Conserved active information

Yanchen Chen, Daniel Andrés Díaz-Pachón, *Member, IEEE,*

***Abstract*—We introduce conserved active information $I^\oplus$, a symmetric extension of active information that quantifies net information gain/loss across the entire search space, respecting No-Free-Lunch conservation. Using Bernoulli and uniform-baseline examples, we show that $I^\oplus$ reveals regimes hidden by KL divergence, such as when strong knowledge reduces global disorder. Such regimes are formally proven under a uniform baseline, distinguishing disorder-increasing mild knowledge from order-imposing strong knowledge. We further illustrate these regimes with examples from Markov chains and cosmological fine-tuning. This resolves a longstanding critique of active information while enabling applications in search, optimization, and beyond.**

***Index Terms*—active information, conservation of information, No-Free-Lunch Theorems.**

I. Introduction

Information theory imposes and is subject to highly consequential limitations. Among the most striking is its self-imposed focus on averages (entropy, Kullback-Leibler divergence, mutual information, etc.). The deep roots of information theory in thermodynamics, the poster child of reductionism, readily explain this trend. Reductionism, the philosophical position that intends to explain the whole as *the sum of its parts*—precisely the definition of average, garnered many adherents because of the spectacular results it brought about. Indeed, thermodynamics has successfully characterized macroscopic phenomena of systems close to equilibrium in terms of their microscopic constituents, and it holds the well-deserved title of the most undisputed theory in science [1], [2].

However, the thermodynamics of equilibrium has a problem: it is incomplete and unsatisfying in most real-life scenarios. As emergent properties accumulated rapidly in the basic sciences, it became impossible to analyze most macrostates as the sum of their microstates, implying that simplifying theory-of-everything-type equations proved insufficient in most interesting cases. Consequently, even Nobelists began to question reductionism some decades ago; for instance, Philip Anderson realized that more is different [3] and Robert Laughlin saw that a theory of everything is hardly a theory of every thing, opening the way to a hierarchical approach [4]. In all, this realization gave rise to the theory of complex systems.

Since statistics also originated in physics (as evidenced by terms like "mass", "density", or "moments"), it comes as no surprise that it has suffered from the same illnesses as information theory. Indeed, the focus on averages has made some problems extremely difficult, if not impossible, to solve. For instance, Vladimir Vapnik, the father of statistical learning, made clear that problems such as the curse of dimensionality, real-life data that refusal to fit a few parametric distributions, and the shortcomings of the maximum likelihood method (closely linked to the Gibbs distribution in statistical physics [5], [6]) imposed strong limitations on the classical approach to statistics. Even robustness, the assumption that led to treating the average as the typical behavior of all the individual parts, became dubious; examples include bump hunting [7]–[11] and density estimation [12]–[14], where assigning the same weight to every state or singleton is useless. Such considerations led Vapnik to develop his highly successful nonparametric theory of statistical learning [15], [16].

In biology, even amid deep disagreements about definitions of life, a fundamental property that cannot be overlooked is homeostasis, the tendency of living entities to resist entropy, which seems to defy all we know about thermodynamics [17], [18]. Therefore, explaining, for example, the origin of life through equilibrium physics is doomed to failure. The reductionist approach grossly underestimates the information content of far-from-equilibrium systems, which include most biological events of interest [19]–[21]. This includes all events for which we have a single observation, such as abiogenesis, eukaryogenesis, the emergence of sexual reproduction, and the emergence of language and intelligence [22]. For this reason, Nobel laureate Jack Szostak proposed functional information—a local measure that focuses exclusively on the fraction of functional configurations in a complex system [23], [24].

Events that, like functional configurations, are highly specific and informative are called *specifications* here. These specifications abound across most fields of science and, by their very nature, cannot be reduced to mean-field assumptions. For instance, in medicine, there appear to be close links between eukaryogenesis and the emergence of intelligence and cognition in cells, and these, in turn, have been linked to regenerative medicine and prospective cures for cancer, among many others [25]–[29]. In cosmology, singularities include the formation and behavior of black holes [2], [30], the origin of the universe [31], [32], and the particular values assumed by the constants of nature and the boundary conditions of the standard models [33]–[36].

In machine learning and artificial intelligence, the study of bias and fairness is precisely the study of these specifications [37]–[40]. For instance, this occurs when a feature of an undersampled group is evaluated using results from an oversampled group—a common situation in health disparity research [41]; in public health, when estimating disease prevalence [42]–[45]; and in GWAS studies based primarily on a European sample, to the near exclusion of other groups [46]–[48].

Therefore, information theory needs to consider techniques based on local events and unaveraged measures that can inform

Y. Chen and D. A. Díaz-Pachón are with the Division of Biostatistics and Bioinformatics, University of Miami, Miami, FL, 33136 USA (e-mail: yxc1378@miami.edu, DDiaz3@miami.edu)

Manuscript received April 19, 2005; revised August 26, 2015.

scientific discovery. Indeed, until very recently, information theory focused only on L_1 measures of information (entropy, mutual information, Kullback-Leibler divergence, etc.). Only recently have Fradilezi, Madiman, and collaborators begun to exploit the importance of the variance of information content, which they beautifully called *varentropy*, to study the concentration of measure phenomenon [49], [50]. The usefulness of this simple concept is epitomized by the extreme success that probabilist Justin Salez has achieved by using it to prove multiple instances of the cut-off phenomenon in Markov chains [51]–[58]. Thus, our aim is for unaveraged information measures to have the same effect in theory and in practice.

In this context, **active information** (AIN) was proposed to quantify the amount of information infused to reach a target in a search problem. The concept can be readily extended beyond search spaces to any probability space (Section II). In this paper, we introduce *conserved active information* (CAIN) based on AIN. We use CAIN to show that, when the amount of AIN in a system exceeds a certain threshold, conservation of information renders it no longer possible to regard AIN as a simple redistribution of the available information in the system (Section III). Rather, new information has been added externally, and CAIN requires that it be accounted for. In search problems, being above this threshold requires a programmer. More generally, CAIN forces us to see a system as open. Therefore, given the laxity with which the word *emergence* is used for events that we do not understand [59], we propose here to define as emergent any target/event whose AIN is beyond the given threshold.

The paper is structured as follows: Section II defines active information and examines its key properties. Section III introduces conserved active information, presents results demonstrating its properties, and illustrates its importance with examples. In Section IV, we conclude with a discussion about the relevance of conserved active information and some open problems.

II. ACTIVE INFORMATION

Another way to see the limitations of averaging is through the highly consequential No Free Lunch theorems (NFLTs), which assert that, on average, no search does better than a random one. The NFLTs imply that, "if an algorithm performs better than random search on some class of problems, then it must perform *worse than random search* on the remaining problems" [60], underscoring "the importance of incorporating problem-specific knowledge into the behavior of the algorithm" [60], [61]. According to the NFLTs, it is precisely this problem-specific knowledge that enables biasing an algorithm to reach the target more effectively than a blind search (a fact that is particularly true in machine learning [62], [63]). Therefore, AIN I^+, was introduced to measure the amount of information I_2 infused by a programmer to reach a target T in a search space $\mathcal{X}$, relative to the information I_1 provided by a baseline [64]–[67]:

$$\begin{aligned} I^+ = I^+_{\mathsf{T}}(\mathbf{P}_2 \parallel \mathbf{P}_1) = I^+(\mathbf{P}_2 \parallel \mathbf{P}_1)(\mathsf{T}) \\ := I_1(\mathsf{T}) - I_2(\mathsf{T}) = \log \frac{\mathbf{P}_2(\mathsf{T})}{\mathbf{P}_1(\mathsf{T})}, \end{aligned} \tag{1}$$

where $I_i(\mathsf{T}) = -\log \mathbf{P}_i(\mathsf{T})$ is the self-information of T under the law $\mathbf{P}_i$, for $i = 1, 2$, and we assume $\log(0/0) = 0$ by continuity. We allow I^+ to take values in the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$. As is customary in search problems, we also assume $|\mathsf{T}| \ll |\mathcal{X}| < \infty$. The probabilities $\mathbf{P}_1$ and $\mathbf{P}_2$ denote random search and search with additional knowledge, respectively. Thus, $I_1(\mathsf{T})$ is the **endogenous information**, measuring the baseline difficulty of finding T, and $I_2(\mathsf{T})$ is the **exogenous information**, measuring the difficulty of the problem given the additional knowledge, making the AIN I^+ the problem-specific information incorporated into the search.

Observe that, in the definition (1) of AIN, only slightly more than T being measurable *by* the two probabilities is required. Formally, we require two probability spaces $(\mathcal{X}_1, \mathcal{F}_1, \mathbf{P}_1)$, $(\mathcal{X}_2, \mathcal{F}_2, \mathbf{P}_2)$, and $\mathsf{T} \in \mathcal{F}_1 \cap \mathcal{F}_2$. In applications, this distinction between the two spaces is key. For instance, in the Brillouin AIN, the programmer learns that the target $\mathsf{T} \subset \mathcal{X}_2 \subset \mathcal{X}_1$ lies within a subset of the original space [68]. Then, assuming that $\mathbf{P}_1$ corresponds to a blind search, $I^+_{\mathsf{T}} = \log \mathbf{P}_2(\mathsf{T}) - \log \mathbf{P}_1(\mathsf{T}) = \log(|\mathcal{X}_1|/|\mathcal{X}_2|)$. However, for theoretical purposes, it is better to work with a single measurable space $(\mathcal{X}, \mathcal{F})$:

Lemma 1. *Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, $\mathcal{F} = \sigma(\mathcal{F}_1, \mathcal{F}_2)$, and extend the measures $\mathbf{P}_i$, for $i = 1, 2$, to*

$$\mathbf{P}^*_i(\mathsf{A}) = \begin{cases} \mathbf{P}_i(\mathsf{A}) & \text{if } \mathsf{A} \in \mathcal{F}_i, \\ 0 & \text{otherwise}. \end{cases}$$

Then the measurable space $(\mathcal{X}, \mathcal{F})$ and the probability measures $\mathbf{P}^*_1, \mathbf{P}^*_2$ associated with it exist and are well-defined.

Proof. For $i = \{1, 2\}$, we need to verify that

- $\mathbf{P}^*_i(\mathcal{X}) = 1$,
- $\mathbf{P}^*_i(\mathsf{A}) \geq 0$,
- Countable additivity: If $\{\mathsf{A}^*_j\}_{j \in \mathbb{N}} \subset \mathcal{F}$ is a disjoint sequence of measurable sets, then $\mathbf{P}^*_i\left(\bigcup_j \mathsf{A}^*_j\right) = \sum_j \mathbf{P}^*_i(\mathsf{A}^*_j)$, for $i \in \{1, 2\}$.

The first two properties are trivial. As for countable additivity, redefine A^*_j as

$$\mathsf{A}^*_j = \begin{cases} \mathsf{A}_j & \text{if } \mathsf{A}^*_j \in \mathcal{F}_i \\ 0 & \text{otherwise}. \end{cases}$$

Then

$$\mathbf{P}^*_i\left(\bigcup_j \mathsf{A}^*_j\right) = \mathbf{P}_i\left(\bigcup_j \mathsf{A}_j\right) = \sum_j \mathbf{P}_i(\mathsf{A}_j) = \sum_j \mathbf{P}^*_i(\mathsf{A}^*_j),$$

as required. □

Remark 1. *With a slight abuse of notation, we will still refer to $\mathbf{P}^*_1$ and $\mathbf{P}^*_2$ as $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively.*

This unified space enables seamless comparison and generalization beyond finite discrete searches. Indeed, the measure-theoretic formulation allows writing AIN as

$$I^+ = \int_{\mathsf{T}} \log \frac{p_2(x)}{p_1(x)} \mathrm{d}\mu, \tag{2}$$

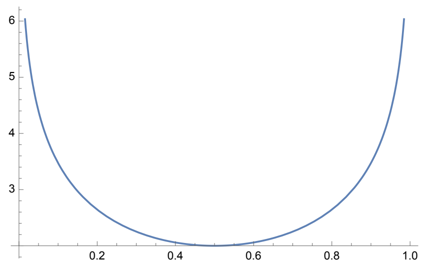

Fig. 1. Total information $\mathcal{H}(X)$ of $X \sim \text{Ber}(p)$. Logs taken in base 2.

where μ is a measure on $(\mathcal{X}, \mathcal{F})$, $\mathbf{P}_i$ is absolutely continuous w.r.t. μ (noted $\mathbf{P}_i \ll \mu$), and $p_i = \mathrm{d}\mathbf{P}_i/\mathrm{d}\mu$ is its Radon–Nikodym derivative, for $i = 1, 2$ (such a measure μ always exists, for example, $\mu = (\mathbf{P}_1 + \mathbf{P}_2)/2$). Therefore, great generality is achieved, depending on the choice of $\mathcal{X}$ (not necessarily finite), $\mathcal{F}$, $\mathbf{P}_2$, $\mathbf{P}_1$, T, and f (defined below):

Definition 1. *Let $f : \mathcal{X} \to \mathbb{R}$ be a real function on $\mathcal{X}$ and assume that we are interested in events T where f is large:*

$$\mathsf{T} = \{x \in \mathcal{X};\, f(x) \geq f_0\} \tag{3}$$

for some $f_0 \in \mathbb{R}$. Then we say that T is a **specification** *and f is a* **specification function**.

We now define the **Kullback-Leibler divergence** (KL):

$$\begin{aligned} &\mathrm{KL}(\mathbf{P}_2 \parallel \mathbf{P}_1) = \mathrm{KL}(X_2 \parallel X_1) \\ &:= \begin{cases} \int p_2(x) \log \frac{p_2(x)}{p_1(x)} \mu(\mathrm{d}x) & \text{if } \mathbf{P}_1, \mathbf{P}_2 \ll \mu, \\ \infty & \text{otherwise,} \end{cases} \end{aligned} \tag{4}$$

where μ is as in (2). Some have proposed KL as the most basic concept in information theory because it can be applied equally to discrete and continuous random variables, being always nonnegative and uniquely defined, whereas entropy, when extended to continuous random variables, can be negative and is not invariant under transformation of the distribution's parameters [69], [70].

Returning to AIN in (2), it is a comparative measure that quantifies how much information is added or removed by viewing T through $\mathbf{P}_2$ rather than $\mathbf{P}_1$, even when the space is not discrete. To illustrate some properties of AIN, let $\mathfrak{P}$ be the class of all probability measures on $(\mathcal{X}, \mathcal{F})$. Then, for all $\mathbf{P}, \mathbf{P}^*, \mathbf{P}' \in \mathfrak{P}$ and all $\mathsf{A} \in \mathcal{F}$, AIN satisfies the following properties:

P1 Can be positive, negative, or zero:
- $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}^*) < 0$ iff $\mathbf{P}(\mathsf{A}) < \mathbf{P}^*(\mathsf{A})$,
- $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}^*) = 0$ iff $\mathbf{P}(\mathsf{A}) = \mathbf{P}^*(\mathsf{A})$,
- $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}^*) > 0$ iff $\mathbf{P}(\mathsf{A}) > \mathbf{P}^*(\mathsf{A})$.

P2 Induces a total ordering of $\mathfrak{P}$, in the sense that $\mathbf{P} \geq \mathbf{P}^*$ iff $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}^*) \geq 0$. It satisfies
- Reflexivity: $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}) \geq 0$ iff $\mathbf{P}(\mathsf{A}) \geq \mathbf{P}(\mathsf{A})$.
- Antisymmetry: $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}') \geq 0$ and $I_{\mathsf{A}}^{+}(\mathbf{P}' \parallel \mathbf{P}) \geq 0$ iff $\mathbf{P}(\mathsf{A}) = \mathbf{P}'(\mathsf{A})$.
- Transitivity: If $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}^*) \geq 0$ and $I_{\mathsf{A}}^{+}(\mathbf{P}' \parallel \mathbf{P}) \geq 0$, then $I_{\mathsf{A}}^{+}(\mathbf{P}' \parallel \mathbf{P}^*) \geq 0$ and $\mathbf{P}'(\mathsf{A}) \geq \mathbf{P}^*(\mathsf{A})$.
- Totality: $I_{\mathsf{A}}^{+}(\mathbf{P} \parallel \mathbf{P}') \geq 0$ or $I_{\mathsf{A}}^{+}(\mathbf{P}' \parallel \mathbf{P}) \geq 0$.

Properties **P1**–**P2** establish other important differences between AIN and KL. First, KL cannot be negative; therefore, it is not a comparative measure and cannot quantify local information changes between $\mathbf{P}_2$ and bfP_1. This is a key consideration in Section III. Second, KL does not induce any ordering whatsoever because, as a relation, it is neither antisymmetric nor transitive. The only comparison we can learn about $\mathbf{P}_2$ and $\mathbf{P}_1$ when looking at $\mathrm{KL}(\mathbf{P}_2 \parallel \mathbf{P}_1)$ is that $\mathrm{KL}(\mathbf{P}_2 \parallel \mathbf{P}_1) = 0$ iff $\mathbf{P}_2 = \mathbf{P}_1$ a.s. Third, KL can be used to obtain AIN, but obtaining AIN from KL is generally not possible. A third important property of AIN, shared by the KL divergence, is:

P3 AIN tensorizes. That is, for an n-dimensional Cartesian product $\mathsf{A} = \mathsf{A}_1 \times \cdots \times \mathsf{A}_n$, and product measures $\mathbf{P}_2 := P_{21} \otimes \cdots \otimes P_{2n}$, $\mathbf{P}_1 := P_{11} \otimes \cdots \otimes P_{1n}$, where $P_{21}, \ldots, P_{2n}$ and $P_{11}, \ldots, P_{1n}$ are independent, and $P_{11}(\mathsf{A}_1) \cdots P_{1m}(\mathsf{A}_n) \neq 0$, we have that $\log[\mathbf{P}_2(\mathsf{A})/\mathbf{P}_1(\mathsf{A})] = \sum_{i=1}^{n} \log[P_{2i}(\mathsf{A}_i)/P_{1i}(\mathsf{A}_i)]$.

In high-dimensional scenarios, property **P3** enables decomposition in product spaces. Moreover, **P3** makes AIN desirable over

$$\mathbf{P}_2(\mathsf{A}) - \mathbf{P}_1(\mathsf{A}) \tag{5}$$

since (5) does not tensorize. Among others, this allows a more practical redefinition of bias of prevalence as follows: If $\mathbf{P}_1(\mathsf{A})$ is the true prevalence of A and $\mathbf{P}_2(\mathsf{A})$ is the expected value of its estimator, I^+ is preferable to the usual definition of bias as (5); see, for instance, [39], [40], [43], [45], [71]. Also, because of **P3**, KL is usually preferable to the total variation distance $\mathrm{TV} = \mathrm{TV}(\mathbf{P}_2, \mathbf{P}_1) := \sup_{\mathsf{A} \in \mathcal{F}} |\mathbf{P}_2(\mathsf{A}) - \mathbf{P}_1(\mathsf{A})|$ because, by Pinsker's inequality, TV can be approximated by KL:

$$\mathrm{TV}(\mathbf{P}_2, \mathbf{P}_1) \leq \sqrt{\mathrm{KL}(\mathbf{P}_2 \parallel \mathbf{P}_1)/2}. \tag{6}$$

For analytical developments and diverse applications of specifications and AIN, see [42]–[45], [71]–[87].

III. Conserved Active Information

When Wolpert and Macready introduced the NFLTs, they realized that improved performance on the target T implied impoverished performance on T^c, and that this should be reported. However, as explained in Section II, AIN reports only the amount of information added or removed by the programmer to reach the target, without accounting for information added or removed from the rest of the space. Thus, we propose here **conserved active information**. To introduce it, we start with a simple question: If AIN can be negative, why is KL nonnegative? To answer this question, we first define the **total information** of a random variable X with law $\mathbf{P}_1$ as

$$\mathcal{H}(X) := \int_{\mathcal{X}} -\log \mathrm{d}\mathbf{P}_1(x). \tag{7}$$

When X discrete, the RHS of (7) becomes $-\sum_{x \in \mathcal{X}} \log p_1(x)$; when X is absolutely continuous, it becomes $-\int_{\mathcal{X}} \log p_1(x) \mathrm{d}x$. As indicated by its name, $\mathcal{H}$ quantifies the total amount of information in a given system.

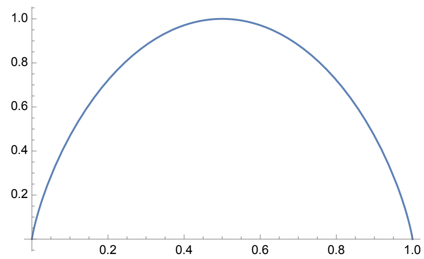

Fig. 2. Entropy $H(X)$ of $X \sim \text{Ber}(p)$. Logs taken in base 2

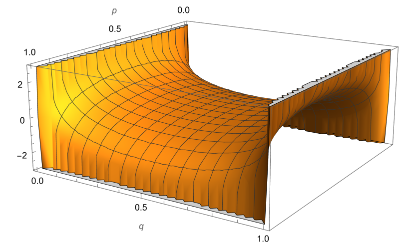

Fig. 3. $I^{\oplus}(\mathbf{P}_1, \mathbf{P}_2)$ when $\mathbf{P}_1 \sim \text{Ber}(p)$ and $\mathbf{P}_2 \sim \text{Ber}(q)$.

Example 1. *Consider a Bernoulli r.v. X in $\mathcal{X} = \{0, 1\}$ with probability of success $p = \mathbf{P}_1(1)$ and entropy $H(X)$. Then*

$$\begin{aligned} \mathcal{H}(X) &= -\log p - \log(1-p), \\ H(X) &= -p\log p - (1-p)\log(1-p) \end{aligned} \tag{8}$$

It is easy to verify that when $p = 0.5$, $\mathcal{H}(X)$ is minimized (see Fig. 1), whereas the entropy $H(X)$ is maximized (see Fig. 2). In base 2, $\mathcal{H}(1/2) = 2$ and $H(1/2) = 1$. On the other hand, as p approaches 0 or 1, $\mathcal{H}(X)$ increases to infinity, while $H(X)$ decreases to 0. In more detail, for $p \ll 1$,

$$\begin{aligned} \mathcal{H}(X) &\approx -\log p \gg 0, \\ H(X) &\approx -p\log p \ll 1, \end{aligned} \tag{9}$$

revealing that the contribution of the second terms in (8) becomes negligible. Still, it is also clear from (9) that the small weight of the first term of $H(X)$ tames the strength of the first term in $\mathcal{H}(X)$.

Example 1 shows that total information and entropy move in opposite directions. When uncertainty is maximal, entropy is maximized, whereas total information is minimized; and when certainty is maximal, entropy is minimized, whereas total information is maximized.

The previous setting can be naturally extended to **conserved active information**, defined as

$$\begin{aligned} I^{\oplus} &= I^{\oplus}(\mathbf{P}_1, \mathbf{P}_2) = I^{\oplus}(X_1, X_2) \\ &:= \mathcal{H}(X_2) - \mathcal{H}(X_1) = \int_{\mathcal{X}} \log \frac{p_1(x)}{p_2(x)} \mathrm{d}\mu, \end{aligned} \tag{10}$$

where μ, p_2, p_1 are as defined in (2). When X_1, X_2 are discrete, the RHS of (10) becomes $-\sum_{x\in\mathcal{X}} \log[p_1(x)/p_2(x)]$; when X_1, X_2 are continuous, the RHS of (10) becomes $-\int_{\mathcal{X}} \log[p_1(x)/p_2(x)]\mathrm{d}x$. Then $I^{\oplus}$ is the amount of additional information needed in the system to replace $X_1 \sim \mathbf{P}_1$ with $X_2 \sim \mathbf{P}_2$. Therefore, when $\mathcal{H}(X_2) \le \mathcal{H}(X_1)$, $I^{\oplus} \le 0$, implying that information is conserved when moving from $\mathbf{P}_1$ to $\mathbf{P}_2$. On the other hand, when $\mathcal{H}(X_2) > \mathcal{H}(X_1)$, $I^{\oplus} > 0$, implying that new information must be infused externally to sustain the new system induced by X_2. The definition of $I^{\oplus}$ and Pinsker inequality (6) directly imply the following Lemma:

Lemma 2. *Let $\mathcal{X} = \{1, \dots, N\}$ be a finite space and define $X_1 \sim \mathcal{U}(N)$ as a uniform random variable in $\mathcal{X}$. Then, for any $X_2 \sim \mathbf{P}_2$ fully supported in $\mathcal{X}$, we have*

1) $N \cdot \text{KL}(\mathbf{P}_1 \parallel \mathbf{P}_2) = I^{\oplus}(\mathbf{P}_1, \mathbf{P}_2)$.
2) $\text{TV}(\mathbf{P}_2, \mathbf{P}_1) \le \sqrt{I^{\oplus}(\mathbf{P}_1, \mathbf{P}_2)/(2N)}$.

As before, a brief example will shed much light on the interpretation of $I^{\oplus}$.

Example 2. *Consider the space $\mathcal{X} = \{0, 1\}$, and two Bernoulli random variables X_2, X_1 with success probabilities $\mathbf{P}_1(\{X_1 = 1\}) = p$ and $\mathbf{P}_2(\{X_2 = 1\}) = q$, respectively, Figure 3 shows that*

$$I^{\oplus} = \log\frac{p}{q} + \log\frac{1-p}{1-q} \tag{11}$$

is a saddle. Three simple scenarios can be discerned:

C1 *If $p \in \{q, 1-q\}$, then $I^{\oplus} = 0$.*
C2 *If $p \neq 0.5 = q$, then $I^{\oplus} < 0$.*
C3 *If $p = 0.5 \neq q$, then $I^{\oplus} > 0$.*

*In scenario **C1**, the system's total information remains constant. In scenario **C2**, a somewhat ordered system decays to maxent, implying that the system redistributes its available information without external input. In scenario **C3**, the system is initially in maximum entropy (maxent), and any update to $\mathbf{P}_2$ orders the entire system, requiring external input. In other words, the conserved active information increases or decreases as the system becomes more or less ordered.*

How then is KL nonnegative if I^{+} and $I^{\oplus}$ can be negative? (Figure 4). Again, let's consider the simplest case of two Bernoulli random variables: $X_1 \sim \text{Ber}(0.5 + \epsilon)$ and $X_2 \sim \text{Ber}(0.5)$, with $\epsilon \in (0, 0.5]$. Then

$$\begin{aligned} I^{\oplus}(\mathbf{P}_1, \mathbf{P}_2) &= \log\frac{0.5-\epsilon}{0.5} + \log\frac{0.5+\epsilon}{0.5} \\ \text{KL}(\mathbf{P}_1 \parallel \mathbf{P}_2) &= (0.5-\epsilon)\log\frac{0.5-\epsilon}{0.5} + (0.5+\epsilon)\log\frac{0.5+\epsilon}{0.5}. \end{aligned} \tag{12}$$

Example 2 and Fig. 3 show that $I^{\oplus} < 0$ in (12). Moreover, the first term of $I^{\oplus}$ is negative, whereas the second is positive. Therefore, the first term determines the sign of $I^{\oplus}$. However, KL is positive not because all its terms are positive (the weighted first term is not), but because the positive, weighted second term outweighs the weighted first term. In fact, as $\epsilon \uparrow 0.5$, $I^{\oplus}$ decreases to $-\infty$ because its first term decreases to $-\infty$, whereas KL approaches $\log 2$ because its weighted first term approaches 0.

As Example 2 makes clear, $I^{\oplus}$ reveals information hidden from KL, with strong implications for the conservation of information. This is made explicit in the theorem below.

Theorem 3 (Regimes of Conserved Active Information under Uniform Baseline)**.** *Let $\mathcal{X} = \{1, \dots, N\}$, where $\mathbf{P}_1 \sim \mathcal{U}(N)$ is the maxent distribution, and the target T satisfies $|\mathsf{T}| \ll$*

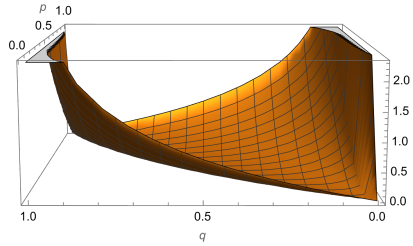

Fig. 4. $\mathrm{KL}(\mathbf{P}_1 \parallel \mathbf{P}_2)$ when $\mathbf{P}_1 \sim \mathrm{Ber}(p)$ and $\mathbf{P}_2 \sim \mathrm{Ber}(q)$.

$|\mathcal{X}|/2$, so that $p = \mathbf{P}_1(\mathsf{T}) = |\mathsf{T}|/|\mathcal{X}| \ll 1/2$. Assume the programmer has additional information yielding $q = \mathbf{P}_2(\mathsf{T})$. If $q < p$ or $q > 1-p$, then $I^{\oplus} > 0$ (the system is more organized); if $p < q < 1-p$, then $I^{\oplus} < 0$ (the system is more disorganized); and if $q \in \{p, 1-p\}$, then $I^{\oplus} = 0$ (the system is equally ordered).

Proof.

$$I^{\oplus}(\mathbf{P}_1, \mathbf{P}_2) = \log \frac{p(1-p)}{q(1-q)},$$

which proves that $I^{\oplus} = 0$ iff $q \in \{p, 1-p\}$. After some algebraic manipulation, we obtain that,

$$\log \frac{p(1-p)}{q(1-q)} < 0 \Leftrightarrow (q-p)(p+q-1) < 0. \tag{13}$$

Thus, the two terms on the RHS of (13) must have opposite signs. If $q-p < 0$, then $p+q-1 > 0$ to satisfy the inequality. However, this is equivalent to $q > 1-p$. But then $q < p < 1/2 < 1-p < q$, a contradiction. Therefore, the RHS of (13) is negative iff $q > p$ and $q < 1-p$. □

Remark 2. *Given the general definition of AIN in (2), note that Theorem 3 holds unchanged when $\mathcal{X} = [0, N]$ and $|\cdot|$ is the Lebesgue measure.*

Theorem 3 has the following important corollary:

Corollary 4. *Under the conditions of Theorem 3, if $I^+ < 0$, then $I^{\oplus} > 0$. On the other hand, if $I^+ > 0$, then $I^{\oplus} > 0$ only when $q > 1-p$. Otherwise, if $I^+ > 0$ and $p < q < 1-p$, then $I^{\oplus} < 0$.*

The findings of Corollary 4 are summarized in Table I.

The implications of Theorem 3 and Corollary 4 are as follows. Consider a search problem in $\mathcal{X} = \{1, \ldots, N\}$, where the system remains in maxent (i.e., $X_1 \sim \mathcal{U}(N)$ is uniformly distributed over $\mathcal{X}$) and the target T satisfies $|\mathsf{T}| \ll |\mathcal{X}|/2$. In the notation of Example 2, $p = |\mathsf{T}|/N$. Assume that an alternative search yields $\mathbf{P}_2(\mathsf{T}) = q$. If $q < p$, then $I^+_{\mathsf{T}} < 0$ but $I^{\oplus} > 0$, meaning the system is more ordered while the target is less probable—an uninteresting scenario of low complexity.

On the other hand, $I^{\oplus} = 0$ when $q \in \{p, 1-p\}$, whereas KL = 0 only when $p = q$. If $q \in (p, 1-p)$, then $I^+_{\mathsf{T}} > 0$ but $I^{\oplus} < 0$, meaning the target search is more probable at the expense of making the whole system more disordered; in other words, this is a system of moderate complexity, where the system's information can be redistributed to make the target more likely. However, if $q > 1-p$, then $I^+_{\mathsf{T}} > 0$ and $I^{\oplus} > 0$, meaning the target becomes much more likely while the order of the whole system increases; this is the jackpot scenario of very high complexity that makes T emergent, where conservation of information requires accounting for the excess total information $I^{\oplus} > 0$ required to produce T. Again, none of the scenarios where $I^{\oplus} \neq 0$ can be discerned by looking only at KL.

Definition 2 (Emergence)**.** *We say that an event $\mathsf{T} \in \mathcal{X}$ is emergent if and only if $I^+(\mathsf{T}) > 0$ and $I^{\oplus} > 0$.*

Example 3 (Markov chains [88])**.** *Let $\mathcal{G} = (\mathcal{X}, E)$ be a connected graph with set of vertices $\mathcal{X}$ such that $|\mathcal{X}| < \infty$. Assume that the graph is d-regular, meaning that the degree of all vertices is d. Then the random walk associated with it has a transition matrix given by*

$$P(x,y) = \begin{cases} 1/d & \text{if } (x,y) \in E \\ 0 & \text{otherwise.} \end{cases}$$

The connectedness and regularity of the graph imply that the random walk is irreducible and has a unique stationary distribution $\pi \sim \mathcal{U}(N)$, which is uniform over N. Let $\mathsf{T} \subset \mathcal{X}$ be a target such that $|\mathsf{T}| < 1/2$, and consider an initial distribution $\mathbf{P}_1$ such that $p = \mathbf{P}_1(\mathsf{T}) \ll |\mathsf{T}|/|\mathcal{X}|$. Then, letting $\mathbf{P}_2$ denote the stationary distribution of the random walk, we have $q = \mathbf{P}_2(\mathsf{T}) = |\mathsf{T}|/|\mathcal{X}|$, greatly increasing the probability of the target through the chain's neutral dynamics. This illustrates the mild-knowledge regime $p < q < 1-p$ in Theorem 3, showing that it is possible to increase the probability of the target via a natural information redistribution within the system.

Example 4 (Cosmological fine-tuning [35])**.** *Cosmological fine-tuning (FT) asserts that some constants of nature must fall within an interval $\ell \subset \mathcal{X} \subset \mathbb{R}$ for carbon-based life to exist [89]. Let $\mathsf{T} \subset \mathcal{X}$ be the event "We observe a universe that exists and permits life." Let x denote the value of a particular constant of nature, and assume that x is a parameter of a model $\mathbf{Q}(\mathsf{T} \mid x)$ that gives the probability of observing a life-permitting universe. According to (3), this corresponds to T when the specification function $f(x) = \mathbb{1}_{\mathsf{T}}(x)$ is binary and satisfies $f(x) = 1$ if the value x of the constant permits a universe with life ($f(x) = 0$ otherwise).*

The tuning probability of the event T is obtained by treating x as an observation of a continuous random variable $X \sim \mathbf{P}$, where $\mathbf{P}$ is in maxent, and averaging $\mathbf{Q}(\mathsf{T} \mid x)$ with respect to x, i.e., $\mathbf{Q}(\mathsf{T}; \xi) = \int_{\mathcal{X}} \mathbf{Q}(\mathsf{T} \mid x) \mathrm{d}\mathbf{P}(x; \xi)$, where $\mathbf{Q}(\mathsf{T} \mid x)$ is the likelihood. If $\mathsf{T} = \ell$, then $\mathbf{Q}(\mathsf{T} \mid x) = \mathbb{1}_{\mathsf{T}}(x)$, and we obtain $\mathbf{Q}(\mathsf{T}; \xi) = \mathbf{P}(\mathsf{T}; \xi)$. To determine the degree of tuning, we maximize $\mathbf{P}(\mathsf{T}; \xi)$ with respect to the hyperparameter ξ as ξ varies over a finite-dimensional space Ξ. That is, the final degree of tuning equals $\mathbf{P}_1(\mathsf{T}) := \sup_{\xi \in \Xi} \mathbf{P}(\mathsf{T}; \xi)$. Since T is observed, this induces a second distribution $\mathbf{P}_2$ s.t. $\mathbf{P}_2(\mathsf{T}) = 1$. Then $\mathbf{P}_0(\mathsf{T}) < \delta \ll 1$ is equivalent to $I^+(\mathsf{T}) > -\log \delta \gg 0$, in which case we infer that X is fine-tuned to level δ for the family of distributions $\mathfrak{F} := \{\mathbf{P}(\mathsf{T}; \xi)\}_{\xi \in \Xi}$.

In this scenario, it has been shown that the gravitational constant and the critical density of the universe are fine-tuned. In contrast, other constants, such as the Higgs vacuum

TABLE I
REGIMES OF CONSERVED ACTIVE INFORMATION UNDER UNIFORM BASELINE

Regime	Condition	$I^+(T)$	$I^\oplus$	Interpretation
Harmful to T	$q < p$	< 0	> 0	Target harder, system more ordered
Mild knowledge	$p \leq q \leq 1 - p$	≥ 0	≤ 0	Target easier, system more disordered
Strong knowledge	$q > 1 - p$	> 0	> 0	Target much easier AND system more ordered (jackpot)

expectation value and the amplitude of the primordial fluctuations, are not [72]–[75]. Therefore, the presence of fine-tuning, $\mathbf{P}_1(\mathsf{T}) \ll 1$ *and* $\mathbf{P}_2(\mathsf{T}) = 1$*, implies that the universe itself is an emergent event of high complexity.*

IV. Discussion

In this paper, we have introduced total information and conserved active information. We have highlighted their properties and contrasted them with the key properties of AIN and KL, which are rarely discussed in the literature. Since neither AIN nor KL can describe changes in the system's total information, $I^\oplus$ provides a necessary criterion for determining whether a specification T requires the infusion of information. When it does, we propose to define the specification as emergent. Therefore, $I^\oplus$ provides the missing scalar for NFL-compliant active information reporting, with extensions unifying information theory and complex systems.

This work can be extended in several ways. First, a natural extension of conserved active information is its thermodynamic implications, where analogies between information theory and thermodynamics are particularly fruitful [90], [91]. Informally, by replacing the term *information* with *energy* in our framework, we can draw direct parallels that illuminate far-from-equilibrium processes central to complex systems. In thermodynamics, energy is conserved by the first law, but its usable form (free energy) decreases via the second law, increasing entropy in isolated systems. Similarly, the NFLTs enforce a conservation principle for information: gains in performance on a target T must be offset by losses elsewhere, averaged over all problems. Conserved active information $I^\oplus = I^+(\mathsf{T}) + I^+(\mathsf{T}^c)$ captures this balance, akin to total energy in a closed system. Positive conserved active information signals an infusion of external "work" (problem-specific knowledge), reducing global disorder much like an engine extracting work to decrease local entropy while increasing it overall [92], [93]. Negative conserved active information, conversely, reflects internal redistribution without net gain, mirroring dissipative processes where entropy rises spontaneously, as in Example 3.

Second, a compelling extension of conserved active information concerns its potential role in evaluating artificial intelligence (AI) systems, particularly large language models (LLMs), and in the pursuit of artificial general intelligence (AGI). In search and optimization, LLMs leverage vast pretrained knowledge to enhance performance on specific tasks, effectively increasing active information by biasing toward relevant targets [71], [76], [77]. We conjecture that this often occurs within the "mild knowledge" regime identified in Theorem 3, where $p < q < 1 - p$ and $I^\oplus < 0$, indicating that the system becomes more disordered overall. Here, the apparent gains in $I^\oplus(\mathsf{T})$ stem from internal redistribution of probabilities—drawing on the programmer's embedded knowledge (e.g., training data curation and architecture design)—rather than a net infusion of order. In contrast, true AGI could be characterized by the machine's autonomous ability—not the programmer's—to transition into the "strong knowledge" regime ($q > 1 - p$ and $I^\oplus > 0$), where it imposes genuine order on the entire space by injecting problem-specific "insights" without external scaffolding. This jump to positive $I^\oplus$ would signify self-generated conservation violations, akin to emergent creativity or adaptation in unexplored landscapes.

Third, in the context of statistical estimation, consider a population $\mathcal{X}$ and a subset $\mathsf{T} \subset \mathcal{X}$ of, say, infected individuals with some disease. The true prevalence is given by $p = \mathbf{P}_1(\mathsf{T}) = |\mathsf{T}|/|\mathcal{X}|$. An estimator introduces bias, yielding $q = \mathbf{P}_2(\mathsf{T})$ [43], [71]. Here, active information $I^+(\mathsf{T}) = \log(q/p)$ quantifies the magnitude and direction of bias: positive for overestimation, negative for underestimation, and 0 for unbiasedness. We believe that the regimes delineated in Theorem 3 provide an objective framework for assessing permissible bias. Specifically, the mild knowledge regime ($p < q < 1/2 < 1 - p$) bounds the extent to which, under unbiased sampling, an estimator can inflate q without violating conservation principles by injecting excessive external knowledge. For rare events where $p \ll 1/2$, the interval $(p, 1/2)$ is vast, allowing substantial overestimation ($q \gg p$) while keeping $I^\oplus < 0$. Consistent with the conservation of information, estimators may appear effective locally but degrade globally, leading to overconfidence in rare-event detection. In contrast, crossing into the "strong knowledge" regime signals that the estimator has imposed genuine order, potentially through unaccounted-for external inputs (e.g., biased sampling or overfitting), thereby violating the conservation of information, unless justified by problem-specific structure. This framework thus offers a diagnostic tool for bias in prevalence studies: compute $I^\oplus$ to determine whether the observed bias is mild (permissible redistribution) or strong (indicating the need for scrutiny of hidden knowledge).

References

[1] A. Eddington, *The Nature of the Physical World*. Cambridge: Cambridge University Press, 1928.
[2] R. Penrose, *The Road to Reality*. Random House, 2004.
[3] P. W. Anderson, "More is Different: Broken symmetry and the nature of the hierachical structure of science," *Science*, vol. 177, no. 4047, pp. 393–396, 1972.
[4] R. B. Laughlin and D. Pines, "The theory of everything," *Proceedings of the National Academy of Sciences*, vol. 97, no. 1, pp. 28–31, 2000.
[5] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.
[6] M. Dudík, "Maximum Entropy Density Estimation with Generalized Regularization and an Application to Species Distribution Modeling," *Journal of Machine Learning Research*, vol. 8, pp. 1217–1260, 2007.

[7] J. E. Chacón, "The Modal Age of Statistics," *International Statistical Review*, vol. 88, pp. 122–141, 2020.
[8] J.-E. Dazard and J. S. Rao, "Local Sparse Bump Hunting," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 900–929, 2010.
[9] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data," *Statistics and Computing*, vol. 9, pp. 123–143, 1999.
[10] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International Journal of Epidemiology*, vol. 41, pp. 200–209, 2012.
[11] W. Polonik and Z. Wang, "PRIM Analysis," *Journal of Multivariate Analysis*, vol. 101, no. 3, pp. 525–540, 2010.
[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer Science, 2009.
[13] B. W. Silverman, "Using kernel density estimates to investigate multimodality," *Journal of the Royal Statistical Society. Series B*, vol. 43, pp. 97–99, 1981.
[14] ——, *Density Estimation for Statistics and Data Analysis*. Boca Raton: Chapman and Hall/CRC, 1986.
[15] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.
[16] ——, *The Nature of Statistical Learning Theory*, 2nd ed. Cham: Springer, 2013.
[17] P. Davies, *The Demon in the Machine*. Great Britain: Allen Lane, 2019.
[18] E. Shrödinger, *What is life?* Cambridge: Cambridge University Press, 1944.
[19] G. Bisker and J. England, "Nonequilibrium associative retrieval of multiple stored self-assembly targets," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. E10 531–E10 538, 2018.
[20] J. England, *Every life is on fire: How thermodynamics explains the origins of living things*. Basic Books, 2020.
[21] ——, "Self-organized computation in the far-from-equilibrium cell," *Biophysics Reviews*, vol. 3, no. 4, p. 041303, 2022.
[22] A. E. Snyder-Beattie, A. Sandberg, K. E. Drexler, and M. B. Bonsall, "The timing of evolutionary transitions suggests intelligent life is rare," *Astrobiology*, vol. 21, no. 3, pp. 265–278, 2021.
[23] R. M. Hazen, P. L. Griffin, J. M. Carothers, and J. W. Szostak, "Functional information and the emergence of biocomplexity," *Proceedings of the National Academy of Sciences*, vol. 104, no. Suppl 1, pp. 8574–8581, 2007.
[24] J. W. Szostak, "Functional information: Molecular messages," *Nature*, vol. 423, no. 6941, p. 689, 2003.
[25] E. Lagasse and M. Levin, "Future medicine: from molecular pathways to the collective intelligence of the body," *Trends in Molecular Medicine*, vol. 29, no. 9, pp. 687–710, 2023.
[26] M. Levin, "Bioelectrical approaches to cancer as a problem of the scaling of the cellular self," *Progress in Biophysics and Molecular Biology*, vol. 165, pp. 102–113, 2021.
[27] C. H. Lineweaver and P. C. W. Davies, "Comparison of the atavistic model of cancer to somatic mutation theory: Phylostratigraphic analyses support the atavistic model," in *The Physics of Cancer*, B. S. Gerstman, Ed. World Scientific, 2021, ch. 12, pp. 243–261.
[28] C. H. Lineweaver, K. J. Bussey, A. C. Blackburn, and P. C. W. Davies, "Cancer progression as a sequence of atavistic reversions," *BioEssays*, vol. 43, no. 7, p. 2000305, 2021.
[29] D. G. Moore, S. I. Walker, and M. Levin, "Cancer as a disorder of patterning information: Computational and biophysical perspectives on the cancer problem," *Convergent Science Physical Oncology*, vol. 3, no. 4, p. 043001, 2017.
[30] S. W. Hawking and R. Penrose, "The singularities of gravitational collapse and cosmology," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 314, no. 1519, pp. 529–548, 1970.
[31] L. A. Barnes and G. F. Lewis, *The Cosmic Revolutionary's Handbook: (Or: How to Beat the Big Bang)*. Cambridge: Cambridge University Press, 2020.
[32] J.-P. Luminet, *The Big Bang Revolutionaries: The Untold Story of Three Scientists Who Reenchanted Cosmology*. Seattle: Discovery Institute Press, 2024.
[33] B. Carter, "The anthropic principle and its implications for biological evolution," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 310, pp. 347–363, 1983.
[34] P. Davies, *The Goldilocks Enigma: Why Is the Universe Just Right for Life?* New York: Mariner Books, 2008.
[35] G. F. Lewis and L. A. Barnes, *A Fortunate Universe: Life In a Finely Tuned Cosmos*. Cambridge: Cambridge University Press, 2016.
[36] M. J. Rees, *Just Six Numbers: The Deep Forces That Shape The Universe*. New York: Basic Books, 2000.
[37] L. E. Celis, V. Keswani, and N. K. Vishnoi, "Data preprocessing to mitigate bias: A maximum entropy based approach," *International Conference on Machine Learning*, pp. 1349–1359, 2020.
[38] G. D. Montañez, "The famine of forte: Few search problems greatly favor your algorithm," *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 477–482, 2017.
[39] G. D. Montañez, J. Hayase, J. Lauw, D. Macias, A. Trikha, and J. Vendemiatti, "The Futility of Bias-Free Learning and Search," in *2nd Australasian Joint Conference on Artificial Intelligence (AI 2019)*, J. Liu and J. Bailey, Eds. Cham: Springer, 2019, pp. 277–288.
[40] G. D. Montañez, D. Bashir, and J. Lauw, "Trading Bias for Expressivity in Artificial Learning," in *Agents and Artificial Intelligence*, A. P. Rocha, L. Steels, and J. ven den Herik, Eds. Cham: Springer, 2021, pp. 332–353.
[41] J. S. Rao, *Statistical Methods in Health Disparity Research*. Boca Raton: Chapman and Hall/CRC, 2023.
[42] D. A. Díaz-Pachón and J. S. Rao, "A simple correction for COVID-19 sampling bias," *Journal of Theoretical Biology*, vol. 512, p. 110556, March 2021.
[43] O. Hössjer, D. A. Díaz-Pachón, Z. Chen, and J. S. Rao, "An Information Theoretic Approach to Prevalence Estimation and Missing Data," *IEEE Transactions on Information Theory*, vol. 70, no. 5, pp. 3567–3582, 2024.
[44] J. S. Rao, T. Liu, and D. A. Díaz-Pachón, ""Back-to-the-future" projections for COVID-19 surges," *PLoS ONE*, vol. 19, no. 1, p. e0296964, 2024.
[45] L. Zhou, D. A. Díaz-Pachón, C. Zhao, J. S. Rao, and O. Hössjer, "Correcting prevalence estimation for biased sampling with testing errors," *Statistics in Medicine*, vol. 42, no. 26, pp. 4713–4737, 2023.
[46] C. D. Bustamante, F. M. De La Vega, and E. G. Burchard, "Genomics for the world," *Nature*, vol. 475, pp. 163–165, 2011.
[47] A. Dominguez Mantes, D. Mas Montserrat, C. D. Bustamante, and A. Ioannidis, "Neural ADMIXTURE for rapid genomic clustering," *Nature Computational Science*, vol. 3, pp. 621–629, 203.
[48] W. Zhou, M. Kanai, K.-H. H. Wu, H. Rasheed, K. Tsuo, J. B. Hirbo, Y. Wang, A. Bhattacharya, H. Zhao, S. Namba, I. Surakka, B. N. Wolford, V. Lo Faro, E. A. Lopera-Maya, K. Läll, M.-J. Favé, J. J. Partanen, S. B. Chapman, J. Karjalainen, M. Kurki, M. Maasha, B. M. Brumpton, S. Chavan, T.-T. Chen, M. Daya, Y. Ding, Y.-C. A. Feng, L. A. Guare, C. R. Gignoux, S. E. Graham, W. E. Hornsby, N. Ingold, S. I. Ismail, R. Johnson, T. Laisk, K. Lin, J. Lv, I. Y. Millwood, S. Moreno-Grau, K. Nam, P. Palta, A. Pandit, M. H. Preuss, C. Saad, S. Setia-Verma, U. Thorsteinsdottir, J. Uzunovic, A. Verma, M. Zawistowski, X. Zhong, N. Afifi, K. M. Al-Dabhani, A. Al Thani, Y. Bradford, A. Campbell, K. Crooks, G. H. de Bock, S. M. Damrauer, N. J. Douville, S. Finer, L. G. Fritsche, E. Fthenou, G. Gonzalez-Arroyo, C. J. Griffiths, Y. Guo, K. A. Hunt, A. Ioannidis, N. M. Jansonius, T. Konuma, M. T. M. Lee, A. Lopez-Pineda, Y. Matsuda, R. E. Marioni, B. Moatamed, M. A. Nava-Aguilar, K. Numakura, S. Patil, N. Rafaels, A. Richmond, A. Rojas-Muñoz, J. A. Shortt, P. Straub, R. Tao, B. Vanderwerff, M. Vernekar, Y. Veturi, K. C. Barnes, M. Boezen, Z. Chen, C.-Y. Chen, J. Cho, G. D. Smith, H. K. Finucane, L. Franke, E. R. Gamazon, A. Ganna, T. R. Gaunt, T. Ge, H. Huang, J. Huffman, N. Katsanis, J. T. Koskela, C. Lajonchere, M. H. Law, L. Li, C. M. Lindgren, R. J. F. Loos, S. MacGregor, K. Matsuda, C. M. Olsen, D. J. Porteous, J. A. Shavit, H. Snieder, T. Takano, R. C. Trembath, J. M. Vonk, D. C. Whiteman, S. J. Wicks, C. Wijmenga, J. Wright, J. Zheng, X. Zhou, P. Awadalla, M. Boehnke, C. D. Bustamante, N. J. Cox, S. Fatumo, D. H. Geschwind, C. Hayward, K. Hveem, E. E. Kenny, S. Lee, Y.-F. Lin, H. Mbarek, R. Mägi, H. C. Martin, S. E. Medland, Y. Okada, A. V. Palotie, B. Pasaniuc, D. J. Rader, M. D. Ritchie, S. Sanna, J. W. Smoller, K. Stefansson, D. A. van Heel, R. G. Walters, S. Zöllner, A. R. Martin, C. J. Willer, M. J. Daly, and B. M. Neale, "Global biobank meta-analysis initiative: Powering genetic discovery across human disease," *Cell Genomics*, vol. 2, no. 10, p. 100192, 2022.
[49] M. Fradelizi, M. Madiman, and L. Wang, "Optimal Concentration of Information Content for Log-Concave Densities," in *High Dimensional Probability VII*, C. Houdré, D. M. Mason, P. Reynaud-Bouret, and J. Rosiński, Eds. Cham: Springer, 2016, pp. 45–60.
[50] M. Fradelizi, J. Li, and M. Madiman, "Concentration of information content for convex measures," *Electronic Journal of Probability*, vol. 25, no. 20, pp. 1–22, 2020.

[51] P. Caputo, F. Münch, and J. Salez, "Entropy and curvature: beyond the peres-tetali conjecture," *Transactions of the American Mathematical Society*, 2024.

[52] P. Caputo and J. Salez, "Entropy factorization via curvature," *arXiv*, 2024.

[53] J. Hermon, X. Huang, F. Pedrotti, and J. Salez, "Concentration of information on discrete groups," *arXiv*, 2024.

[54] F. Pedrotti and J. Salez, "A new cutoff criterion for non-negatively curved chains," *arXiv*, 2025.

[55] J. Salez, "Cutoff for non-negatively curved Markov chains," *Journal of the European Mathematical Society*, 2023.

[56] ——, "Spectral gap curvature of monotone Markov chains," *Annals of Probability*, vol. 52, no. 3, pp. 1153–1161, 2024.

[57] ——, "The varentropy criterion is sharp on expanders," *Annales Henri Lebesgue*, vol. 7, pp. 239–250, 2024.

[58] ——, "Cutoff for non-negatively curved diffusions," *arXiv*, 2025.

[59] M. Levin, "Ingressing Minds: Causal Patterns Beyond Genetics and Environment in Natural, Synthetic, and Hybrid Embodiments," *PsyArXiv*, 2025.

[60] D. H. Wolpert and W. G. MacReady, "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[61] ——, "No Free Lunch Theorems for Search," Santa Fe Institute, Tech. Rep. SFI-TR-95-02-010, 1995.

[62] G. D. Montañez, "Why machine learning works," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, May 2017.

[63] D. H. Wolpert, "What is important about the No Free Lunch theorems?" in *Black Box Optimization, Machine Learning and No-Free Lunch Theorems*, P. M. Pardalos, V. Rasskazova, and M. N. Vrahatis, Eds. Springer, 2021.

[64] W. A. Dembski and R. J. Marks II, "Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search," *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, TX*, pp. 2647–2652, October 2009.

[65] ——, "Conservation of Information in Search: Measuring the Cost of Success," *IEEE Transactions Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 5, no. 5, pp. 1051–1061, September 2009.

[66] ——, "The Search for a Search: Measuring the Information Cost of Higher Level Search," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 14, no. 5, pp. 475–486, 2010.

[67] R. J. Marks, W. A. Dembski, and W. Ewert, *Introduction to Evolutionary Informatics*. Singapore: World Scientific, 2017.

[68] L. Brillouin, *Science and Information Theory*. New York: Dover, 1956.

[69] P. S. Chodrow, "Divergence, entropy, information: An opinionated introduction to information theory," *arXiv*, 2019.

[70] S. Verdú, "Relative entropy," *NIPS Conference*, 2009. [Online]. Available: https://videolectures.net/videos/nips09_verdu_re

[71] D. A. Díaz-Pachón, R. Gallegos, O. Hössjer, and J. S. Rao, "Statistical learning does not always entail knowledge," *Bayesian Analysis*, 2025.

[72] D. A. Díaz-Pachón and O. Hössjer, "Assessing, testing and estimating the amount of fine-tuning by means of active information," *Entropy*, vol. 24, no. 10, p. 1323, 2022.

[73] D. A. Díaz-Pachón, O. Hössjer, and R. J. Marks II, "Is Cosmological Tuning Fine or Coarse?" *Journal of Cosmology and Astroparticle Physics*, vol. 2021, no. 07, p. 020, 2021.

[74] ——, "Sometimes size does not matter," *Foundations of Physics*, vol. 53, p. 1, 2023.

[75] D. A. Díaz-Pachón, O. Hössjer, and C. Matthew, "Is it possible to know cosmological fine-tuning?" *The Astrophysical Journal Supplement Series*, vol. 271, no. 2, p. 56, April 2024.

[76] O. Hössjer, D. A. Díaz-Pachón, and J. S. Rao, "A formal framework for knowledge acquisition: Going beyond machine learning," *Entropy*, vol. 24, no. 10, p. 1469, 2022.

[77] O. Hössjer, D. A. Díaz-Pachón, and J. S. Rao, "Mathematical Modelling of Knowledge Acquisition and Consensus Formation in Populations," *Under review*, 2025.

[78] D. A. Díaz-Pachón and R. J. Marks II, "Generalized active information: Extensions to unbounded domains," *BIO-Complexity*, vol. 2020, no. 3, pp. 1–6, 2020.

[79] D. A. Díaz-Pachón, J. P. Sáenz, and J. S. Rao, "Hypothesis testing with active information," *Statistics & Probability Letters*, vol. 161, p. 108742, 2020.

[80] D. A. Díaz-Pachón, J. P. Sáenz, J. S. Rao, and J.-E. Dazard, "Mode hunting through active information," *Applied Stochastic Models in Business and Industry*, vol. 35, no. 2, pp. 376–393, 2019.

[81] T. Liu, D. A. Díaz-Pachón, J. S. Rao, and J.-E. Dazard, "High Dimensional Mode Hunting Using Pettiest Component Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4637–4649, April 2023.

[82] D. A. Díaz-Pachón, T. Liu, and J. S. Rao, "Lenses of variation," *Submitted*, 2025.

[83] D. A. Díaz-Pachón and R. J. Marks II, "Active Information Requirements for Fixation on the Wright-Fisher Model of Population Genetics," *BIO-Complexity*, vol. 2020, no. 4, pp. 1–6, 2020.

[84] S. Thorvaldsen and O. Hössjer, "Using statistical methods to model the fine-tuning of molecular machines and systems," *Journal of Theoretical Biology*, vol. 501, p. 110352, 2020.

[85] ——, "Estimating the Information Content of Genetic Sequence Data," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 72, no. 5, pp. 1310–1338, 2023.

[86] ——, "Use of directed quasi-metric distances for quantifying the information of gene families," *BioSystems*, vol. 243, p. 105256, 2024.

[87] S. Thorvaldsen, P. Øhrstrøm, and O. Hössjer, "The representation, quantification, and nature of genetic information," *Synthese*, vol. 204, p. 15, 2024.

[88] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*. Providence: American Mathematical Society, 2017.

[89] B. Carter, "Large Number Coincidences and the Anthropic Principle in Cosmology," in *Confrontation of Cosmological Theories with Observational Data*, M. S. Longhair, Ed. Dordrecht: D. Reidel, 1974, pp. 291–298.

[90] E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, May 1957.

[91] ——, "Information Theory and Statistical Mechanics II," *Physical Review*, vol. 108, no. 2, pp. 171–190, 1957.

[92] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[93] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research and Development*, vol. 5, no. 3, pp. 183–191, 1961.

Yanchen Chen obtained a Summa Cum Laude B.S. in Statistics from Stony Brook University, he also possesses a M.A. in Statistics from Columbia University, and is currently pursuing a Ph.D. in Biostatistics at the University of Miami. His research interests are machine learning, modeling, and information theory.

Daniel Andrés Díaz–Pachón received his Ph.D. in probability from the Institute of Mathematics and Statistics at the University of São Paulo. He is currently an Assistant Professor in the Division of Biostatistics at the University of Miami. His research lies at the intersection of probability, mathematical statistics, and information theory.