INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Kolchinsky A, Corominas-Murtra B. 2020 Decomposing information into copying versus transformation. *J. R. Soc. Interface* **17**: 20190623.
http://dx.doi.org/10.1098/rsif.2019.0623

Received: 6 September 2019
Accepted: 2 January 2020

Subject Category:
Life Sciences–Physics interface

Subject Areas:
systems biology, biophysics, biomathematics

Keywords:
mutual information, replication, thermodynamics of replication, copy information

Author for correspondence:
Bernat Corominas-Murtra
e-mail: bernat.corominas-murtra@ist.ac.at

Electronic supplementary material is available online at http://doi.org/10.6084/m9.figshare.c.4808931.

THE ROYAL SOCIETY PUBLISHING

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, 3400 Klosterneuburg, Austria

AK, 0000-0002-3518-9208; BC-M, 0000-0001-9806-5643

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 mutational biases occur during genetic replication. 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 versus the information transmitted by transformation. We begin with a decomposition that applies when the source and destination of the channel have the same set of messages and a notion of message identity exists. We then generalize our decomposition to other kinds of channels, which can involve different source and destination sets and broader notions of similarity. In addition, we show that copy information can be interpreted as the minimal work needed by a physical copying process, which is relevant for understanding 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.

Significance statement

Analysing and understanding the flow of information is crucial in many fields, from biology to physics to artificial intelligence. In many situations, it is crucial to disentangle how much information is transmitted by exact copying and how much information is transmitted by systematic transformations, a distinction that is not captured by standard information-theoretic measures. Here, we derive such a decomposition by starting from a few simple and intuitive assumptions. Our decomposition is easy to compute and has fundamental interpretations in terms of the thermodynamic costs of physical replication. Our measures apply to any information-processing scenario in which the faithfulness of the copy, not just the overall amount of mutual information, is of functional importance.

1. Introduction

Shannon's information theory provides a powerful set of tools for quantifying and analysing information transmission. A particular measure of interest is *mutual information* (MI), which is the most common way of quantifying the amount of information transmitted from a source to a destination. MI has fundamental interpretations and operationalizations in a variety of domains, ranging from telecommunications [1,2] to gambling and investment [3–5], biological evolution [6], statistical physics [7,8] and many others. Nonetheless, it has long been observed [9,10] that MI 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

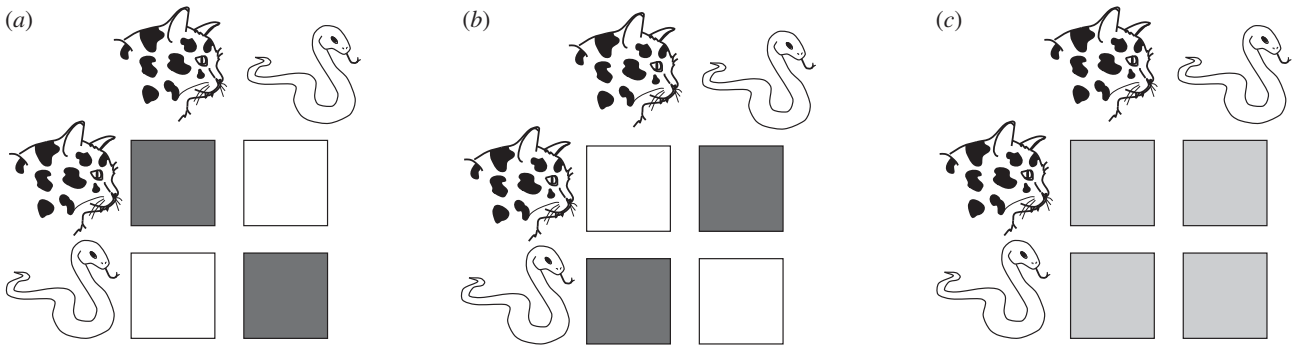

Figure 1. An illustration of the problem of copy and transformation. Consider three channels, each of which can transmit two messages, indicated by *cat* and *snake* (e.g. alarm calls in an animal communication system). In all panels, the rows indicate the message selected at the source, the columns 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 a given message at the destination does not depend on the message selected at the source, and the mutual information between source and destination is 0. Observe that transformation is different from noise, in that it still involves the transmission of information.

source message (where 'systematic' refers 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 modelled as a communication channel from parent to offspring, the amount of transmitted genetic information is often quantified by MI [11–15]. During replication, however, genetic information is not only copied but can also undergo systematic transformations in the form of non-random mutational biases. For instance, in the DNA of most organisms, $\mathtt{A} \leftrightarrow \mathtt{G}$ and $\mathtt{C} \leftrightarrow \mathtt{T}$ mutations occur more frequently than $\mathtt{A} \leftrightarrow \mathtt{C}$, $\mathtt{A} \leftrightarrow \mathtt{T}$, $\mathtt{G} \leftrightarrow \mathtt{C}$ and $\mathtt{G} \leftrightarrow \mathtt{T}$ mutations [16–18]. That means that some information about parent nucleotides is preserved even when those nucleotides undergo mutations. MI does not distinguish which part of genetic information is transmitted by exact copying and which part is transmitted by mutational biases. However, these two modes of information transmission are driven by different mechanisms and have dramatically different evolutionary and functional implications, given that mutations are more likely to lead to deleterious 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 figure 1, we provide a schematic that illustrates the problem. Essentially, we seek a decomposition of transmitted information into copy and transformation that distinguishes the example provided in figure 1*a*, where the copy is perfect, from the one provided in figure 1*b*, where the message has been systematically scrambled, from the one provided in figure 1*c*, where the channel is completely noisy. Of course, we also want 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 [10,19–25]. 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 MI between them may be high. As another example, the distinction between copying and transformation may be 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 [26]. In this scenario, information is copied when developmental signals are decoded correctly, and transformed when they are systematically decoded in an incorrect manner. Yet other examples are provided by Markov chain models, which are commonly used to study computation and other dynamical processes in physics [27], biology [28] or sociology [29], among other fields. In fact, a Markov chain 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 previously considered in the information-theoretic literature. This may be partly due to the different roles that information theory has historically played: on the 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 describing and analysing intrinsic properties of real-world systems. Because of its origins in engineering, much of information theory—including Shannon's channel-coding theorem, which established MI as a fundamental measure of transmitted information [2,30,31]—is formulated under the assumption of an external agent who can appropriately encode and decode information for transmission across 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 [10,32], and one is interested in quantifying the ability of a channel to copy information without any additional encoding or decoding. This latter problem 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' [2], which we indicate as ϵ. In particular, ϵ is the probability that the destination receives a different message from the one that was sent by the source, while $1-\epsilon$ is 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 quantify the propensity to transmit information by transformation, since ϵ increases both in the presence of transformation and in the presence of noise (in other words, ϵ is high both in a channel like figure 1*b* and in a channel like figure 1*c*). 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 measure an *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 fibre-optic line. In the language of physics, one might say that ϵ is an intensive property, rather than an 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.

In this paper, 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 satisfies these axioms. Our resulting measure is easy to compute and can be used to decompose either the total MI flowing across a channel, or the specific MI 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 §3, we 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 channels with different source and destination messages, to continuous-valued channels, and to other definitions of copying. We also discuss how our approach relates to *rate distortion* in information theory [2]. In §4, we show that our measure can be used to quantify the thermodynamic efficiency of physical copying processes, a central topic in biological physics. In §5, we demonstrate our measures on a real-world dataset of amino acid substitution rates.

2. Copy and transformation information

2.1. Preliminaries

We briefly present some basic concepts from information theory that will be useful for our further developments.

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 $p_Y, q_Y, \ldots \in \Delta$ to indicate marginal distributions over Y, and $p_{Y|x}, q_{Y|x}, \ldots \in \Delta$ 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 some distribution p over random variable X, we write the Shannon entropy as $H(p(X)) := -\sum_x p(x)\log p(x)$, or simply $H(X)$. For any two distributions s and q over the same set of outcomes, the *Kullback–Leibler* (KL) divergence is defined as

$$D_{\mathsf{KL}}(s\|q) := \sum_x s(x)\log\frac{s(x)}{q(x)}. \tag{2.1}$$

KL is non-negative and equal to 0 if and only if $s(x) = q(x)$ for all x. It is infinite when the support of s is not a subset of the support of q. In this paper, we will also 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,

$$\mathsf{d}(a, b) := a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}. \tag{2.2}$$

We will in general assume that logs are in base 2 (so 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 distribution. The amount of intrinsic noise in the channel, given some probability distribution of source messages $s_X(x)$, is the conditional Shannon entropy $H(Y\,|\,X) := -\sum_x s(x)\sum_y p(y|x)\log\, p(y|x)$. The amount of information transferred across a communication channel is quantified using the MI between the source and the destination [2],

$$I_p(Y:X) := \sum_x s(x)\sum_y p(y|x)\log\frac{p(y|x)}{p(y)}, \tag{2.3}$$

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

$$p(y) := \sum_x s(x)p(y|x). \tag{2.4}$$

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 decreases by a large amount, given the source message. MI can also be written as a weighted sum of so-called *specific MI*[1] terms [33–35], one for each outcome of X,

$$I(Y:X) = \sum_x s(x)I(Y:X{=}x), \tag{2.5}$$

where the specific MI for outcome x is given by

$$I(Y:X{=}x) := \sum_y p(y|x)\log\frac{p(y|x)}{p(y)} = D_{\mathsf{KL}}(p_{Y|x}\|p_Y). \tag{2.6}$$

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 equation (2.3), 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 $p_Y(y)$. The agent then updates their beliefs conditioned on the event $X = x$, resulting in the posterior distribution $p_{Y|x}$. The KL divergence between the posterior and the prior, $D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$ (equation (2.6)), is called *Bayesian surprise* [36], 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 [37–40] and the selection of 'non-informative priors' [41,42]. 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 equation (2.4). 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 communication channel $p_{Y|X}$.

Because Bayesian surprise 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 Bayesian surprise, $D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$. 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 exists 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 electronic supplementary material.

2.2. Axioms for copy information

We propose that any measure of copy information should satisfy a set of four axioms. Our set-up is motivated in the following way. First, our decomposition should apply at the level of the individual source message, i.e. we wish to be able to decompose each specific MI term (or, more generally, Bayesian surprise) into a non-negative *(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 $p_{Y|X}(x|x)$ (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'). This postulate can also be interpreted in a Bayesian way. Imagine two Bayesian agents with the same prior distribution over beliefs, p_Y, who update their beliefs conditioned on the event $X = x$. We postulate that the agent with the larger posterior probability on $Y = x$ should have greater copy information.

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,

$$F^{\mathsf{trans}}(p_{Y|x}, p_Y, x) := D_{\mathsf{KL}}(p_{Y|x}\|p_Y) - F(p_{Y|x}, p_Y, x). \tag{2.7}$$

We now propose a set of axioms that any measure of copy information F 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 equation (2.7), this guarantees that both F and F^{trans} are non-negative.

Axiom 2.1. $F(p_{Y|x}, p_Y, x) \geq 0$.

Axiom 2.2. $F(p_{Y|x}, p_Y, x) \leq D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$.

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 2.3. 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 electronic supplementary material, section B, we show that any measure of copy information that satisfies the above three axioms must obey $F(p_{Y|x}, p_Y, x) = 0$ whenever $p_{Y|x}(x) \leq p_Y(x)$. We also show that one particular measure of copy information, which is called D_x^{copy} and is discussed in the next section, is the largest measure that satisfies the above three axioms. However, the three axioms do not uniquely determine what happens when $p_{Y|x}(x) > p_Y(x)$. This means that D_x^{copy} is not unique, and, in fact, there are some trivial measures (such as $F(p_{Y|x}, p_Y, x) = 0$ for all $p_{Y|x}$, p_Y and x) that also satisfy the above axioms. Such trivial cases are excluded by our final axiom, which states that for all prior distributions and all posterior probabilities $p_{Y|x}(x) > p_Y(x)$, there are posterior distributions that contain *only* copy information. As we will see below, D_x^{copy} is the unique satisfying measure once this axiom is added.

Axiom 2.4. 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 $F(p_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$.

2.3. The measure D_x^{copy}

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

$$D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) = \begin{cases} \mathsf{d}(p_{Y|x}(x), p_Y(x)) & \text{if } p_{Y|x}(x) > p_Y(x) \\ 0 & \text{otherwise,} \end{cases} \tag{2.8}$$

where we have used the notation of equation (2.2). We now state the main result of our paper, which is as follows.

Theorem 2.5. *D_x^{copy} is the unique measure which satisfies axioms* 2.1–2.4.

In electronic supplementary material, section A, we demonstrate that D_x^{copy} satisfies all the axioms, and in electronic supplementary material, section B, we prove that it is the only measure that satisfies them. We further show that if one drops axiom 2.4, then D_x^{copy} is the largest possible measure that can satisfy the remaining axioms.

Given the definition of F^{trans} in equation (2.7), D_x^{copy} also defines a non-negative measure of transformation information, which we call D_x^{trans},

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

2.4. Decomposing mutual information

We now show that D_x^{copy} and D_x^{trans} allow for a decomposition of 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 equation (2.6). Each specific MI term can be seen as a Bayesian surprise, where the prior distribution is the marginal distribution at the destination (see equation (2.4)), and the posterior distribution is the conditional distribution of destination messages given a particular source message. Thus, our definitions of D_x^{copy} and D_x^{trans} provide a non-negative decomposition of each specific MI term,

$$I_p(Y:X{=}x) = D_x^{\text{copy}}(p_{Y|x}\|p_Y) + D_x^{\text{trans}}(p_{Y|x}\|p_Y). \tag{2.9}$$

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*,

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

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

$$I_p^{\text{copy}}(X \to Y) := \sum_x s(x) D_x^{\text{copy}}(p_{Y|x}\|p_Y) \tag{2.10}$$

and

$$I_p^{\text{trans}}(X \to Y) := \sum_x s(x) D_x^{\text{trans}}(p_{Y|x}\|p_Y). \tag{2.11}$$

(When writing I^{copy} and I^{trans}, we will often omit the subscript p when the channel is clear from the context.) By a simple manipulation, we can also decompose the marginal entropy of the destination $H(Y)$ into three non-negative components,

$$H(Y) = I^{\text{copy}}(X \to Y) + I^{\text{trans}}(X \to Y) + H(Y|X). \tag{2.12}$$

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.

For illustration purposes, we plot the behaviour of I^{copy} and I^{trans} in the classical binary symmetric channel (BSC) in figure 2 (see caption for details). More detailed analysis of copy and transformation information in the BSC is discussed in electronic supplementary material, section E.

It is worthwhile pointing out several important differences between our proposed measures and MI.

First, in the definitions of $I^{\text{copy}}(X \to Y)$ and $I^{\text{trans}}(X \to Y)$, the notation $X \to Y$ indicates that 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 $I^{\text{copy}}(X \to Y) \neq I^{\text{copy}}(Y \to X)$, and similarly for I^{trans}. We also note that the above forms of I^{copy} and I^{trans}, where they are written as sums over the individual source message, are sometimes referred to as the trace-like forms in the literature, and are commonly desired characteristics of information-theoretic functionals [43,44].

Second, I^{copy} and I^{trans} do not obey the data-processing inequality [2], and can either decrease or increase as the destination undergoes further operations. In this respect, they are different from MI (the sum of I^{copy} and I^{trans}). As an example, consider the case where channel $p_{Y|X}$ first transforms source message X into an encrypted message Y, and then another channel $p_{X'|Y}$ decrypts Y back into a copy of X (so $X' = X$). In this example, $I^{\text{copy}}(X \to X') > I^{\text{copy}}(X \to Y)$ even though the Markov condition $X - Y - X'$ holds.

Finally, unlike MI, I^{copy} and I^{trans} are generally non-additive when multiple independent channels are concatenated. As an example, imagine that the source messages are bit strings of length n, which are transmitted through a product of n independent channels, $p(y|x) = \prod_i p_i(y_i|x_i)$. If the source bits are independent, $s(x) = \prod_i s_i(x_i)$, it is straightforward to show that the MI between X and Y has the additive form $I(Y:X) = \sum_i I(Y_i:X_i)$. However, I^{copy} will generally not have this additive form, because copy information is defined in terms of the probability of exactly copying the entire source message (e.g. the entire n-bit-long string). Imagine that, in the above example, one of the bit-wise channels carries out a bit flip, $p_i(x_i|y_i) = 1 - \delta(x_i, y_i)$. In that case, the probability of receiving an exact copy of the source message at the destination is zero, and therefore I^{copy} is also zero regardless of the nature of the other bit-wise channels p_j for $j \neq i$. If desired, it is possible to derive an additive version of I^{copy} by generalizing our measure with an appropriate 'loss function', as discussed in more detail in §3 and electronic supplementary material, section C.3.

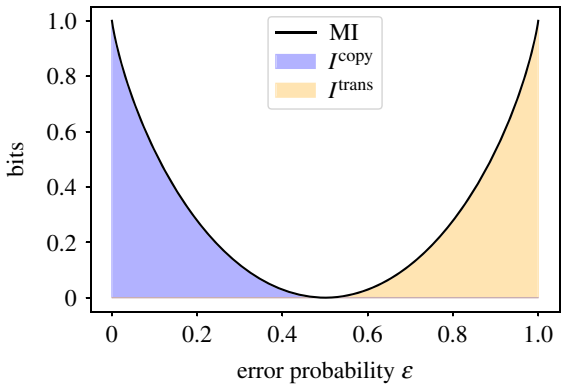

Figure 2. The binary symmetric channel (BSC) with a uniform source distribution. We plot values of the MI $I(Y:X)$, copy information $I^{\text{copy}}(X \to Y)$ (equation (2.10)) and transformation information $I^{\text{trans}}(X \to Y)$ (equation (2.11)) for the BSC along the whole range of error probabilities $\epsilon \in [0, 1]$. When $\epsilon \leq 1/2$, all mutual information is I^{copy} (blue shading); when $\epsilon \geq 1/2$, all mutual information is I^{trans} (orange shading). (Online version in colour.)

2.5. Copying efficiency

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

$$\eta_p(x) := \frac{D_x^{\text{copy}}(p_{Y|x}\|p_Y)}{D_{\text{KL}}(p_{Y|x}\|p_Y)} \in [0, 1], \tag{2.13}$$

where the bounds come directly from axioms 2.1 and 2.2. It can also be defined at the level of a channel as a whole as

$$\eta_p := \frac{I^{\text{copy}}(X \to Y)}{I(Y:X)} \in [0, 1]. \tag{2.14}$$

The bounds follow simply given the above results.

For equations (2.13) and (2.14) 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 equation (2.13), the bounds can be saturated because of axiom 2.4, which guarantees that, for any source message x, prior p_Y and desired posterior probability $p_{Y|x}(x) \geq p_Y(x)$, there exists a posterior $p_{Y|x}$ such that the Bayesian surprise $D_{\text{KL}}(p_{Y|x}\|p_Y)$ is composed entirely of copy information (for example, see electronic supplementary material, Eq. (A.1)).

One can show that the bounds in equation (2.14) can also be saturated. First, it can be verified that completely inefficient channels exist, since any channel which has $p_{Y|x}(x) \leq p_Y(x)$ for

all $x \in \mathcal{A}$ will have $I^{\text{copy}}(X \to Y) = 0$ (note that such channels exist at all levels of MI). We also show that maximally efficient channels exist, using the following result, which is proved in electronic supplementary material, section D.

Proposition 2.6. *For any source distribution s_X with $H(X) < \infty$, there exist channels $p_{Y|x}$ for all levels of mutual information $I_p(Y:X) \in [0, H(X)]$ such that $I_p^{\text{copy}}(X \to Y) = I_p(Y:X)$.*

Proposition 2.6 shows that it is possible to achieve all values of total copy information, which is defined at the level of a channel. Note that this proposition does not follow immediately from axiom 2.4, which is a statement about copy information at the level of a prior p_Y and posterior $p_{Y|x}$, where no particular relationship between p_Y and $p_{Y|x}$ is assumed.

3. Generalization and relation to rate distortion

We now 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'.

As we showed above, D_x^{copy} is the unique measure that satisfies the four axioms proposed in §2.2. In particular, it satisfies axiom 2.3, which states that, given the same prior p_Y, copy information should be larger for $q_{Y|x}$ than for $p_{Y|x}$ whenever $q_{Y|x}(x) \geq p_{Y|x}(x)$. It also satisfies axiom 2.4, which states that there exist posterior distributions that have only copy information for all possible $p_{Y|x}(x) \in [p_Y(x), 1]$.

These axioms are based on one particular definition of copying, which states that copying occurs when the source and destination messages match perfectly. In fact, this can be generalized to other definitions of copying and transformation by using a *loss function* $\ell(x, y)$, which quantifies the dissimilarity between source message x and destination message y. For a given loss function, $\ell(x, y) = 0$ indicates that x and y should be considered a perfect copy of each other, while $\ell(x, y) > 0$ indicates that x and y should be considered as somewhat different. Importantly, $\ell(x, y)$ can quantify similarity in a graded manner, so that $\ell(x, y') > \ell(x, y)$ indicates that y is closer to being a copy of x than y' (even though neither y nor y' may be a perfect copy of x).

Given an externally specified loss function $\ell(x, y)$, one can define axioms 2.3 and 2.4 in a generalized manner. The generalized version of axiom 2.3 states that posterior distribution $q_{Y|x}$ should have higher copy information than $p_{Y|x}$ whenever its expected loss is lower:

Axiom 3.1*. 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 2.4 states that, at all values of the expected loss which are lower than the expected loss achieved by p_Y, there are channels which transmit information only by copying.

Axiom 3.2*. For any p_Y and $c \in [\min_y \ell(x, y), \mathbb{E}_{p_Y}[\ell(x, Y)]]$, there exists a posterior distribution $p_{Y|x}$ such that $\mathbb{E}_{p_{Y|x}}[\ell(x, Y)] = c$ and $F(p_{Y|x}, p_Y, x) = D_{\text{KL}}(p_{Y|x} \| p_Y)$.

Note that, in defining axiom 3.2*, we used that $\min_y \ell(x, y)$ is the lowest expected loss that can be achieved by any posterior distribution.

Each particular loss function induces its own measure of copy information. In fact, as we show in electronic supplementary material, section C.1, there is a unique measure of copy information which satisfies axioms 2.1 and 2.2, as defined in §2.2, plus the generalized axioms 3.1* and 3.2*, as defined here in terms of the loss function $\ell(x, y)$. This generalized measure of copy information has the following form:

$$G_x^{\text{copy}}(p_{Y|x} \| p_Y) := \min_{r_Y} D_{\text{KL}}(r_Y \| p_Y) \tag{3.1}$$

$$\text{s.t.} \quad \mathbb{E}_{r_Y}[\ell(x, Y)] \leq \mathbb{E}_{p_{Y|x}}[\ell(x, Y)]. \tag{3.2}$$

Recall that the KL divergence $D_{\text{KL}}(r_Y \| p_Y)$ reflects the amount of information acquired by an agent in going from prior distribution p_Y to posterior distribution r_Y. Thus, $G_x^{\text{copy}}(p_{Y|x} \| p_Y)$ quantifies the minimum information that must be acquired by an agent in order to match the copying performance of the actual posterior $p_{Y|x}$, as measured by the expected loss.

Equation (3.1) is an instance of a 'minimum cross-entropy' problem, which is closely related to the 'maximum entropy' principle [45–47]. The distribution that optimizes equation (3.1) can be written in a simple form [48, pp. 299–300],

$$w(y) = \frac{1}{Z(\lambda)} p_Y(y)\, \mathrm{e}^{-\lambda \ell(x,y)},$$

where $\lambda \geq 0$ is a Lagrange multiplier chosen so that the constraint in equation (3.1) is satisfied, and $Z(\lambda) = \sum_y p_Y(y) e^{-\lambda \ell(x,y)}$ is a normalization constant. Note that, whenever $\mathbb{E}_{p_{Y|x}}[\ell(x, Y)] \geq \mathbb{E}_{p_Y}[\ell(x, Y)]$, $\lambda = 0$ and $w_Y = p_Y$ [48]. Otherwise, $\lambda > 0$ and the constraint in equation (3.1) will be tight up to equality. In practice, equation (3.1) can be solved by sweeping across the one-dimensional space of possible $\lambda \geq 0$ values (it can also be solved by standard convex optimization techniques). Once λ is determined, the value of copy information is given by

$$G_x^{\text{copy}} = -\lambda \mathbb{E}_{p_{Y|x}}[\ell(x, Y)] - \log Z(\lambda).$$

It can be verified that D_x^{copy}, the measure derived above, corresponds to the special case $\ell(x, y) := 1 - \delta(x, y)$, which is called '0–1 loss' in statistics [49] and 'Hamming distortion' in information theory [2] (see electronic supplementary material, section C.2).

The generalized measure G_x^{copy} has many similarities to D_x^{copy}. Like D_x^{copy}, it naturally leads to a non-negative measure of generalized transformation information,

$$G_x^{\text{trans}}(p_{Y|x} \| p_Y) = D_{\text{KL}}(p_{Y|x} \| p_Y) - G_x^{\text{copy}}(p_{Y|x} \| p_Y). \tag{3.3}$$

G_x^{copy} can also be used to decompose total MI into (generalized) total copy and transformation information, akin to equations (2.10) and (2.11). Finally, one can use G_x^{copy} to define a generalized measure of copying efficiency, following the approach described in §2.5.

While we believe 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, it may also be useful to consider other loss functions. One important example is when the source and destination have different sets of outcomes. Recall that D_x^{copy} assumes that the source and destination share the same set of possible outcomes, $\mathcal{A}$. When this assumption does not hold, generalized measures of copy and transformation information can still be defined, as long as an appropriate loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is provided

(where $\mathcal{X}$ and $\mathcal{Y}$ indicate the outcomes of the source and destination, respectively).

Another important use case occurs when the loss function specifies continuously varying degrees of functional similarity between source and destination messages. For example, imagine that $p_{Y|X}$ is an image compression algorithm which maps raw images X to compressed outputs Y. Research in computer vision has developed sophisticated loss functions for image compression which correlate strongly with human perceptual judgements [50]. By defining copy information in terms of such a loss function, one could measure how much perceptual information is copied by a particular image compression algorithm.

Our generalized approach can also be used to define copy and transformation information for random variables with continuous-valued outcomes. The 0–1 loss function, as used in D_x^{copy}, is not very meaningful for continuous-valued outcomes, since it depends on a measure-0 property of $p_{Y|x}$. A more natural measure of copy information is produced by the squared-error loss function $\ell(x, y) := (x - y)^2$, giving

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

This particular 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 [51–53].

Finally, it is also possible to generalize this approach to vector-valued loss functions $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^n$, which allow one to specify dissimilarity in a multi-dimensional way. We discuss the relevant axioms and resulting copy information measure for vector-valued loss functions in electronic supplementary material, section C.3. We also demonstrate that vector-valued loss functions can be used to define measures of copy and transformation information that are additive for independent channels, in the sense discussed in §2.4.

After what we have discussed so far, it is natural to briefly review the similarities between our generalized approach and *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 $r_{Y|X}$ which minimizes MI without exceeding some constraint on the expected distortion,

$$\min_{r_{Y|X}} D_{\mathsf{KL}}(r_{Y|X} \| r_Y) \quad \text{s.t.} \quad \mathbb{E}_r[\ell(X, Y)] \leq \alpha, \tag{3.4}$$

where α is an externally determined parameter. The prototypical application of rate distortion is compression, i.e. to find a compression channel $r_{Y|X}$ that has both low MI and low expected distortion. As can be seen by comparing equations (3.1) and (3.4), the optimization problem considered in our definition of generalized 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. In our approach, the aim is to decompose the information transmitted by a fixed externally specified channel into copy and transformation. In rate distortion, there is no externally specified channel and the aim is instead to find an optimal channel *de novo*. 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 which are 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.

We finish by noting that one can also define equation (3.1) in a channel-wise manner (by minimizing $D_{\mathsf{KL}}(r_{Y|X} \| r_Y)$, as in equation (3.4)) rather than a pointwise manner (minimize $D_{\mathsf{KL}}(r_{Y|X=x} \| p_Y)$, as in equation (3.1)). Under that formulation, one could no longer decompose specific MI into non-negative copy and information terms, though total MI could still be decomposed in that way. Interestingly, this alternative formulation would become equivalent to the so-called *minimum information principle*, a previous proposal for quantifying how much information about source messages is carried by different properties of destination messages [54].

4. 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 difference between information transfer by copying versus by transformation in a fundamental, biologically inspired physical set-up.

Consider a physical system coupled to a heat bath at temperature T, and which is initially in an equilibrium distribution $\pi(i) \propto e^{-E(i)/(kT)}$ with respect to some Hamiltonian E (k 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 [55],

$$W \geq kT \; D_{\mathsf{KL}}(p \| \pi). \tag{4.1}$$

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 analysed the fundamental thermodynamic constraints on copying in a physical system, for example for an information-carrying polymer like DNA [56,57]. Here, we will generally follow the model described in [56], 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 polymer to be copied), and the destination Y represents the state of the replicate (e.g. the polymer produced by the copying mechanism). We make several assumptions. First, the source X is not modified during the copying process. Second, X and Y have the same Hamiltonian before and after the copying process. Finally, we follow [56] in assuming that Y is a *persistent* copy of X, meaning that before and after the copying process Y is physically separated from X and there is no interaction

energy between them. This does not preclude X and Y from coming into contact and interacting energetically during intermediate stages of the copying process (for instance by template binding). The assumption of persistent copying means that there are no unaccounted energetic costs involved in preparing the copying system and transporting the produced replicate (e.g. moving the replicate Y to a daughter cell).

Assume that Y starts in the equilibrium distribution, indicated as π_Y (note that, by our persistent copy assumption, the equilibrium distribution cannot depend on the state of X). Let $p_{Y|x}(x)$ indicate the conditional distribution of replicates after the end of the copying process, where x is the state of the original system X. Following equation (4.1), the minimal work required to bring Y out of equilibrium and produce replicates according to $p_{Y|x}(x)$ is given by

$$W(x) \geq kT\, D_{\mathsf{KL}}(p_{Y|x} \| \pi_Y). \tag{4.2}$$

Note that equation (4.2) specifies the minimal work required to create the overall distribution $p_{Y|x}$. However, in many real-world scenarios, probably including DNA copying, the primary goal is to create exact copies of the original state, not transformed versions of it (such as non-random 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 a 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 begins and ends with the same equilibrium distribution, π_Y. The answer is given by the minimum of the RHS of equation (4.2) under a constraint on the exact-copy yield, which is exactly proportional to D_x^{copy},

$$W_{\min}^{\mathrm{exact}}(x) = kT \left[\min_{r_Y : r_Y(x) \geq p_{Y|x}(x)} D_{\mathsf{KL}}(r_Y \| \pi_Y) \right] \tag{4.3}$$

$$= kT\, D_x^{\mathsf{copy}}(p_{Y|x} \| \pi_Y), \tag{4.4}$$

where equation (4.4) follows from electronic supplementary material, section C.2. The additional work that is expended by the replication process is then lower bounded by a quantity proportional to D_x^{trans},

$$W(x) - W_{\min}^{\mathrm{exact}}(x) \geq kT D_x^{\mathsf{trans}}(p_{Y|x} \| \pi_Y). \tag{4.5}$$

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 (for example, the DNA replication system can copy a huge ensemble of source DNA sequences, not just one). Assume that X is distributed according to some $s_X(x)$. Across this 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

$$\langle W \rangle \geq kT \sum_x s(x) D_{\mathsf{KL}}(p_{Y|x} \| \pi_Y) \tag{4.6}$$

$$= kT[I_p(Y:X) + D_{\mathsf{KL}}(p_Y \| \pi_Y)]. \tag{4.7}$$

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 equation (4.4),

$$\langle W_{\min}^{\mathrm{exact}} \rangle = kT \sum_x s(x) D_x^{\mathsf{copy}}(p_{Y|x}(x) \| \pi_Y) \tag{4.8}$$

$$= kT[I_p^{\mathsf{copy}}(X \to Y) + D_{\mathsf{KL}}(p_Y \| \pi_Y)]. \tag{4.9}$$

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,

$$\langle W \rangle - \langle W_{\min}^{\mathrm{exact}} \rangle \geq kTI_p^{\mathsf{trans}}(X \to Y). \tag{4.10}$$

When the equilibrium distribution π_Y matches the marginal distribution p_Y, $\langle W_{\min}^{\mathrm{exact}} \rangle$ is exactly equal to kTI^{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 equation (2.14),

$$\frac{\langle W_{\min}^{\mathrm{exact}} \rangle}{\langle W \rangle} = \frac{I_p^{\mathsf{copy}}(X \to Y)}{I_p(Y:X)} = \eta_p. \tag{4.11}$$

As can be seen, standard information-theoretic measures, such as equation (4.2), 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 DNA copying that we discussed throughout this section, our 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 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 the 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 [17], as analysed 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.

5. Copy and transformation in amino acid substitution matrices

In the previous section, we saw how D_x^{copy} and I^{copy} arise 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 [17,58]. 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 *non-random* mutations that carry information but do not preserve the identity of the original message [16–18]. We also quantify the effect of purely random mutations, which correspond to the conditional entropy of the channel, $H(Y|X)$.

We demonstrate these measures on empirical data of *point accepted mutations* (PAMs) of amino acids. PAM data represent the rates of substitutions between different amino acids during the course of biological evolution, and have various applications, including evolutionary modelling, phylogenetic reconstructions and protein alignment [58]. We emphasize that amino acid PAM matrices do not reflect the direct physical transfer of information from protein to protein, but rather the effects of underlying processes of DNA-based replication and selection, followed by translation.

Formally, an amino acid PAM matrix Q is a continuous-time rate matrix. 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 particular PAM matrix Q which was published by Le & Gascuel [58] (this matrix was provided by the `pyvolve` Python package [59]). We calculated a discrete-time conditional probability distribution $p_{Y|X}$ from this matrix by computing 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 time scale $\tau = 1$. We used the stationary distribution of Q as the source distribution s_X, which correlates closely with empirically observed amino acid frequencies [58, fig. 1]. Using the decomposition presented in equation (2.11), we arrived at the following values for the communication channel described by the conditional probabilities $p_{Y|X}$:

$$I(Y:X) = I^{\text{copy}}(X \to Y) + I^{\text{trans}}(X \to Y) \approx 1.2\,\text{bits},$$

where

$$I^{\text{copy}}(X \to Y) = \sum_x s(x) D_x^{\text{copy}}(p_{Y|x} \| p_Y) \approx 0.88\,\text{bits}$$

and $$I^{\text{trans}}(X \to Y) = \sum_x s(x) D_x^{\text{trans}}(p_{Y|x} \| p_Y) \approx 0.32\,\text{bits}.$$

We also computed the intrinsic noise for this channel (see equation (2.12)),

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

Finally, we computed the specific copy and transformation information, D_x^{copy} and D_x^{trans}, for different amino acids. The results are shown in figure 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 of equation (2.9).

While we do not dive deeply in the biological significance of these results, we highlight several interesting findings. First, for this PAM matrix and time scale ($\tau = 1$), a considerable fraction of the information ($\approx 1/4$) 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

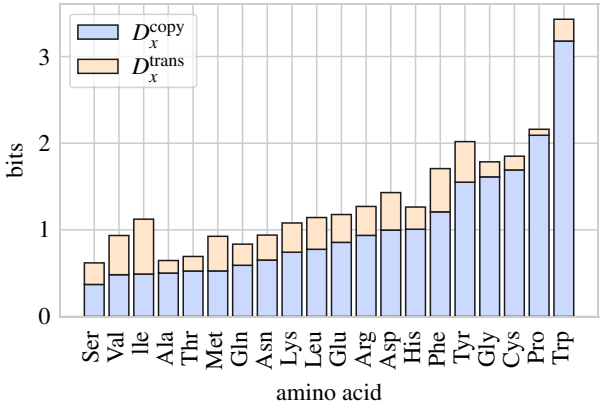

Figure 3. Copy and transformation information for different amino acids, based on an empirical PAM matrix [58]. We show magnitude of $D_x^{\text{copy}}(p_{Y|x} \| p_Y)$ in blue; on top of this is the amount of transformation information in orange. The sum of both is the specific MI for each amino acid, according to the decomposition given in equation (2.9). (Online version in colour.)

different amino acids, as well as different ratios of copy information to transformation information. In general, amino acids with more copy information are conserved unchanged over evolutionary time scales. At the same time, it is known that conserved amino acids tend to be 'outliers' in terms of their physio-chemical properties (such as hydrophobicity, volume, polarity, etc.), since mutations to such outliers are likely to alter protein function in deleterious ways [60,61]. To analyse this quantitatively, we used Miyata's measure of distance between amino acids, which is based on differences in volume and polarity [62]. 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 figure 3) was 0.57 ($p = 0.009$). On the other hand, the rank correlation between outlierness and transformation information was 0.22 ($p = 0.352$). Similar results were observed for other chemically motivated measures of amino acid distance, such as Grantham's distance [63] and Sneath's index [64]. This demonstrates that amino acids with unique chemical characteristics tend to have more copy information, but not more transformation information.

6. Discussion

Although MI 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 MI, and more generally Bayesian surprise, into two non-negative terms corresponding to copy and transformation, $I = I^{\text{copy}} + I^{\text{trans}}$. We derive our decomposition using an axiomatic framework: we propose a set of four axioms that 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 versus 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, from analysis of biological and neural information transmission to the study of the thermodynamics of self-replication, a fundamental and challenging problem in biophysics [65].

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 MI to include other information-theoretic measures, including conditional MI and multi-information. Decomposing conditional MI is of particular interest, since it will permit a decomposition of the commonly used *transfer entropy* [66] measure into copy and transformation components, thus separating two different modes of dynamical information flow between systems.

Finally, we point out that our proposed decomposition has some high-level similarities to other recent proposals for information-theoretic decomposition, such as the 'partial information decomposition' of multivariate information into redundant and synergistic components [67], integrated information decompositions [68,69] and decompositions of MI into 'semantic' (valuable) and 'non-semantic' (non-valuable) information [70]. We also mention another recent proposal for an alternative information-theoretic notion of 'copying' [71], in which copying is said to occur in a multivariate system when information that is present in one variable spreads to other variables (regardless of any transformations that information may undergo). Further research should explore if and how the decomposition proposed in this paper relates to these other approaches.

Data accessibility. Data used for §5 comes from [58], provided by the `pyvolve` Python package [59].

Authors' contributions. A.K. and B.C.-M. contributed equally to the manuscript.

Competing interests. We declare we have no competing interests.

Funding. B.C.-M. was supported by the Institute for Science and Technology-Austria. A.K. was supported by grant no. FQXi-RFP-1622 from the FQXi foundation, and grant no. CHE-1648973 from the US National Science Foundation.

Acknowledgements. A.K. thanks the Santa Fe Institute for supporting this research. B.C.-M. thanks the Institute for Science and Technology-Austria for supporting his research. The authors thank Jordi Fortuny, Rudolf Hanel, Joshua Garland and Blai Vidiella for helpful discussions, as well as the anonymous reviewers for their insightful suggestions.

Endnote

[1]Readers should be aware that the term 'specific MI' has been used to refer to two different measures in the literature [33]. The version of specific MI used here, as specified by equation (2.6), is also sometimes called 'specific surprise'.

References

1. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Technical J.* **27**, 379–423. (doi:10.1002/j.1538-7305.1948.tb01338.x)
2. Cover TM, Thomas JA. 2006 *Elements of information theory*. New York, NY: John Wiley & Sons.
3. Kelly JL. 1956 A new interpretation of information rate. *Bell Syst. Technical J.* **35**, 917–926.
4. Barron AR, Cover TM. 1988 A bound on the financial value of information. *IEEE Trans. Inf. Theory* **34**, 1096–1101. (doi:10.1109/18.21241)
5. Cover TM, Ordentlich E. 1996 Universal portfolios with side information. *IEEE Trans. Inf. Theory* **42**, 348–363. (doi:10.1109/18.485708)
6. Donaldson-Matasci MC, Bergstrom CT, Lachmann M. 2010 The fitness value of information. *Oikos* **119**, 219–230. (doi:10.1111/j.1600-0706.2009.17781.x)
7. Sagawa T, Ueda M. 2008 Second law of thermodynamics with discrete quantum feedback control. *Phys. Rev. Lett.* **100**, 080403. (doi:10.1103/PhysRevLett.100.080403)
8. Parrondo JMR, Horowitz JM, Sagawa T. 2015 Thermodynamics of information. *Nat. Phys.* **11**, 131–139. (doi:10.1038/nphys3230)
9. Pierce JR. 1980 *An introduction to information theory: symbols, signals & noise*. New York, NY: Dover.
10. Corominas-Murtra B, Fortuny J, Solé RV. 2014 Towards a mathematical theory of meaningful communication. *Sci. Rep.* **4**, 4587. (doi:10.1038/srep04587)
11. Bergstrom CT, Rosvall M. 2011 The transmission sense of information. *Biol. Phil.* **26**, 159–176. (doi:10.1007/s10539-009-9180-z)
12. Penner O, Grassberger P, Paczuski M. 2011 Sequence alignment, mutual information, and dissimilarity measures for constructing phylogenies. *PLoS ONE* **6**, e14373. (doi:10.1371/journal.pone.0014373)
13. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. 2013 MISTIC: mutual information server to infer coevolution. *Nucleic Acids Res.* **41**, W8–W14. (doi:10.1093/nar/gkt427)
14. Butte AJ, Kohane IS. 1999 Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pp. 418–429. Singapore: World Scientific.
15. Ramani AK, Marcotte EM. 2003 Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284. (doi:10.1016/S0022-2836(03)00114-1)
16. Li W-H, Wu C-I, Luo C-C. 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58–71. (doi:10.1007/BF02100628)
17. Yang Z. 1994 Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111.
18. Graur D, Li W-H. 2000 *Fundamentals of molecular evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.
19. Seyfarth RM, Cheney DL, Marler P. 1980 Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Anim. Behav.* **28**, 1070–1094. (doi:10.1016/S0003-3472(80)80097-2)
20. Hurford JR. 1989 Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187–222. (doi:10.1016/0024-3841(89)90015-6)
21. Nowak MA, Krakauer DC. 1999 The evolution of language. *Proc. Natl Acad. Sci. USA* **96**, 8028–8033. (doi:10.1073/pnas.96.14.8028)
22. Cangelosi A, Parisi D. 2002 *Simulating the evolution of language*. London, UK: Springer Science & Business Media.
23. Komarova N, Niyogi P. 2004 Optimizing the mutual intelligibility of linguistic agents in a shared world. *Artif. Intell.* **154**, 1–42. (doi:10.1016/j.artint.2003.08.005)
24. Niyogi P. 2006 *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.
25. Steels L. 2003 Evolving grounded communication for robots. *Trends Cogn. Sci.* **7**, 308–312. (doi:10.1016/S1364-6613(03)00129-3)

26. Petkova M, Tkačik G, Bialek W, Wieschaus EF, Gregor T. 2019 Optimal decoding of cellular identities in a genetic network. *Cell* **176**, 844–855.E15. (doi:10.1016/j.cell.2019.01.007)
27. Van Kampen NG. 1992 *Stochastic processes in physics and chemistry*, vol. 1. Amsterdam, The Netherlands: Elsevier.
28. Jong HD. 2002 Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103. (doi:10.1089/10665270252833208)
29. Sorensen AB. 1978 Mathematical models in sociology. *Annu. Rev. Sociol.* **4**, 345–371. (doi:10.1146/annurev.so.04.080178.002021)
30. Shannon CE. 1959 Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.* **4**, 1.
31. Ash RB. 1990 *Information theory*. New York, NY: Dover.
32. Hopfield JJ. 1994 Physics, computation, and why biology looks so different. *J. Theor. Biol.* **171**, 53–60. (doi:10.1006/jtbi.1994.1211)
33. DeWeese MR, Meister M. 1999 How to measure the information gained from one symbol. *Network: Comput. Neural Syst.* **10**, 325–340. (doi:10.1088/0954-898X_10_4_303)
34. Butts DA. 2003 How much information is associated with a particular stimulus? *Network* **14**, 177–187. (doi:10.1088/0954-898X_14_2_301)
35. Wibral M, Lizier JT, Priesemann V. 2015 Bits from brains for biologically inspired computing. *Front. Rob. AI* **2**, 5. (doi:10.3389/frobt.2015.00005)
36. Itti L, Baldi PF. 2006 Bayesian surprise attracts human attention. In *Advances in neural information processing systems 18* (eds Y Weiss, B Schölkopf, JC Platt), pp. 547–554. Cambridge, MA: MIT Press.
37. Lindley DV. 1956 On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005. (doi:10.1214/aoms/1177728069)
38. Stone M. 1959 Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Stat.* **30**, 55–70. (doi:10.1214/aoms/1177706359)
39. Bernardo JM. 1979 Expected information as expected utility. *Ann. Stat.* **7**, 686–690. (doi:10.1214/aos/1176344689)
40. Chaloner K, Verdinelli I. 1995 Bayesian experimental design: a review. *Stat. Sci.* **10**, 273–304. (doi:10.1214/ss/1177009939)
41. Bernardo JM. 1979 Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc. Ser. B (Methodological)* **41**, 113–147.
42. Berger JO, Bernardo JM, Sun D. 2009 The formal definition of reference priors. *Ann. Stat.* **37**, 905–938. (doi:10.1214/07-AOS587)
43. Hanel R, Thurner S. 2011 A comprehensive classification of complex statistical systems and an ab initio derivation of their entropy and distribution functions. *Europhys. Lett.* **93**, 20006. (doi:10.1209/0295-5075/93/20006)
44. Thurner S, Corominas-Murtra B, Hanel R. 2017 The three faces of entropy for complex systems—information, thermodynamics and the maximum entropy principle. *Phys. Rev. E* **96**, 032124. (doi:10.1103/PhysRevE.96.032124)
45. Kullback S. 1959 *Information theory and statistics*. New York, NY: John Wiley & Sons.
46. Kapur JN, Kesavan HK. 1992 Entropy optimization principles and their applications. In *Entropy and energy dissipation in water resources*, pp. 3–20. Dordrecht, The Netherlands: Springer.
47. Shore J, Johnson R. 1981 Properties of cross-entropy minimization. *IEEE Trans. Inf. Theory* **27**, 472–482. (doi:10.1109/TIT.1981.1056373)
48. Rubinstein RY, Kroese DP. 2017 *Simulation and the Monte Carlo method*. Hoboken, NJ: John Wiley & Sons.
49. Friedman J, Hastie T, Tibshirani R 2001 *The elements of statistical learning*, vol. 1. Springer Series in Statistics: New York, NY: Springer.
50. Wang Z, Bovik AC. 2006 Modern image quality assessment. *Synth. Lect. Image Video Multimedia Process.* **2**, 1–156. (doi:10.2200/S00010ED1V01Y200508IVM003)
51. Altun Y, Smola A. 2006 Unifying divergence minimization and statistical inference via convex duality. In *Proc. 19th Annu. Conf. on Learning Theory, COLT 2006, Pittsburgh, PA, 22–25 June 2006*, pp. 139–153. Berlin, Germany: Springer.
52. Dudík M, Schapire RE. 2006 Maximum entropy distribution estimation with generalized regularization. In *Proc. 19th Annu. Conf. on Learning Theory, COLT 2006, Pittsburgh, PA, 22–25 June 2006*, pp. 123–138. Berlin, Germany: Springer.
53. Koyejo O, Ghosh J. 2013 A representation approach for relative entropy minimization with expectation constraints. In *Proc. of the 30th Int. Conf. on Machine Learning, Atlanta, GA, 16–21 June 2013*. JMLR: W&CP, vol. 28.
54. Globerson A, Stark E, Vaadia E, Tishby N. 2009 The minimum information principle and its application to neural code analysis. *Proc. Natl Acad. Sci. USA* **106**, 3490–3495. (doi:10.1073/pnas.0806782106)
55. Esposito M, Van den Broeck C. 2010 Three faces of the second law. I. Master equation formulation. *Phys. Rev. E* **82**, 011143. (doi:10.1103/PhysRevE.82.011143)
56. Ouldridge TE, ten Wolde PR. 2017 Fundamental costs in the production and destruction of persistent polymer copies. *Phys. Rev. Lett.* **118**, 158103. (doi:10.1103/PhysRevLett.118.158103)
57. Poulton J, ten Wolde PR, Ouldridge TE. 2018 Non-equilibrium correlations in minimal dynamical models of polymer copying. (http://arxiv.org/abs/1805.08502)
58. Le SQ, Gascuel O. 2008 An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320. (doi:10.1093/molbev/msn067)
59. Spielman SJ, Wilke CO. 2015 Pyvolve: a flexible Python module for simulating sequences along phylogenies. *PLoS ONE* **10**, e0139047. (doi:10.1371/journal.pone.0139047)
60. Graur D. 1985 Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* **22**, 53–62. (doi:10.1007/BF02105805)
61. Yang Z, Nielsen R, Hasegawa M. 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611. (doi:10.1093/oxfordjournals.molbev.a025888)
62. Miyata T, Miyazawa S, Yasunaga T. 1979 Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236. (doi:10.1007/BF01732340)
63. Grantham R. 1974 Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864. (doi:10.1126/science.185.4154.862)
64. Sneath PHA. 1966 Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195. (doi:10.1016/0022-5193(66)90112-3)
65. Corominas-Murtra B. 2019 Thermodynamics of duplication thresholds in synthetic protocell systems. *Life* **9**, 9. (doi:10.3390/life9010009)
66. Schreiber T. 2000 Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464. (doi:10.1103/PhysRevLett.85.461)
67. Williams PL, Beer RD. 2010 Nonnegative decomposition of multivariate information. (http://arxiv.org/abs/1004.2515)
68. Kahle T, Olbrich E, Jost J, Ay N. 2009 Complexity measures from interaction structures. *Phys. Rev. E* **79**, 026201. (doi:10.1103/PhysRevE.79.026201)
69. Oizumi M, Tsuchiya N, Amari S-i. 2016 Unified framework for information integration based on information geometry. *Proc. Natl Acad. Sci. USA* **113**, 14 817–14 822. (doi:10.1073/pnas.1603583113)
70. Kolchinsky A, Wolpert DH. 2018 Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* **8**, 20180041. (doi:10.1098/rsfs.2018.0041)
71. Mediano PAM, Rosas F, Carhart-Harris RL, Seth AK, Barrett AB. 2019 Beyond integrated information: a taxonomy of information dynamics phenomena. (http://arxiv.org/abs/1909.02297).