y) \\ \text{s.t.} \;\; & r_Y(x) \geq p_{Y|x}(x). \end{aligned} \tag{8}$$

The minimization problem of Eq. (8) turns out to have a simple closed-form solution, as shown in Appendix A:

$$D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) = \begin{cases} \mathsf{d}(p_{Y|x}(x), p_Y(x)) & \text{if } p_{Y|x}(x) > p_Y(x) \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where we have used the notation of Eq. (2). It turns out that D_x^{copy} is the *unique* measure that satisfies the four proposed axioms. We can now state the main result of our paper:

Theorem 1. D_x^{copy} *is the unique measure which satisfies Axioms 1 to 4.*

In the Appendix B we demonstrate that D_x^{copy} satisfies all the axioms, and in the Appendix C we prove that it is actually the only measure that satisfies them. We further show that if we drop Axiom 4, then D_x^{copy} is the largest possible measure that can satisfy the remaining axioms. Finally, note that given the definition of F^{trans} in Eq. (7), D_x^{copy} also defines a particular, non-negative measure of transformation information, which we call D_x^{trans},

$$D_x^{\mathsf{trans}}(p_{Y|x}\|p_Y) = D_{\mathsf{KL}}(p_{Y|x}\|p_Y) - D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y). \tag{10}$$

D. Decomposing mutual information

We now show that D_x^{copy} and D_x^{trans} provide the ingredients for a decomposition of mutual information (MI) into *MI due to copying* and *MI due to transformation*. Recall that MI can be written as an expectation over specific MI terms, as shown in Eq. (6). Each specific MI term can be seen as a Bayesian surprise, where the prior distribution is the marginal distribution at the destination (see Eq. (4)), and the posterior distribution is the conditional distribution of destination given a particular source message under the channel. Thus, our definitions of D_x^{copy} and D_x^{trans} provide a non-negative decomposition of each specific MI term,

$$I_p(Y:X{=}x) = D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) + D_x^{\mathsf{trans}}(p_{Y|x}\|p_Y). \tag{11}$$

In consequence, they also provide a non-negative decomposition of the total MI into two non-negative terms: the *total copy information* and the *total transformation information*,

$$I_p(Y:X) = I_p^{\mathsf{copy}}(X{\to}Y) + I_p^{\mathsf{trans}}(X{\to}Y), \tag{12}$$

where $I_p^{\mathsf{copy}}(X{\to}Y)$ and $I_p^{\mathsf{trans}}(X{\to}Y)$ are given by

$$I_p^{\mathsf{copy}}(X{\to}Y) := \sum_x s(x) D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y), \tag{13}$$

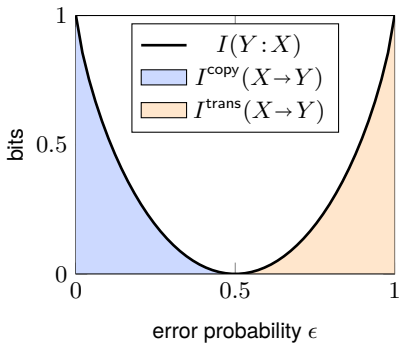

Figure 2. The Binary Symmetric Channel (BSC) with a uniform source distribution. Values of mutual information, $I^{\mathsf{copy}}(X \to Y)$ (Eq. (13)) and $I^{\mathsf{trans}}(X\to Y)$ (Eq. (14)) for the BSC along the whole range of error probabilities $\epsilon \in [0, 1]$. When $\epsilon \leq 1/2$, all mutual information is I^{copy} (blue shading), while when $\epsilon \geq 1/2$, all mutual information is I^{trans} (red shading).

$$I_p^{\mathsf{trans}}(X\to Y) := \sum_x s(x) D_x^{\mathsf{trans}}(p_{Y|x} \| p_Y). \tag{14}$$

(When writing I^{copy} and I^{trans}, we will often omit the subscript p where the channel is clear from context.)

In the definitions of $I_p^{\mathsf{copy}}(X \to Y)$ and $I_p^{\mathsf{trans}}(X \to Y)$, the notation $X \to Y$ indicates that I^{copy} and I^{trans} are defined as the copy and transformation information when X is the source and Y is the destination. This is necessary because, unlike MI, I^{copy} and I^{trans} are in general non-symmetric, so it is possible that $I^{\mathsf{copy}}(X\to Y) \neq I^{\mathsf{copy}}(Y \to X)$, and similarly for I^{trans}. We also note that the above form of I^{copy} and I^{trans}, where they are written as sums over individual source message, is sometimes referred to as *trace-like* form in the literature, and is a common desired characteristic of information-theoretic functionals [41, 42].

For illustration purposes, in Fig. 2 we plot the behavior of I^{copy} and I^{trans} in the classical binary symmetric channel (BSC) (see caption for details). More detailed analysis of copy and transformation informatoin in the BSC is discussed in Appendix F.

Finally, we note that by a simple manipulation, we can decompose the marginal the entropy of the destination $H(Y)$ into three non-negative components:

$$H(p_Y) = I_p^{\mathsf{copy}}(X\to Y) + I_p^{\mathsf{trans}}(X\to Y) + H(p_{Y|X}). \tag{15}$$

Thus, given a channel from X to Y, the uncertainty in Y can be written as the sum of the copy information from X, the transformed information from X, and the intrinsic noise in that channel from X to Y.

E. Copying efficiency

As mentioned in the Introduction, our approach provides a way to quantify which portion of the information transmitted across a channel is due to copying rather than transformation, which we refer to as "copying efficiency". Copying efficiency is defined at the level of individual source messages as

$$\eta_p(x) := \frac{D_x^{\mathsf{copy}}(p_{Y|x} \| p_Y)}{D_{\mathsf{KL}}(p_{Y|x} \| p_Y)} \in [0, 1], \tag{16}$$

where the bounds come directly from Axioms 1 and 2. It can also be defined at the level of a channel as whole as

$$\eta_p := \frac{I^{\mathsf{copy}}(X\to Y)}{I(Y:X)} \in [0, 1]. \tag{17}$$

The bounds follow simply given the above results.

For Eq. (16) and Eq. (17) to be useful efficiency measures, there should exist channels which are either "completely inefficient" (have efficiency 0) or "maximally efficient" (achieve efficiency 1). For the case of Eq. (16), the bounds can be saturated because of Axiom 4, which guarantees that for any source message x, prior p_Y, and desired posterior probability $p_{Y|x}(x) > p_Y(x)$, there exists a posterior $p_{Y|x}$ such that the Bayesian surprise $D_{\mathsf{KL}}(p_{Y|x} \| p_Y)$ is composed entirely of copy information (for example, see Eq. (B1)).

One can show that the bounds in Eq. (17) can also be saturated. First, it can be verified that completely inefficient channels exists, since any channel which has $p_{Y|x}(x) \leq p_Y(x)$ for all $x \in \mathcal{A}$ will have $I^{\mathsf{copy}}(X\to Y) = 0$ (note that such channels exists at all levels of mutual information). We also prove that maximally efficient channels exist, using the following result which is proved in Appendix E.

Proposition 1. *For any source distribution s_X with full support and $H(s_X) < \infty$, there exist channels p for all levels of mutual information $I_p(Y:X) \in [0, H(s_X)]$ such that $I_p^{\mathsf{copy}}(X\to Y) = I_p(Y:X)$.*

III. GENERALIZATION AND RELATION TO RATE DISTORTION

In this section, we show that D_x^{copy} can be written as a particular element among a broad family of copy information measures, which generalize the formal definition of what is meant by "copying". All elements of this family of copying information measures have the following form:

$$\hat{F}(p_{Y|x}, p_Y, x) := \min_{r_Y} D_{\mathsf{KL}}(r_Y \| p_Y) \quad \text{s.t. } \mathbb{E}_{r_Y}[\ell(x, Y)] \leq \mathbb{E}_{p_{Y|x}}[\ell(x, Y)], \tag{18}$$

where $\ell(x, y)$ is some function that specifies the "loss" the occurs when source message x is received by the destination as message y. D_x^{copy} has the form of Eq. (18) if one takes $\ell(x, y) := 1 - \delta(x, y)$, which is called "0-1 loss" in statistics [43] and "Hamming distortion" in information theory [2], and which formalizes copy information in terms of exact identity between source and destination message.

Each particular loss function induces its own measure of copy information, allowing us to define copy information in a more general manner. While D_x^{copy}, as defined via the 0-1 loss function, is a simple and reasonable choice in a variety

of applications, in some cases copy information could also be defined in a more general way, in terms of another loss function. One important application of this generalization is the situation where copying is not defined in terms of simple identity between source and destination messages, but rather in terms of some more continuous measure of similarity. This is especially useful for continuous-valued random variables, because in this case the constraint in Eq. (8) is no longer meaningful (it only constraints a measure-0 property of r_Y). Here a natural measure of copy information is given by the squared error loss function, $\ell(x,y) := (x-y)^2$, giving

$$\min_{r_Y} D_{\mathsf{KL}}(r_Y \| p_Y) \quad \text{s.t.} \quad \mathbb{E}_{r_Y}[(Y-x)^2] \leq \mathbb{E}_{p_{Y|x}}[(Y-x)^2].$$

Note that this optimization problem has been investigated in the maximum entropy literature, and has been shown to be particularly tractable when p_Y belongs to an exponential family [44–46].

Recall that we have so far assumed that the source and destination share the same set of possible outcomes, $\mathcal{A}$. Importantly, with a generalized copy information the case where the source and destination do not share the same set of outcomes can be studied. In that case, some measure of similarity, must be defined between source and destination messages.

The optimization problem of Eq. (18) has some interesting properties. First, it is an example of a "minimum cross-entropy" distribution, which itself is a generalization of the "maximum entropy" principle [47, 48]. Using this fact, in Appendix D we show that the optimal solution has the following simple form

$$w_Y(y) = \begin{cases} \frac{p_Y(y)}{Z} e^{-\lambda \ell(x,y)} & \text{if } \mathbb{E}_{p_{Y|x}}[\ell(x,Y)] < \mathbb{E}_{p_Y}[\ell(x,Y)] \\ p_Y(y) & \text{otherwise} \end{cases}, \tag{19}$$

where Z is an appropriate normalization constant and $\lambda > 0$ is a Lagrange multiplier chosen so that the constraint Eq. (18) is satisfied. Thus, Eq. (18) can be solved efficiently by searching across the 1-dimensional space of possible λ values.

In Appendix D, we also show that Eq. (18) emerges as the unique measure that satisfies our aforementioned set of four axioms, as long as Axiom 3 and Axiom 4 are replaced by their "generalized" versions. The generalized version of Axiom 3 states that one posterior distribution should have higher copy information than another whenever its expected loss is lower:

Axiom 3*. *If* $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] \geq \mathbb{E}_{q_{Y|x}}[\ell(x,Y)]$, *then* $F(p_{Y|x}, p_Y, x) \leq F(q_{Y|x}, p_Y, x)$.

The generalized version of Axiom 4 states that at all expected loss values lower than that achieved by p_Y, there are channels which transmit information only by occupying.

Axiom 4*. *For any* p_Y *and* $c \in [\mathbb{E}_{p_Y}[\ell(x,Y)], \min_y \ell(x,y)]$, *there exists a posterior distribution* $p_{Y|x}$ *such that* $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] = c$ *and* $F(p_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(p_{Y|x} \| p_Y)$.

Note that we used that $\min_y \ell(x,y)]$ is the lowest possible expected loss that can be achieved by any distribution.

We finish by noting that Eq. (18) shows a similarity between our approach and so called *rate-distortion theory* [2]. In rate distortion theory, one is given a distribution over source messages s_X and a "distortion function" $\ell(x,y)$ which specifies the loss incurred when source message x is encoded with destination message y. The problem is to find the channel $p_{Y|X}$ which minimizes mutual information without exceeding some constraint on the expected distortion,

$$\min_{p_{Y|X}} D_{\mathsf{KL}}(p_{Y|X} \| p_Y) \quad \text{s.t.} \quad \mathbb{E}_p[\ell(X,Y)] \leq \alpha\,, \tag{20}$$

where α is an externally-determined parameter. The prototypical application of rate-distortion theory is compression, i.e., finding representations of the source that are both compressed (have low mutual information) and have low distortion. As can be seen by comparing Eq. (18) and Eq. (20), the optimization problem considered in our definition of copy information and the optimization found in rate distortion are quite similar: they both involve minimizing a KL divergence subject to an expected loss constraint. Nonetheless, there are some important differences. First and foremost, the goals of the two approaches are different: we aim to decompose the information transmitted by a fixed externally-specified channel into copy and transformation (it is this given channel that determines the constraint α in Eq. (18)), while in rate-distortion there is no externally-specified channel and the aim is instead to find an optimal channel *de novo*. In this respect, our approach also has some similarities to the so-called *minimum information principle*, which aims to quantify how much information about sent messages is carried by different properties of received messages [49]. Second, our approach is motivated by a set of axioms which postulate how a measure of copy information should behave, rather than from channel-coding considerations as is used to derive the optimization problem in rate-distortion [2]. Lastly, copy information is defined in a point-wise manner for each source message x, rather than for an entire set of source messages at once as is rate-distortion. However, we do note that in principle it should be possible to define Eq. (18) in a way that include copy information measures defined in a channel-wise manner (like Eq. (20)) rather than a pointwise manner (like Eq. (8)).

IV. THERMODYNAMIC COSTS OF COPYING

Given the close connection between information theory and statistical physics, many information-theoretic quantities can be interpreted in thermodynamic terms [8]. As we show here, this includes our proposed measure of copy information, D_x^{copy}. Specifically, we will show that D_x^{copy} reflects the minimal amount of thermodynamic work necessary to copy a physical entity such as a polymer molecule. This latter example emphasizes the important difference between information transfer by copying versus by transformation in a fundamental, biologically-inspired physical setup.

Consider a physical system coupled to a heat bath at temperature T, and which is initially in equilibrium distribution $\pi(i) \propto e^{-E_i/(k_B T)}$ with respect to some Hamiltonian E (k_B is

Boltzmann's constant). Now imagine that the system is driven to some non-equilibrium distribution p by a physical process, and that by the end of the process the Hamiltonian is again equal to E. The minimal amount of work required by any such process is related to the KL divergence between p and π [50],

$$W \geq k_B T \; D_{\mathsf{KL}}(p\|\pi). \tag{21}$$

The limit is achieved by thermodynamically-reversible processes. (In this subsection, in accordance with the convention in physics, we assume that all logarithms are in base e, so information is measured in nats.)

Recent work has analyzed the fundamental thermodynamic constraints on copying in a physical system, for example for an information-carrying polymer like DNA [51, 52]. Here we will generally follow the model described in [51], while using our notation and omitting some details that are irrelevant for our purposes (such as the microstate/macrostate distinction). In this model, the source X represents the state of the original system (e.g., the sequence of the polymer to be copied), and the destination Y represents the state of the product (e.g., the sequence that is produced by the copying mechanism). We imagine that the original system is in state $X = x$, and that replicates are produced according to the conditional distribution $p_{Y|x}(x)$. Finally, assume that X and Y are statistically independent under the equilibrium distribution (in other words, that there the energetic coupling between X and Y is negligible compared to their internal energies), and let π_Y indicate the equilibrium distribution of the replicate Y. Following Eq. (21), the minimal work required to bring system Y out of equilibrium and produce replicates of x according to $p_{Y|x}(x)$ is given by

$$W(x) \geq k_B T \; D_{\mathsf{KL}}(p_{Y|x}\|\pi_Y). \tag{22}$$

Note that Eq. (22) specifies the minimal work required to create the overall distribution $p_{Y|x}$. However, in many real-world scenarios, likely including DNA copying, the primary goal is to create exact copies of the original state, not transformed versions it (such as nonrandom mutations). That means that for a given source state x, the quality of the replication process can be quantified by the probability of making an exact copy, $p_{Y|x}(x)$. We can now ask: what is the minimal work required by any physical replication process whose probability of making exact copies is at least as large as $p_{Y|x}(x)$? To make the comparison fair, we require that the process operate under the same equilibrium distribution, π_Y. The answer is given by the minimum of the RHS of Eq. (22) under a constraint on the exact-copy yield, which is exactly proportional to D_x^{copy}:

$$W_{\text{min}}^{\text{exact}}(x) = k_B T \left[\min_{r_Y : r_Y(x) \geq p_{Y|x}(x)} D_{\mathsf{KL}}(r_Y\|\pi_Y) \right] \tag{23}$$

$$= k_B T \; D_x^{\mathsf{copy}}(p_{Y|x}\|\pi_Y). \tag{24}$$

The additional work that is expended by the replication process is then lower bounded by a quantity proportional D_x^{trans},

$$W(x) - W_{\text{min}}^{\text{exact}}(x) \geq D_x^{\mathsf{trans}}(p_{Y|x}\|\pi_Y)\,. \tag{25}$$

This shows formally the intuitive idea that transformation information contributes to thermodynamic costs but not to the accuracy of correct copying.

In most cases, a replication system is designed for copying not just one source state x, but an entire ensemble of source states, which can be assumed to be distributed according to some $s_X(x)$ (for example, the DNA replication system can copy a huge of ensemble of source DNA sequences, not just one). Across the entire ensemble of source states, the minimal amount of *expected* thermodynamic work required to produce replicates according to conditional distribution $p_{Y|X}$ is given by

$$\langle W \rangle \geq k_B T \sum_x s(x) D_{\mathsf{KL}}(p_{Y|x}\|\pi_Y) \tag{26}$$

$$= k_B T \left[I_p(Y:X) + D_{\mathsf{KL}}(p_Y\|\pi_Y) \right]. \tag{27}$$

Since KL is non-negative, the minimum expected work is lowest when the equilibrium distribution π_Y matches the marginal distribution of replicates, $p_Y(y) = \sum_x s(x)p(y|x)$. Using similar arguments as above, we can ask about the minimum expected work required to produce replicates, assuming each source state x achieves an exact-copy yield of at least $p_{Y|x}(x)$. This turns out to be the expectation of Eq. (24),

$$\langle W_{\text{min}}^{\text{exact}} \rangle = k_B T \sum_x s(x) D_x^{\mathsf{copy}}(p_{Y|x}(x)\|\pi_Y) \tag{28}$$

$$= k_B T \left[I_p^{\mathsf{copy}}(X \to Y) + D_{\mathsf{KL}}(p_Y\|\pi_Y) \right]. \tag{29}$$

The additional expected work that is needed by the replication process, above and beyond an optimal process that achieves the same exact-copy yield, is lower bounded by the transformation information,

$$\langle W \rangle - \langle W_{\text{min}}^{\text{exact}} \rangle \geq k_B T I_p^{\mathsf{trans}}(X \to Y)\,. \tag{30}$$

When the equilibrium distribution π_Y matches the marginal distribution p_Y, $\langle W_{\text{min}}^{\text{exact}} \rangle$ is exactly equal $k_B T I^{\mathsf{copy}}$. Furthermore, in this special case the thermodynamic efficiency of exact copying, defined as the ratio of minimal work to actual work, becomes equal to the information-theoretic copying efficiency of p, as defined in Eq. (17):

$$\frac{\langle W_{\text{min}}^{\text{exact}} \rangle}{\langle W \rangle} = \frac{I_p^{\mathsf{copy}}(X \to Y)}{I_p(Y:X)} = \eta_p. \tag{31}$$

As can be seen, standard information-theoretic measures, such as Eq. (22), bound the minimal thermodynamic costs of transferring information from one physical system to another, whether that transfer happens by copying or by transformation. However, as we have argued above, the difference between copying and transformation is essential in many biological scenarios, as well as other domains. In such cases, D_x^{copy} arises naturally as the minimal thermodynamic work required to replicate information by copying. Concerning the example of the DNA copying that we followed throughout this section, we note that these results should be interpreted with some care. We have generally imagined that the source system represents the state of an entire polymer, e.g., the state of an

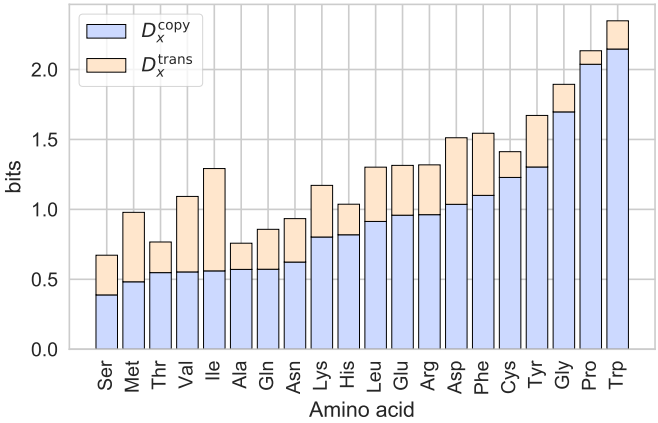

Figure 3. Copy and transformation information for different amino acids, based on an empirical PAM matrix [54]. In blue, we show magnitude of $D_x^{\text{copy}}(p_{Y|x}\|p_Y)$, on top, the amount of transformed information, in red. The sum of both is the specific MI for each amino acid, according to the decomposition given in equation (11)

entire DNA molecule, and that the probability of exact copying refers to the probability that the entire sequence is reproduced without any errors. Alternatively, one can use the same framework to consider probability of copying a single monomer in a long polymer (assuming that the thermodynamics of polymerization can be disregarded), as might be represented for instance by a single-nucleotide DNA substitution matrix [53], as analyzed in the last section. Generally speaking, D_x^{copy} computed at the level of single monomers will be different from D_x^{copy} computed at the level of entire polymers, since the probability of exact copying means different things in these two formulations.

V. COPY AND TRANSFORMATION IN AMINO ACID SUBSTITUTION MATRICES

In the previous section, we saw how D_x^{copy} and I^{copy} arises naturally when studying the fundamental limits on the thermodynamics of copying, which includes the special case of replicating information-bearing polymers. Here we demonstrate how these measures can be used to characterize the information-transmission properties of a real-world biological replication system, as formalized by a communication channel $p_{Y|X}$ from parent to offspring [53, 54]. In this context, we show how I^{copy} can be used to quantify precisely how much information is transmitted by copying, without mutations. At the same time, we will use I^{trans} to quantify how much information is transmitted by transformation, that is by systematic *nonrandom* mutations that carry information but do not preserve the identity of the original message [16, 17]. We also quantify the effect of purely-random mutations, which correspond to the conditional entropy of the channel, $H(p_{Y|X})$.

We demonstrate these measures on empirical data of *point accepted mutations* (PAM) of amino acids. PAM data represents the rates of substitutions between different pairs of amino acids during the course of biological evolution, and has various applications, including evolutionary modeling, phylogenetic reconstructions, and protein alignment [54]. We emphasize that PAM matrices between amino acids do not reflect the direct physical transfer of information from protein to protein, but rather the result effects of underlying processes of DNA-based replication and selection, followed by translation.

Formally, a PAM matrix Q for amino acids is a continuous-time rate matrix, where Q_{yx} represents the instantaneous rates of substitutions from amino acid x to amino acid y, where both x and y belong to $\mathcal{A} = \{1, \ldots, 20\}$, representing the 20 standard amino acids. We performed our analysis on a fixed PAM matrix Q, which was published by Le and Gascuel [54] (this matrix was provided by the `pyvolve` Python package [55]). We calculated a discrete-time conditional probability distribution $p_{Y|X}$ from this matrix by taking the matrix exponential, $p_{Y|X} = \exp(\tau Q)$. Thus, $p(y|x)$ represents the probability that amino acid x is replaced by amino acid y over time scale τ. For simplicity, we used timescale $\tau = 1$ and assumed that the source distribution over amino acids is uniform, $s(x) = 1/20$ for all x. Using the decomposition presented in Eq. (14), we get the following values for the communication channel described by the conditional probabilities $p_{Y|X}$:

$$I(Y:X) = I^{\text{copy}}(X\to Y) + I^{\text{trans}}(X\to Y) \approx 1.3 \text{ bits},$$

where

$$I^{\text{copy}}(X\to Y) = \sum_x s(x) D_x^{\text{copy}}(p_{Y|x}\|p_Y) \approx 0.965 \text{ bits},$$
$$I^{\text{trans}}(X\to Y) = \sum_x s(x) D_x^{\text{trans}}(p_{Y|x}\|p_Y) \approx 0.335 \text{ bits}.$$

We also computed the intrinsic noise for this channel (see Eq. (15)),

$$H(p_{Y|X}) = \sum_x s(x) H(p_{Y|x}(x)) \approx 2.97 \text{ bits}.$$

Finally, we computed the specific copy and transformation information, D_x^{copy} and D_x^{trans}, for different amino acids. The results for different amino acids are shown in Fig. 3. We remind the reader that the sum of $D_x^{\text{copy}}(p_{Y|x}\|p_Y)$ and $D_x^{\text{trans}}(p_{Y|x}\|p_Y)$ for each amino acid x — that is, the total height of the stacked bar plots in the figure — is equal to the specific MI $I(Y:X=x)$ for that x, as explained in the decomposition shown in Eq. (11).

While we do not dive deeply in the biological significant of these results, we highlight several interesting findings. First, for this PAM matrix and timescale ($\tau = 1$), a considerable fraction of the information ($\approx 1/3$) is transmitted not by copying but by non-random mutations. Generally, such non-random mutations represent underlying physical, genetic, and biological constraints that allow some pairs of amino acids to substitute each other more readily than other pairs.

Second, we observe considerable variation in the amount of specific MI, copy information, and transformation between different amino acids, as well as different the ratio of copy information to transformation information. In general, amino acids with more copy information are conserved unchanged

over evolutionary timescales. At the same time, it is known that conserved amino acids tend to be "outliers" in terms of their physiochemical properties (such as hydrophobicity, volume, polarity, etc.), since mutations to such outliers are likely to alter protein function in deleterious ways [56, 57]. To analyze this quantitatively, we used Miyata's measure of distance between amino acids, which is based on differences in volume and polarity [58]. For each amino acid, we quantified its degree of "outlierness" in terms of its mean Miyata distance to all 19 other amino acids. The Spearman rank correlation between this outlierness measure and copy information (as shown in Fig. 3) was 0.66 ($p = 0.0035$). On the other hand, the rank correlation between outlierness and transformation information was 0.02 ($p = 0.92$). Similar results were observed for other chemically-motivated measures of amino acid distance, such as Grantham's distance [59] and Sneath's index [60]. This demonstrates that amino acids with unique chemical characteristics tend to have more copy information, but not more transformation information.

VI. DISCUSSION

Although mutual information is a very common and successful measure of transmitted information, it is insensitive to the distinction between information that is transmitted by copying versus information that is transmitted by transformation. Nonetheless, as we have argued, this distinction is of fundamental importance in many real-world systems.

In this paper we propose a rigorous and practical way to decompose specific mutual information, and more generally Bayesian surprise, into two non-negative terms corresponding to copy and transformation, $I = I^{\mathsf{copy}} + I^{\mathsf{trans}}$. We derive our decomposition using an axiomatic framework: we propose a set of four axioms any measure of copy information should obey, and then identify the unique measure that satisfies those axioms. At the same time, we show that our measure of copy information is one of a family of functionals, each of which corresponds to a different way of quantifying error in transmission. We also demonstrate that our measures have a natural interpretation in thermodynamic terms, which suggests novel approaches for understanding the thermodynamic efficiency of biological replication processes, in particular DNA and RNA duplication. Finally, we demonstrate our results on real-world biological data, exploring copy and transformation information of amino acid substitution rates. We find significant variation among the amount of information transmitted by copying vs. transformation among different amino acids.

Several directions for future work present themselves. First, there is a large range of practical and theoretical application of our measures, ranging form analysis of biological and neural information transmission to the study of the thermodynamics of self-replication, a fundamental and challenging problem in biophysics [61]. Second, we suspect our measures of copy and transformation information have further connections to existing formal treatments in information theory, in particular rate-distortion theory [2], whose connections we started to explore here. We also believe that our decomposition may be generalizable beyond Bayesian surprise and mutual information to include other information-theoretic measures, including conditional mutual information and multi-information. Decomposing conditional mutual information is of particular interest, since it will permit a decomposition of the commonly-used *transfer entropy* [62] measure into copy and transformation components, thus separating two different modes of information flow between systems. Finally, our decomposition of mutual information into copy and transformation components has some high-level similarities to other recent proposals for non-negative decompositions of information measures, for example non-negative decompositions of multivariate information into redundant and synergistic components [63], integrated information decompositions [64, 65], and decompositions of mutual information into "semantic" (valuable) and "non-semantic" (non-valuable) information [66]. Further research should explore if and how our decomposition into copy and transformation information relates to these existing approaches.

ACKNOWLEDGMENTS

AK was supported by Grant No. FQXi-RFP-1622 from the FQXi foundation, and Grant No. CHE-1648973 from the U.S. National Science Foundation. AK would like to thank the Santa Fe Institute for supporting this research. The authors thank Jordi Fortuny, Rudolf Hanel, Joshua Garland, and Blai Vidiella for helpful discussions on the manuscript.

[1] Claude Elwood Shannon, "A mathematical theory of communication," The Bell System Technical Journal **27**, 379–423 (1948).
[2] Thomas M. Cover and Joy A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).
[3] J L Kelly, "A New Interpretation of Information Rate," The Bell System Technical Journal , 10 (1956).
[4] Andrew R. Barron and T. M. Cover, "A Bound on the Financial Value of Information," IEEE Transactions on Information Theory **34**, 1096–1101 (1988).
[5] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," IEEE Transactions on Information Theory **42**, 348–363 (1996).
[6] Matina C. Donaldson-Matasci, Carl T. Bergstrom, and Michael Lachmann, "The fitness value of information," Oikos **119**, 219–230 (2010).
[7] Takahiro Sagawa and Masahito Ueda, "Second law of thermodynamics with discrete quantum feedback control," Physical review letters **100**, 080403 (2008).

[8] Juan MR Parrondo, Jordan M. Horowitz, and Takahiro Sagawa, "Thermodynamics of information," Nature Physics **11**, 131–139 (2015).

[9] John Robinson Pierce, *An Introduction to Information Theory: Symbols, Signals & Noise* (Courier Corporation, 1961).

[10] Bernat Corominas-Murtra, Jordi Fortuny, and Ricard V Solé, "Towards a mathematical theory of meaningful communication," Scientific reports **4**, 4587 (2014).

[11] Carl T. Bergstrom and Martin Rosvall, "The transmission sense of information," Biology & Philosophy **26**, 159–176 (2011).

[12] Orion Penner, Peter Grassberger, and Maya Paczuski, "Sequence Alignment, Mutual Information, and Dissimilarity Measures for Constructing Phylogenies," PLOS ONE **6**, e14373 (2011).

[13] Franco L. Simonetti, Elin Teppa, Ariel Chernomoretz, Morten Nielsen, and Cristina Marino Buslje, "MISTIC: mutual information server to infer coevolution," Nucleic Acids Research **41**, W8–W14 (2013).

[14] Atul J. Butte and Isaac S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Biocomputing 2000* (World Scientific, 1999) pp. 418–429.

[15] Arun K. Ramani and Edward M. Marcotte, "Exploiting the Coevolution of Interacting Proteins to Discover Interaction Specificity," Journal of Molecular Biology **327**, 273–284 (2003).

[16] Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo, "Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications," Journal of Molecular Evolution **21**, 58–71 (1984).

[17] Barbara E Wright, "A biochemical mechanism for nonrandom mutations and evolution," Journal of bacteriology **182**, 2993–3001 (2000).

[18] Robert M Seyfarth, Dorothy L Cheney, and Peter Marler, "Vervet monkey alarm calls: semantic communication in a free-ranging primate," Animal Behaviour **28**, 1070–1094 (1980).

[19] James R Hurford, "Biological evolution of the saussurean sign as a component of the language acquisition device," Lingua **77**, 187–222 (1989).

[20] Martin A Nowak and David C Krakauer, "The evolution of language," Proceedings of the National Academy of Sciences **96**, 8028–8033 (1999).

[21] Angelo Cangelosi and Domenico Parisi, *Simulating the evolution of language* (Springer Science & Business Media, 2012).

[22] Natalia Komarova and Partha Niyogi, "Optimizing the mutual intelligibility of linguistic agents in a shared world," Artificial Intelligence **154**, 1–42 (2004).

[23] Partha Niyogi, *The computational nature of language learning and evolution* (MIT press Cambridge, MA, 2006).

[24] Luc Steels, "Evolving grounded communication for robots," Trends in cognitive sciences **7**, 308–312 (2003).

[25] Mariela Petkova, Gašper Tkačik, William Bialek, Eric F Wieschaus, and Thomas Gregor, "Optimal decoding of cellular identities in a genetic network," Cell **176**, 844–855.E15 (2019).

[26] Nicolaas Godfried Van Kampen, *Stochastic processes in physics and chemistry*, Vol. 1 (Elsevier, 1992).

[27] Hidde De Jong, "Modeling and simulation of genetic regulatory systems: a literature review," Journal of computational biology **9**, 67–103 (2002).

[28] Aage B Sorensen, "Mathematical models in sociology," Annual review of sociology **4**, 345–371 (1978).

[29] Claude E Shannon, "Coding theorems for a discrete source with a fidelity criterion," IRE Nat. Conv. Rec **4**, 1 (1959).

[30] Robert B. Ash, *Information Theory* (Courier Corporation, 2012).

[31] JJ Hopfield, "Physics, computation, and why biology looks so different," Journal of Theoretical Biology **171**, 53–60 (1994).

[32] The reader should note that, in the literature, the term "specific MI" can refer to two different measures [33]. The version of specific MI used here, as specified by Eq. (6), is also sometimes called "specific surprise".

[33] Michael R DeWeese and Markus Meister, "How to measure the information gained from one symbol," Network: Computation in Neural Systems **10**, 325–340 (1999).

[34] Laurent Itti and Pierre F. Baldi, "Bayesian Surprise Attracts Human Attention," in *Advances in Neural Information Processing Systems 18*, edited by Y. Weiss, B. Schölkopf, and J. C. Platt (MIT Press, 2006) pp. 547–554.

[35] Dennis V. Lindley, "On a measure of the information provided by an experiment," The Annals of Mathematical Statistics , 986–1005 (1956).

[36] M. Stone, "Application of a Measure of Information to the Design and Comparison of Regression Experiments," The Annals of Mathematical Statistics **30**, 55–70 (1959).

[37] José M. Bernardo, "Expected information as expected utility," The Annals of Statistics **7**, 686–690 (1979).

[38] Kathryn Chaloner and Isabella Verdinelli, "Bayesian Experimental Design: A Review," Statistical Science **10**, 273–304 (1995).

[39] Jose M. Bernardo, "Reference Posterior Distributions for Bayesian Inference," Journal of the Royal Statistical Society. Series B (Methodological) **41**, 113–147 (1979).

[40] James O. Berger, José M. Bernardo, and Dongchu Sun, "The formal definition of reference priors," The Annals of Statistics **37**, 905–938 (2009).

[41] R. Hanel and S. Thurner, "A comprehensive classification of complex statistical systems and an ab initio derivation of their entropy and distribution functions," Europhysics Letters **93**, 20006 (2011).

[42] S. Thurner, B. Corominas-Murtra, and R. Hanel, "The three faces of entropy for complex systems – information, thermodynamics and the maxent principle," Physical Review E **96**, 032124 (2017).

[43] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, Vol. 1 (Springer series in statistics New York, 2001).

[44] Yasemin Altun and Alex Smola, "Unifying divergence minimization and statistical inference via convex duality," in *International Conference on Computational Learning Theory* (Springer, 2006) pp. 139–153.

[45] Miroslav Dudík and Robert E Schapire, "Maximum entropy distribution estimation with generalized regularization," in *International Conference on Computational Learning Theory* (Springer, 2006) pp. 123–138.

[46] Oluwasanmi Koyejo and Joydeep Ghosh, "A representation approach for relative entropy minimization with expectation constraints," in *ICML WDDL workshop* (2013).

[47] Solomon Kullback, *Information Theory and Statistics* (John Riley and Sons, 1959).

[48] Jagat Narain Kapur and Hiremaglur K Kesavan, "Entropy optimization principles and their applications," in *Entropy and energy dissipation in water resources* (Springer, 1992) pp. 3–20.

[49] Amir Globerson, Eran Stark, Eilon Vaadia, and Naftali Tishby, "The minimum information principle and its application to neural code analysis," Proceedings of the National Academy of Sciences **106**, 3490–3495 (2009).

[50] Massimiliano Esposito and Christian Van den Broeck, "Three faces of the second law. i. master equation formulation," Physical

Review E **82**, 011143 (2010).
[51] Thomas E. Ouldridge and Pieter Rein ten Wolde, "Fundamental Costs in the Production and Destruction of Persistent Polymer Copies," Physical Review Letters **118** (2017), 10.1103/PhysRevLett.118.158103.
[52] Jenny Poulton, Pieter Rein ten Wolde, and Thomas E. Ouldridge, "Non-equilibrium correlations in minimal dynamical models of polymer copying," arXiv:1805.08502 [cond-mat] (2018), arXiv: 1805.08502.
[53] Ziheng Yang, "Estimating the pattern of nucleotide substitution," Journal of molecular evolution **39**, 105–111 (1994).
[54] Si Quang Le and Olivier Gascuel, "An improved general amino acid replacement matrix," Molecular biology and evolution **25**, 1307–1320 (2008).
[55] Stephanie J Spielman and Claus O Wilke, "Pyvolve: a flexible python module for simulating sequences along phylogenies," PloS one **10**, e0139047 (2015).
[56] Dan Graur, "Amino acid composition and the evolutionary rates of protein-coding genes," Journal of molecular evolution **22**, 53–62 (1985).
[57] Ziheng Yang, Rasmus Nielsen, and Masami Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution." Molecular biology and evolution **15**, 1600–1611 (1998).
[58] Takashi Miyata, Sanzo Miyazawa, and Teruo Yasunaga, "Two types of amino acid substitutions in protein evolution," Journal of Molecular Evolution **12**, 219–236 (1979).
[59] Richard Grantham, "Amino acid difference formula to help explain protein evolution," Science **185**, 862–864 (1974).
[60] PHA Sneath, "Relations between chemical structure and biological activity in peptides," Journal of theoretical biology **12**, 157–195 (1966).
[61] Bernat Corominas-Murtra, "Thermodynamics of duplication thresholds in synthetic protocell systems," Life **9**, 9 (2019).
[62] Thomas Schreiber, "Measuring information transfer," Phys. Rev. Lett. **85**, 461–464 (2000).
[63] Paul L Williams and Randall D Beer, "Nonnegative decomposition of multivariate information," arXiv preprint arXiv:1004.2515 (2010).
[64] Thomas Kahle, Eckehard Olbrich, Jürgen Jost, and Nihat Ay, "Complexity measures from interaction structures," Physical Review E **79**, 026201 (2009).
[65] Masafumi Oizumi, Naotsugu Tsuchiya, and Shun-ichi Amari, "Unified framework for information integration based on information geometry," Proceedings of the National Academy of Sciences **113**, 14817–14822 (2016).
[66] Artemy Kolchinsky and David H. Wolpert, "Semantic information, autonomous agency and non-equilibrium statistical physics," Interface Focus **8**, 20180041 (2018).
[67] Imre Csiszar and János Körner, *Information theory: coding theorems for discrete memoryless systems* (Cambridge University Press, 2011).

Appendix A: Derivation of Eq. 9

Consider the optimization problem:

$$\min_{r_Y \in \Delta : r_Y(x) \geq p_{Y|x}(x)} D_{\mathsf{KL}}(r_Y \| p_Y). \tag{A1}$$

It is clear that if $p_{Y|x}(x) \leq p_Y(x)$, then the solution $r_Y(y) = p_Y(y)$ for all y satisfies the constraint, and achieves $D_{\mathsf{KL}}(p_Y \| p_Y) = 0$, the minimum possible. We now consider the case when $p_{Y|x}(x) > p_Y(x)$. In this case, we use the chain rule for KL divergence [2] to write

$$\begin{aligned} &D_{\mathsf{KL}}(r_Y \| p_Y) \\ &= \mathsf{d}(r_Y(x), p_Y(x)) + D_{\mathsf{KL}}(r_Y(Y|Y \neq x) \| p_Y(Y|Y \neq x)). \end{aligned}$$

The second term is minimized by setting $r_Y(y) \propto p_Y(y)$ for $y \neq x$, so that $r_Y(y|Y \neq x) = p_Y(y|Y \neq x)$ and $D_{\mathsf{KL}}(r_Y(Y|Y \neq x) \| p_Y(Y|Y \neq x)) = 0$. Thus, in the case that $p_{Y|x}(x) > p_Y(x)$, we have reduced the optimization problem of Eq. (A1) to the equivalent problem

$$\min_{a \in \mathbb{R} : a \geq p_{Y|x}(x)} \mathsf{d}(a, p_Y(x)). \tag{A2}$$

Note that the derivative $\mathsf{d}(a, b)$ with respect to a is $\frac{d}{da}\mathsf{d}(a, b) = \log \frac{a}{b} - \log \frac{1-a}{1-b}$, which is strictly positive when $a > b$. Given the assumption that $p_{Y|x}(x) > p_Y(x)$, it must be that Eq. (A2) is minimized by setting $a = p_{Y|x}(x)$, the minimum possible. Thus, $\mathsf{d}(p_{Y|x}(x), p_Y(x))$ is the solution to Eq. (A1) when $p_{Y|x}(x) > p_Y(x)$.

Appendix B: D_x^{copy} satisfies the four axioms

It satisfies Axiom 1 by non-negativity of KL. It satisfies Axiom 2 when $p_{Y|x}(x) > p_Y(x)$ because $\mathsf{d}(p_{Y|x}(x), p_Y(x)) \leq D_{\mathsf{KL}}(p_{Y|x} \| p_Y)$ by the data processing inequality for KL divergence [67, Lemma 3.11]; otherwise, when $p_{Y|x}(x) \leq p_Y(x)$, D_x^{copy} vanishes and thus satisfies Axiom 2 trivially. It satisfies Axiom 3 because if $p_Y(x) \leq q_Y(x)$, then the minimization problem of Eq. (8) for q_Y is more constrained than for p_Y, so the minimum of the former must be larger than the minimum of the latter. Finally, to show that D_x^{copy} satisfies Axiom 4, for any prior distribution p_Y consider the following posterior distribution $p^{\alpha}_{Y|x}(y)$:

$$p^{\alpha}_{Y|x}(y) = \begin{cases} \alpha & \text{if } y = x \\ \frac{1-\alpha}{1-p_Y(x)} p_Y(y) & \text{if } y \neq x \end{cases}, \tag{B1}$$

where α is a parameter that can vary from 0 to 1. It is easy to verify that for all $\alpha \in [p_Y(x), 1]$,

$$D_{\mathsf{KL}}(p^{\alpha}_{Y|x} \| p_Y) = \mathsf{d}(\alpha, p_Y(x)) = D_x^{\mathsf{copy}}(p^{\alpha}_{Y|x} \| p_Y), \tag{B2}$$

and that $D_x^{\mathsf{copy}}(p^{\alpha}_{Y|x} \| p_Y)$ ranges in a continuous manner from 0 (for $\alpha = p_Y(x)$) to $-\log p_Y(x)$ (for $\alpha = 1$).

Appendix C: Proof of Theorem 1

Before proceeding, we first prove two useful lemmas.

Lemma C.1. *Given Axiom 3,* $F(p_{Y|x}, p_Y, x) = F(q_{Y|x}, p_Y, x)$ *if* $p_{Y|x}(x) = q_{Y|x}(x)$.

Proof. Follows from applying Axiom 3 twice. □

Lemma C.2. *Given Axioms 1 to 3, if $p_{Y|x}(x) \leq p_Y(x)$, then $F(p_{Y|x}, p_Y, x) = 0$.*

Proof. If $p_{Y|x}(x) \leq p_Y(x)$, then $F(p_{Y|x}, p_Y, x) \leq F(p_Y, p_Y, x)$ by Axiom 3. By Axiom 2, $F(p_Y, p_Y, x) \leq D_{\mathsf{KL}}(p_Y \| p_Y) = 0$. Combining gives $F(p_{Y|x}, p_Y, x) \leq 0$, while $F(p_{Y|x}, p_Y, x) \geq 0$ by Axiom 1. □

We then show that D_x^{copy} is the largest possible measure that satisfies Axioms 1 to 3.

Proposition C.1. *Any F which satisfies Axioms 1 to 3 must obey $F(p_{Y|x}, p_Y, x) \leq D_x^{\mathsf{copy}}(p_{Y|x} \| p_Y)$.*

Proof. Given Lemma C.2, without loss of generality we restrict our attention to the case where $p_{Y|x}(x) > p_Y(x)$. Define the posterior $p^{\alpha}_{Y|x}$ as in Eq. (B1), while taking $\alpha = p_{Y|x}(x)$. Then, by Lemma C.1,

$$F(p_{Y|x}, p_Y, x) = F(p^{\alpha}_{Y|x}, p_Y, x).$$

At the same time,

$$\begin{aligned} F(p^{\alpha}_{Y|x}, p_Y, x) &\leq D_{\mathsf{KL}}(p^{\alpha}_{Y|x} \| p_Y) \\ &= \mathsf{d}(p_{Y|x}(x) \| p_Y(x)) = D_x^{\mathsf{copy}}(p_{Y|x} \| p_Y), \end{aligned}$$

where the first inequality follows from Axiom 2, and the second equality from Eq. (B2). □

We are now ready to prove the main result from Section II B.

Proof of Theorem 1. Consider some $p_{Y|x}, p_Y, x$, and assume $p_{Y|x}(x) > p_Y(x)$ (without loss of generality by Lemma C.2). By Axiom 4, there must exist a posterior $q_{Y|x}$ such that $q_{Y|x}(x) = p_{Y|x}(x)$ and

$$F(q_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(q_{Y|x} \| p_Y). \tag{C1}$$

Note that by definition of D_x^{copy} (and in consistency with Axiom 2), $D_{\mathsf{KL}}(q_{Y|x} \| p_Y) \geq D_x^{\mathsf{copy}}(q_{Y|x} \| p_Y)$.

Then, by Lemma C.1, $F(p_{Y|x}, p_Y, x) = F(q_{Y|x}, p_Y, x)$ since $p_{Y|x}(x) = q_{Y|x}(x)$. Similarly, it can be verified that $D_x^{\mathsf{copy}}(q_{Y|x} \| p_Y) = D_x^{\mathsf{copy}}(p_{Y|x} \| p_Y)$. Combining the above results shows that $F(p_{Y|x}, p_Y, x) \geq D_x^{\mathsf{copy}}(p_{Y|x} \| p_Y)$. the theorem follows by combining with Proposition C.1.

□

Appendix D: Solution and axiomatic derivation of Eq. 18

In this appendix, we first show that Eq. (19) is the solution to Eq. (18), the generalized copy information measure considered in Section III.

First, note that if $\mathbb{E}_{p_{Y|x}}[\ell(x, Y)] \geq \mathbb{E}_{p_Y}[\ell(x, Y)]$, then the solution $r_Y = p_Y$ satisfies the constraint and achieves the smallest possible KL value of 0. Therefore, without loss of generality, we will consider the case when $\mathbb{E}_{p_{Y|x}}[\ell(x, Y)] < \mathbb{E}_{p_Y}[\ell(x, Y)]$. From the MaxEnt literature [48], it is known that the solution to the following equality-constrained optimization problem,

$$\min_{r_Y} D(r_Y \| p_Y) \quad \text{s.t.} \quad \mathbb{E}_{r_Y}[\ell(x, Y)] = \alpha, \tag{D1}$$

is given by

$$r(y) = \frac{q(y) e^{-\lambda \ell(x,y)}}{Z(\lambda)}, \tag{D2}$$

where $Z(\lambda)$ is the normalization constant and λ is a Lagrange multiplier that enforces the constraint. This can be seen as a 1-dimensional manifold of distributions parameterized by λ. We show that as λ increases, $\mathbb{E}_{r_Y}[\ell(x, Y)]$ monotonically decreases while $D(r_Y \| p_Y)$ monotonically increases. That means that solutions to the inequality-constrained optimization problem, Eq. (18), will make the constraint $\mathbb{E}_{r_Y}[\ell(x, Y)] = \alpha$ tight, and thus have the form of Eq. (D2). This recovers Eq. (19).

To show that $\mathbb{E}_{r_Y}[\ell(x, Y)]$ monotonically decreases while $D(r_Y \| p_Y)$ monotonically increases with λ, we first write the derivatives of $Z(\lambda)$ and $r(y)$ with respect to λ,

$$\begin{aligned} \tfrac{d}{d\lambda} Z(\lambda) &= \tfrac{d}{d\lambda} \sum_y q(y) e^{-\lambda \ell(x,y)} = -Z(\lambda) \mathbb{E}_{r_Y}[\ell(x, Y)] \\ \tfrac{d}{d\lambda} r(y) &= \frac{\mathbb{E}_{r_Y}[\ell(x, Y)] q(y) e^{-\lambda \ell(x,y)} - q(y) e^{-\lambda \ell(x,y)} \ell(x, y)}{Z(\lambda)} \\ &= r(y) \left[\mathbb{E}_{r_Y}[\ell(x, Y)] - \ell(x, y)\right] \end{aligned}$$

Clearly, the expectation of $\ell(x, y)$ decreases with λ,

$$\begin{aligned} \tfrac{d}{d\lambda} \mathbb{E}_{r_Y}[\ell(x, Y)] &= \sum_y r(y) \left[\mathbb{E}_{r_Y}[\ell(x, Y)] - \ell(x, y)\right] f(y) \\ &= \mathbb{E}_p[\ell(x, Y)]^2 - \mathbb{E}_p[\ell(x, Y)^2] \\ &= -\mathrm{Var}(\ell(x, Y)) \leq 0 \end{aligned}$$

At this point, note that $r_Y = p_Y$ when $\lambda = 0$. Since we are considering the case where $\mathbb{E}_{p_{Y|x}}[\ell(x, Y)] < \mathbb{E}_{p_Y}[\ell(x, Y)]$, it must be that $\lambda > 0$. We can then show that the KL term increases with λ,

$$\begin{aligned} \tfrac{d}{d\lambda} D(r_Y \| p_Y) &= \tfrac{d}{d\lambda} \left[\sum_y r(y) \ln \frac{r(y)}{q(y)}\right] \\ &= \sum_y \tfrac{d}{d\lambda} r(y) \left[\ln r(y) + 1 - \ln q(y)\right] \\ &= \sum_y r(y) \left[\mathbb{E}_{r_Y}[\ell(x, Y)] - \ell(x, y)\right] \left[-\ln Z(\lambda) - \lambda f(y) + 1\right] \\ &= \sum_y r(y) \left[\mathbb{E}_{r_Y}[\ell(x, Y)] - \ell(x, y)\right] \left[-\lambda f(y)\right] \\ &= \lambda \left[\mathbb{E}_{r_Y}[\ell(x, Y)^2] - \mathbb{E}_{r_Y}[\ell(x, Y)]^2\right] \\ &= \lambda \mathrm{Var}(\ell(x, Y)) \geq 0. \end{aligned}$$

We now demonstrate that the generalized copy information defined in Eq. (18), $\hat{F}(p_{Y|x}, p_Y, x)$, is the unique measure

that satisfies Axioms 1 and 2 and our modified Axioms 3* and 4*. Our derivation has the same structure as the one in Appendix C, and we proceed more quickly.

First, by applying applying Axiom 3* twice, it is clear that

$$\text{If } \mathbb{E}_{p_{Y|x}}[\ell(x,Y)] = \mathbb{E}_{q_{Y|x}}[\ell(x,Y)], \quad \text{then } F(p_{Y|x}, p_Y, x) = F(q_{Y|x}, p_Y, x) \tag{D3}$$

In addition, by Axioms 1 and 2, $0 \leq F(p_Y, p_Y, x) \leq D_{\mathsf{KL}}(p_Y \| p_Y) = 0$, so it must be that $F(p_Y, p_Y, x) = 0$. By Axiom 3*, that means that $F(p_{Y|x}, p_Y, x) = 0$ for any $p_{Y|x}$ with $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] \geq \mathbb{E}_{p_Y}[\ell(x,Y)]$ (i.e., whose loss is bigger than that of the prior distribution). In the rest of this appendix, we will assume without loss of generality that $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] < \mathbb{E}_{p_Y}[\ell(x,Y)]$.

We now show that $F(p_{Y|x}, p_Y, x) \leq \hat{F}(p_{Y|x}, p_Y, x)$. Given a choice of $p_{Y|x}$, p_Y, and x, let w_Y be the solution to Eq. (18), so

$$\hat{F}(p_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(w_Y \| p_Y). \tag{D4}$$

Given the derivations at the beginning of this appendix, we know that $\mathbb{E}_{w_Y}[\ell(x,Y)] = \mathbb{E}_{p_Y}[\ell(x,Y)]$, which implies

$$F(p_{Y|x}, p_Y, x) = F(w_Y, p_Y, x) \leq D_{\mathsf{KL}}(w_Y \| p_Y) = \hat{F}(w_Y, p_Y, x),$$

where the first equality uses Eq. (D3), the inequality uses Axiom 2, and the last equality uses Eq. (D4).

Finally, we show that $F(p_{Y|x}, p_Y, x) \geq \hat{F}(p_{Y|x}, p_Y, x)$. By Axiom 4*, there must exist a posterior $q_{Y|x}$ such that $\mathbb{E}_{q_{Y|x}}[\ell(x,Y)] = \mathbb{E}_{p_{Y|x}}[\ell(x,Y)]$ and

$$F(q_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(q_{Y|x} \| p_Y). \tag{D5}$$

Note that by definition of $\hat{F}$ (and in consistency with Axiom 2), $D_{\mathsf{KL}}(q_{Y|x} \| p_Y) \geq \hat{F}(q_{Y|x} \| p_Y)$. Then, by Eq. (D3), $F(p_{Y|x}, p_Y, x) = F(q_{Y|x}, p_Y, x)$. Similarly, it can be verified that $\hat{F}(p_{Y|x}, p_Y, x) = \hat{F}(q_{Y|x}, p_Y, x)$. Combining the above results shows that $F(p_{Y|x}, p_Y, x) \geq \hat{F}(p_{Y|x}, p_Y, x)$.

Thus, $\hat{F}$ is the only function that satisfies Axioms 1 and 2 and our modified Axioms 3* and 4*

Appendix E: Proof of Prop. 1

Before proving Proposition 1, we prove several intermediate results.

We start with a lemma that derives some useful properties of the roots of the quadratic polynomial $ax^2 - (a+s)x + sc$.

Lemma E.1. *Consider the function*

$$f_{\pm}(a,s,c) = \frac{a + s \pm \sqrt{(a+s)^2 - 4asc}}{2a} \tag{E1}$$

where $a \in \mathbb{R}$, $s \in (0,1]$, $c \in [0,1]$. Then,

1. *If $a < 0$, $f_+(a,s,c) \leq 0$ and the inequality is strict when $c > 0$. Otherwise, when $a \geq 0$, $f_+(a,s,c) \geq 1$, and the inequality is strict when $a = 0$.*
2. $f_-(a,s,c) \leq \begin{cases} \frac{s}{a} & \text{if } a \geq s \\ 1 & \text{otherwise} \end{cases}$.
3. $f_-(a,s,c) \geq \begin{cases} 1 - \frac{s}{|a|} & \text{if } a \leq -s \\ 0 & \text{otherwise} \end{cases}$.
4. $\lim_{a \to 0} f_-(a,s,c) = c$.
5. $\frac{\partial}{\partial a} f_-(a,s,c) \leq 0$, *and the inequality is strict when $c < 1$.*

Proof. Proof of 1. First consider $a < 0$, in which case $f_+(a,s,c) \leq f_-(a,s,c)$. Vieta's formula states that

$$f_-(a,s,c) f_+(a,s,c) = \frac{sc}{a} \leq 0. \tag{E2}$$

This implies that $f_+(a,s,c)$ must be less than 0. The inequality is strict when $c > 0$.

In the case $a \geq 0$, we lower bound the determinant,

$$(a+s)^2 - 4asc \geq a^2 + 2as + s^2 - 4as = (a-s)^2, \tag{E3}$$

This implies

$$f_+(a,s,c) \geq \frac{a + s + |a - s|}{2a} = \begin{cases} 1 & \text{if } a \geq s \\ \frac{s}{a} > 1 & \text{if } s > a \geq 0 \end{cases}$$

Proof of 2 and 3. When $a < 0$,

$$\begin{aligned} f_-(a,s,c) &= \frac{|a| - s + \sqrt{(s - |a|)^2 + 4sc|a|}}{2|a|} \\ &> \frac{|a| - s + ||a| - s|}{2|a|} = \begin{cases} 1 - \frac{s}{|a|} & \text{if } |a| \geq s \\ 0 & \text{otherwise} \end{cases} \\ f_-(a,s,c) &= \frac{|a| - s + \sqrt{(s - |a|)^2 + 4sc|a|}}{2|a|} \\ &\leq \frac{|a| - s + ||a| + s|}{2|a|} = 1, \end{aligned}$$

and the last inequality is strict when $c < 1$.

When $a > 0$,

$$\begin{aligned} f_-(a,s,c) &= \frac{a + s - \sqrt{(a+s)^2 - 4asc}}{2a} \\ &\geq \frac{a + s - |a + s|}{2a} = 0 \\ f_-(a,s,c) &= \frac{a + s - \sqrt{(a+s)^2 - 4asc}}{2a} \\ &\leq \frac{a + s - |a - s|}{2a} = \begin{cases} \frac{s}{a} & \text{if } a \geq s \\ 1 & \text{otherwise} \end{cases} \end{aligned}$$

where we've used Eq. (E3). Note that that inequality is strict when $c < 1$. Then, 2 and 3 follow by combining and simplifying.

Proof of 4. We use L'Hôpital's rule,

$$\begin{aligned}\lim_{a\to 0} f_-(a,s,c) &= \lim_{a\to 0} \frac{\frac{d}{da}\left(a+s-\sqrt{(a+s)^2-4asc}\right)}{\frac{d}{da}(2a)} \\ &= \frac{1}{2} - \lim_{a\to 0} \frac{2(a+s-4sc}{2\cdot 2\sqrt{(a+s)^2-4asc}} \\ &= \frac{1}{2} - \frac{s-2sc}{2s} = c.\end{aligned}$$

Proof of 5. If $c = 1$,

$$f_-(a,s,c) = \frac{a+s-|a-s|}{2a} = \begin{cases} \frac{s}{a} & \text{if } a \geq s \\ 1 & \text{otherwise} \end{cases}$$

where we've used the proof of 2 and 3. Thus, when $c = 1$, $\frac{\partial}{\partial a} f_-(a,s,c) = 0$ for $a < s$ and $\frac{\partial}{\partial a} f_-(a,s,c) < 0$ for $a > s$.

When $c < 1$, for notational convenience define

$$\eta := \sqrt{(a+s)^2 - 4asc} > 0.$$

(The inequality comes from the fact that Eq. (E3) is strict when $c < 1$). Now, we consider the derivative,

$$\begin{aligned}&\tfrac{\partial}{\partial a} f_-(a,s,c) \\ &= \frac{1}{4a^2}\left[\left(1 - \frac{1}{2}\frac{2a+2s-4sc}{\eta}\right)2a - 2\left(a+s-\eta\right)\right] \\ &= \frac{1}{2a^2}\left[\frac{a^2+sa-2sca}{\eta} - s + \eta\right] \\ &\overset{(1)}{\propto} -a^2 - sa + 2sca - s\eta + \eta^2 \\ &\overset{(2)}{=} sa - 2asc + s^2 - s\eta \\ &\overset{(3)}{\propto} \frac{a-2c+s}{\eta} - 1 \\ &= \sqrt{\frac{(a-2ac+s)^2}{\eta^2}} - 1 \\ &= \sqrt{1 - 4a^2c\frac{1-c}{\eta^2}} - 1 < 0,\end{aligned}$$

where in (1) we multiplied by the (positive) term $2a^2\eta$, in (2) we plugged in the definition of η and simplified, and in line (3) we divided by the (positive) term ηs. The inequality in the last line is strict since $0 < c < 1$.

□

Using Lemma E.1, we now prove the following.

Lemma E.2. *Let $c(x) \in [0,1]$ indicate a set of values for all $x \in \mathcal{A}$. Then, for any source distribution s_X with full support, there is a channel $p_{Y|X}$ that satisfies*

$$p(y|x) = \begin{cases} c(x) & \text{if } x = y \\ \frac{1-c(x)}{1-p_Y(x)} p_Y(y) & \text{otherwise,} \end{cases} \tag{E4}$$

where p_Y is the marginal $p_Y(y) = \sum_x s(x)p(y|x)$. The channel is unique if $c(x) > 0$ for all x.

Proof. We will show that there exists a marginal p_Y that satisfies the consistency conditions of Eq. (E4). The marginal is given by Eq. (E5), where a is a unique constant between 0 and 1 that makes $\sum_x p_Y(x) = 1$. Without loss of generality, we assume that there exists at least one x such that $c(x) < 1$ (otherwise, the result is trivially found by setting $p_{Y|X}$ to the identity map).

We now proceed in several steps. First, we plug Eq. (E4) into $p_Y(y) = \sum_x s(x)p(y|x)$ to write

$$\begin{aligned}p_Y(x) &= s(x)c(x) + p_Y(x)\sum_{x':x'\neq x} s(x')\frac{1-c(x')}{1-p_Y(x')} \\ 0 &= s(x)c(x) + p_Y(x)\left(-a - s(x)\frac{1-c(x)}{1-p_Y(x)}\right)\end{aligned}$$

where we've defined the constant

$$a := 1 - \sum_{x'} s(x')\frac{1-c(x')}{1-p_Y(x')}.$$

Multiplying both sides by $1 - p_Y(x)$ and simplifying gives

$$\begin{aligned}0 &= s(x)c(x) - s(x)c(x)p_Y(x) - ap_Y(x) + ap_Y(x)^2 - \\ &\quad [p_Y(x)s(x) - p_Y(x)s(x)c(x)] \\ &= ap_Y(x)^2 - (a+s(x))p_Y(x) + s(x)c(x).\end{aligned}$$

We solve this this quadratic equation for $p_Y(x)$,

$$p_Y^a(x) = \frac{a+s(x) - \sqrt{(a+s(x))^2 - 4a\,s(x)\,c(x)}}{2a}, \tag{E5}$$

where we include the superscript a in p_Y^a to make the dependence on a explicit, and we choose the negative solution of the quadratic equation. Our reason for choosing the negative solution are as follows:

- In case $c(x) = 0$, both branches may be valid and the solution is not necessarily unique, but for simplicity of analysis we choose the negative branch.
- In case $c(x) > 0$ and $a \leq 0$, then by Lemma E.1 1, only the negative branch gives valid probability values.
- In case $a > 0$, then (by arguments that follow below) it must be the case that $\sum_x c(x) > 1$, meaning that at least two $x', x'' \in \mathcal{A}$ have $c(x') > 0$ and $c(x'') > 0$. Then, since s_X has full support, both

$$p_Y(x') = \sum_x s(x)p(x'|x) \geq s(x')p(x'|x) = s(x')c(x) > 0$$

and (by a similar derivation) $p_Y(x'') > 0$, thus no $p_Y(x)$ can be equal to 1. Thus, again by Lemma E.1 1, only the negative branch is valid, since it allows $p_Y(x) < 1$.

Note that these considerations imply that the solution is unique when $c(x) > 0$ for all x.

We now find a^*, the value of the value of a that satisfies $\sum_x p_Y^a(x) = 1$. In other words, a^* is defined implicitly via:

$$\sum_x \frac{a^* + s(x) - \sqrt{(a^* + s(x))^2 - 4a^*\, s(x)\, c(x)}}{2a^*} = 1. \tag{E6}$$

By Lemma E.1 2, Lemma E.1 3, and Lemma E.1 4,

$$\sum_x p_Y^a(x)|_{a=-1} > \sum_x (1 - s(x)) = |\mathcal{A}| - 1 \geq 1$$
$$\lim_{a\to 0} \sum_x p_Y^a(x) = \sum_x c(x)$$
$$\sum_x p_Y^a(x)|_{a=1} \leq \sum_x s(x) = 1.$$

Moreover, $p_Y^a(x)$ is continuous over $a \in [-1, 0)$ and $(0, 1]$. This means that the a^* that satisfies Eq. (E6) must be:

1. within $[-1, 0)$ if $\sum_x c(x) < 1$,
2. equal to 0 if $\sum_x c(x) = 1$,
3. within $(0, 1]$ if $\sum_x c(x) > 1$.

(The first and last claims follow from the Intermediate Value Theorem.) Finally, using Lemma E.1 5, we have

$$\tfrac{d}{da} \textstyle\sum_x p_Y^a(x) < 0\,,$$

where we've used the assumption that at least one x has $c(x) < 1$. Thus, the a^* that satisfies Eq. (E6) must be unique. □

In practice, the value a^* in the proof of Lemma E.2 can be found by a numerical root finding algorithm, or by trying values from -1 to 1 in small intervals and selecting the first value that makes the LHS of Eq. (E6) less than or equal to 1. The marginal p_Y and channel $p_{Y|X}$ can then be computed in closed form using Eqs. (E4) and (E5).

We now prove the following theorem.

Theorem E.1. *Let $c(x) \in [0, 1]$ indicate a set of values for all $x \in \mathcal{A}$, and s_X any source distribution with full support. If and only if $\sum_x c(x) \geq 1$, there is a channel $p_{Y|X}$ with diagonals $p_{Y|x}(x) = c(x)$ and $I_p(Y:X) = I_p^{\mathsf{copy}}(X \to Y)$.*

Proof. It is easy to verify that a given channel $p_{Y|X}$ will have given diagonals $c(x)$ and $I_p(Y:X) = I_p^{\mathsf{copy}}(X \to Y)$ if and only if it has the form of the channel constructed in Lemma E.2, and all $c(x) \geq p_Y(x)$ for all x.

We now show that $c(x) \geq p_Y(x)$ for all x if $\sum_x c(x) \geq 1$, and $c(x) < p_Y(x)$ otherwise. As in the proof of Lemma E.2, we assume that we are not in the trivial case where all $c(x) = 1$.

In the proof of Lemma E.2, we found that the a^* that satisfies Eq. (E6) obeys $a^* < 0$ if $\sum_x c(x) < 1$, and $a^* \geq 0$ otherwise. At the same time, by Lemma E.1 5, $p_Y^a(x)$ is strictly decreasing in a and, by Lemma E.1 5, $\lim_{a\to\infty} p_Y^a(x) = c(x)$. Thus, if $\sum_x c(x) < 1$ it must be that $a^* < 0$ and $p_Y(x) = p_Y^{a^*}(x) < c(x)$. If $\sum_x c(x) \geq 1$ then $a^* \geq 0$ and $p_Y(x) = p_Y^{a^*}(x) \geq c(x)$.

□

We are now ready to prove Proposition 1.

Proposition 1. *For any source distribution s_X with full support and $H(s_X) < \infty$, there exist channels p for all levels of mutual information $I_p(Y:X) \in [0, H(s_X)]$ such that $I_p^{\mathsf{copy}}(X \to Y) = I_p(Y:X)$.*

Proof. First, consider some source distribution s_X and mutual information level $I_p(Y:X) = k$. We begin by showing that $p_{Y|X}$, as defined in Lemma E.2 and Theorem E.1, is a continuous function of s_X and $\vec{c}$ (in this proof, we use the notation $\vec{c}$ to indicate the vector of values $c(x)$ for all $x \in \mathcal{A}$). Without loss of generality, we assume that $0 < k < H(s_X)$, since it is trivial to find solutions for these edge cases (e.g., the channel $p(y|x) = s_X(y)$ for $k = 0$ and the identity channel $p(y|x) = \delta(y, x)$ for $k = H(s_X)$).

Consider the channel $p_{Y|X}$ from Lemma E.2. This channel is defined in terms of s_X, $\vec{c}$, and a^*. In turn, a^* is implicitly defined in terms of s_X and $\vec{c}$ in Eq. (E6). We rewrite this implicit definition as

$$g(a^*, s_X, \vec{c}) := \sum_x \frac{a^* + s(x) - \sqrt{(a^* + s(x))^2 - 4a^*\, s(x)\, c(x)}}{2a^*} - 1 = 0. \tag{E7}$$

As shown in the proof of Lemma E.2, for a fixed s_X and $\vec{c}$ (assuming that $\sum_x c(x) > 1$ and there is at least one x with $c(x)$), $g(\cdot, s_X, \vec{c})$ is strictly decreasing, and thus invertible. By the implicit function theorem, that means that the a^* that satisfies Eq. (E7) is a continuous function of s_X and $\vec{c}$. Thus, $p_{Y|X}$ is a continuous function of s_X and $\vec{c}$.

Next, define a manifold of $c(x)$ values parameterized by $\alpha \in (0, 1)$ as

$$c_\alpha(x) = (1 - \alpha)(1/|\mathcal{A}|) + \alpha.$$

For a given source s_X and set of values $c_\alpha(x)$ for a given α, let $p_{Y|X}^\alpha$ indicate the corresponding channel, as defined in Lemma E.2. Observe that for $\alpha = 0$, $p_{Y|X}^\alpha$ is completely noisy and thus $I_{p^\alpha}(Y:X) = I_{p^\alpha}^{\mathsf{copy}}(Y:X) = 0$. At the same time, for $\alpha = 1$, the corresponding channel $p_{Y|X}^\alpha$ is the identity map, and it is easy to verify that $I_{p^\alpha}(Y:X) = I_{p^\alpha}^{\mathsf{copy}}(Y:X) = H(s_X)$.

Note that, for a fixed source distribution s_X, mutual information is continuous in the conditional probabilities specified $p_{Y|X}$. Clearly, $c_\alpha(x)$ is continuous in α and (as we've shown above) $p_{Y|X}^\alpha$ is continuous in $c_\alpha(x)$. Thus, the mutual information $I_{p^\alpha}(Y:X)$ is continuous in $\alpha \in (0, 1)$ (so that $\sum_x c_\alpha(x) > 1$), and must sweep all values in $(0, H(s_X))$ (for $\alpha \to 0$ and $\alpha \to 1$ respectively). Simultaneously, we know by Theorem E.1 that $I_{p^\alpha}(Y:X) = I_{p^\alpha}^{\mathsf{copy}}(Y:X)$. □

Appendix F: The binary symmetric channel

The BSC is a channel over a two-state space ($\mathcal{A} = \{0, 1\}$) parameterized by a "probability of error" $\epsilon \in [0, 1]$. The BSC can be represented in matrix form as

$$p^{\epsilon}_{Y|X} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}.$$

When $\epsilon = 0$, the BSC is a noiseless channel which copies the source without error. In this extreme case, MI is large, and we expect it to consist entirely of copy information. On the other hand, when $\epsilon = 1$, the BSC is a noiseless "inverted" channel, where messages are perfectly switched between the source and the destination. In this case, MI is again large, but we now expect it to consist entirely of transformation information. Finally, $\epsilon = 1/2$ defines a completely noisy channel, for which mutual information (and thus copy and transformation information) must be 0.

For simplicity, we assume a uniform source distribution, $s_X(0) = s_X(1) = 1/2$, which by symmetry implies a marginal probability $p_Y(0) = p_Y(1) = 1/2$ at the destination for any ϵ. For the BSC with this source distribution, Eq. (9) states that for both $x = 0$ and $x = 1$, $D^{\mathsf{copy}}_x(p^{\epsilon}_{Y|x} \| p_Y) = I_{p^{\epsilon}}(Y : X{=}x)$ and $D^{\mathsf{trans}}_x(p^{\epsilon}_{Y|x} \| p_Y) = 0$ when $\epsilon \leq 1/2$, and $D^{\mathsf{copy}}_x(p^{\epsilon}_{Y|x} \| p_Y) = 0$ and $D^{\mathsf{trans}}_x(p^{\epsilon}_{Y|x} \| p_Y) = I_{p^{\epsilon}}(Y : X{=}x)$ otherwise. Using the definition of the (total) copy and transformation components of total MI, Eqs. (13) and (14), it then follows that

$$I^{\mathsf{copy}}_{p^{\epsilon}}(X{\to}Y) = \begin{cases} I_{p^{\epsilon}}(Y : X) & \text{if } \epsilon \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$I^{\mathsf{trans}}_{p^{\epsilon}}(X{\to}Y) = \begin{cases} I_{p^{\epsilon}}(Y : X) & \text{if } \epsilon \geq 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

This confirms intuitions about the BSC discussed in the beginning of this section. The behavior of MI, $I^{\mathsf{copy}}(X{\to}Y)$ and $I^{\mathsf{trans}}(X{\to}Y)$ for the BSC with a uniform source distribution is shown visually in Fig. 2 of the main text.