Which Similarity-Sensitive Entropy (S-entropy)?

Phuc Nguyen, Josiah Couch, Rahul Bansal, Alexandra Morgan,
Chris Tam, Miao Li, Rima Arnaout, and Ramy Arnaout

November 7, 2025

Contents

1 Introduction **2**

2 Related Work **2**
2.1 Traditional, similarity-insensitive entropy . . . 2
2.2 Similarity-sensitive entropy (S-entropy) . . . 4
2.2.1 The Leinster-Cobbold-Reeve framework (LCR) . . . 4
2.2.2 The Vendi score and its "cousins" (VS) . . . 4

3 Methods **5**
3.1 Datasets and definitions . . . 5
3.2 Similarity matrices . . . 6
3.3 Entropy calculations . . . 6
3.4 Clustering . . . 6
3.5 Numerical experiments . . . 6
3.6 Speed tests . . . 7

4 Results **7**
4.1 LCR and VS of major imaging and tabular ML datasets . . . 7
4.2 Trends with LCR and VS . . . 8
4.3 Bounds on LCR and VS . . . 9

5 Discussion **12**
5.1 Richer measures of the information that ML datasets contain . . . 12
5.2 The effect of scaling the similarity measure/similarity matrix . . . 13

6 Conclusion **14**

7 Acknowledgements **16**

A Appendix **16**
A.1 Python code for finding counterexamples to $VS \geq LCR$ . . . 16
A.2 Proof that $VS \geq LCR$ for $q \in \{2, 3, \infty\}$ . . . 17
A.3 Approach to a proof of $VS \geq LCR$ for all q . . . 18
A.4 Other bounds . . . 19

Abstract

A canonical step in quantifying a system is to measure its entropy. Shannon entropy and other traditional entropy measures capture only the information encoded in the frequencies of a system's elements. Recently, Leinster, Cobbold, and Reeve (LCR) introduced a method that also captures the far richer additional information encoded in the similarities and differences among elements, yielding similarity-sensitive entropy (S-entropy). More recently, the Vendi score (VS) was introduced as an alternative, raising the practical questions of how LCR and VS compare and which one to use. Here

we address these questions conceptually, analytically, and experimentally, using 53 machine-learning datasets. We show that LCR and VS values can differ by orders of magnitude and can capture complementary information about a system, except in limiting cases. We demonstrate that both LCR and VS depend on how similarities are scaled and introduce the concept of "half distance" to parameterize this dependence. We prove that VS provides an upper bound on LCR for several values of the Rényi-Hill order parameter and present evidence that this bound holds for all values. We conclude that VS is preferable only when interpreting elements as linear combinations of a more fundamental set of "ur-elements" or when the system or dataset possesses a quantum-mechanical character. In the broader circumstance where one seeks simply to capture the rich information encoded by similarity as part of one's entropy calculation, LCR is favored; nevertheless, for certain half-distances the two methods can complement each other.

1 Introduction

Entropy is the foundational quantitative descriptor of information, disorder, and uncertainty in a system and finds applications across science, engineering, commerce, and law [1, 2, 3, 4]. Traditional entropy, as formulated by Shannon [5] and generalized by Rényi [6] (and later Tsallis [7]), depends exclusively on the frequency distribution of a system's unique elements: these entropies represent the number of bits or nats required to encode the shape of the distribution, with smaller values for more skew distributions and larger values for flatter ones (Fig. 1). These measures rely solely on frequencies and therefore ignore a rich additional source of information about the system: the pairwise similarities and differences among its elements (Fig. 1).

Incorporating similarities into the calculation of entropy results in so-called similarity-sensitive entropy or S-entropy. This was first accomplished by Leinster and Cobbold [8] and extended by Reeve and colleagues [9], in a framework we refer to here as LCR [10]. By incorporating similarity, LCR can differentiate between systems that have identical frequency distributions but whose elements vary in how alike they are (Fig. 1). This capability is valuable in many domains [11], especially in machine learning (ML) [12, 13] (for example, to achieve state-of-the-art performance more efficiently [14]), where datasets generally consist of all-unique elements, meaning any two datasets that are the same size will have the same traditional (i.e. similarity-insensitive) entropies because their frequency distributions are both flat. S-entropy has proven to provide valuable insights into highly heterogeneous biological systems such as antibody and T-cell receptor (TCR) repertoires [15, 16], in which the vast majority of elements are unique, making the distributions heavy/long-tailed or nearly flat and therefore traditional entropies less informative.

More recently, an alternative form of S-entropy has been described, consisting of the Vendi score [17, 18] and its variants [19]—hereafter collectively VS—separate from LCR. This raises the practical question of which form of S-entropy to choose for a given situation: LCR or VS. The current work addresses this question from several angles—conceptual, empirical, and mathematical—including by measuring LCR and VS on a wide variety of medical and non-medical imaging and tabular datasets used for ML, and comparing their values.

2 Related Work

Entropies fall into two classes: traditional/similarity-insensitive and similarity-sensitive. Although one often sees reference to "the" entropy (either as a shorthand or referring specifically to Shannon entropy), entropies are actually families of measures; individual members are distinguished by how they weight frequencies and, for S-entropy, by their similarity matrix (see Table 1).

2.1 Traditional, similarity-insensitive entropy

The best known traditional entropy is Shannon entropy: $-\sum_i p_i \log p_i$ [5], where p_i is the frequency of unique element i. Shannon entropy can be thought of as a weighted average of element frequencies, with the weights being the logarithms of the frequencies themselves ($\log p_i$). Rényi [6] generalized this into a family of entropies H_α of order α—"deformations" of Shannon entropy—in which $\alpha > 1$ represents greater up-weighting of more-frequent elements relative to Shannon entropy; Shannon entropy itself is $H_{\alpha=1}$.

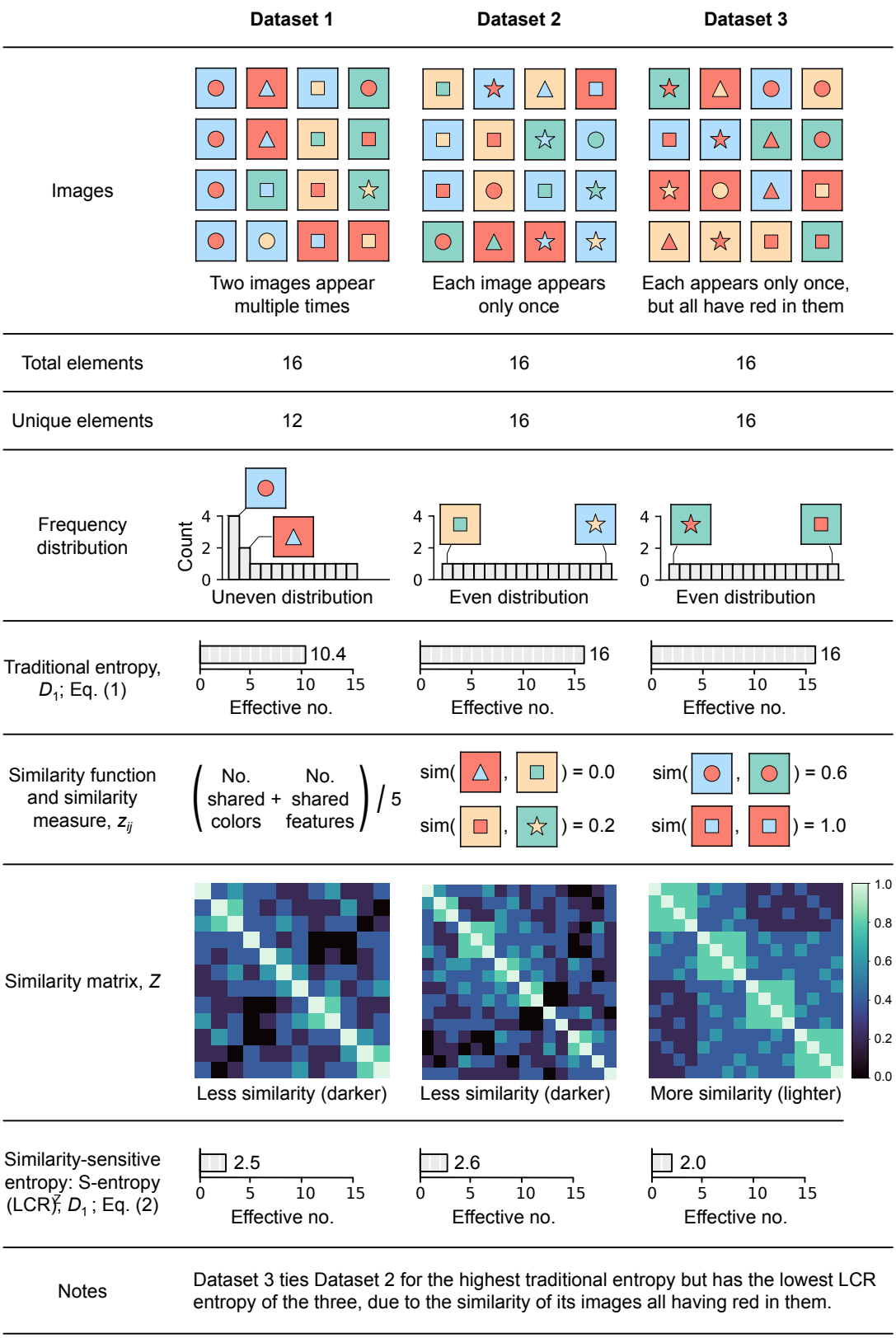

Figure 1: The concepts of element, frequency distribution, traditional entropy (at $q = 1$), similarity function, similarity measure (z_ij, with examples), similarity matrix (Z), and S-entropy (here, LCR, also at $q = 1$). Entropy values are expressed in effective-number form, i.e. in units of effective number of images present in the dataset. Different similarity measures can be chosen (see Section 5.2); in this example, the similarity measure is the normalized sum of the shared colors (0, 1, or 2) and shared features (outside color, inside color, and shape).

Hill showed that exponentiating Rényi entropy yields the "effective number" of distinct elements in the system [20]. Effective numbers, denoted D_q, use a parameter q (identical to Rényi's α) such that $q = 0$ counts the distinct types (by giving zero consideration to frequencies), while larger q give greater weight to more frequent elements. $D_1 = \exp(H_1)$ is Shannon entropy in effective-number form. Hill demonstrated that several other familiar statistics correspond to special cases of D_q for certain q, including Simpson's index for $q = 2$ and the Berger–Parker index for $q = \infty$ [21]. For this reason, Hill's formulation is often described as a unifying framework. D_q are known as the Hill numbers or D numbers (for "diversity") and are given by

$$D_q(\mathbf{p}) = \exp(H_\alpha(\mathbf{p})) = \begin{cases} \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{1-q}}, & q \neq 1, \\ \exp\left(-\sum_{i=1}^{n} p_i \ln p_i\right), & q = 1. \end{cases} \tag{1}$$

where $\mathbf{p}$ denotes the frequency distribution of unique elements. Other generalizations such as Tsallis entropy [7] also exist. What these traditional entropies have in common is that they are similarity-insensitive: they depend solely on $\mathbf{p}$ and ignore any relationships among the elements.

2.2 Similarity-sensitive entropy (S-entropy)

2.2.1 The Leinster-Cobbold-Reeve framework (LCR)

LCR [10] extends Hill's framework by incorporating information about the similarities and differences of the elements within the system, which traditional entropy does not. A similarity matrix Z is introduced, with entries $z_{ij} \in [0, 1]$ that quantify the similarity between elements i and j. The similarity of each element to itself is set to 1, making Z's diagonal entries 1. The resulting quantities, denoted D_q^Z, are the exponentials of similarity-sensitive versions of the Rényi entropies H_α^Z and are given by

$$D_q^Z(\mathbf{p}; Z) = \exp\big(H_\alpha^Z(\mathbf{p}; Z)\big) = \begin{cases} \left(\sum_{i=1}^{n} p_i (Z\mathbf{p})_i^{q-1}\right)^{\frac{1}{1-q}}, & q \neq 1, \\ \exp\left(-\sum_{i=1}^{n} p_i \ln (Z\mathbf{p})_i\right), & q = 1. \end{cases} \tag{2}$$

Here $(Z\mathbf{p})_i = \sum_j z_{ij} p_j$ is the frequency-weighted average similarity of element i to all the elements (including itself). When this average is large, element i is considered ordinary, making Eq. 2 interpretable as the average "ordinariness" [8] across all elements. $D_q^Z(\mathbf{p})$ appears widely in the recent literature, where it is variously known as:

- *phylogenetic diversity* [22] in the special case where z_{ij} forms an ultrametric, typically derived from a phylogenetic tree;
- *functional diversity* [23] when the similarity pertains to elements' function, for example the binding similarity between pairs of antibodies or TCRs [16];
- *attribute diversity* [23] when interpreting the system as a set of attribute contributions instead of frequencies; and
- *similarity-sensitive* or *similarity-aware diversity* more generally [11].

LCR has proven useful for describing many complex systems whose empirical samples are uniform or close to uniform—where unique elements' frequency distribution is flat or nearly so—a regime where traditional entropies become uninformative (Fig. 1). Representative applications in the life sciences include high-throughput immunology (immunomes) [16], microbiome research (metagenomics) [11], and medical imaging [12] Specifically in ML contexts, where training sets (e.g., image collections) are often composed of unique observations (e.g. images), LCR has been shown to help identify performance predictors beyond simple dataset size or class balance [12].

2.2.2 The Vendi score and its "cousins" (VS)

VS constitutes a related but separate class of similarity-sensitive entropy measures. Like LCR, VS entropies are functions of a similarity matrix, but now the matrix has dimensions $n \times n$, where n is the number of observations or *total* elements; for this reason we refer to it as Z_n to distinguish it from the Z used in LCR, which has one row/column per *unique* element. (When all elements are unique, Z and

Entropy type and formula	Frequency weighting	Similarity-sensitive?	Notes and applications
Shannon (Boltzmann–Gibbs) $H_1(\mathbf{p}) = -\sum_{i=1}^{S} p_i \log p_i$	$q = 1$	No	Information theory, Ecology, ML loss functions, Thermodynamics
Rényi entropy (α) $H_\alpha(mathbfp) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha$	Any q	No	Info-theoretic security, Ecology, Fractal analysis, Physics
Tsallis entropy (q-entropy) $S_q(\mathbf{p}) = \frac{1}{q-1}\left(1 - \sum_i p_i^q\right)$	Any q	No	Non-extensive statistical mechanics, Turbulence, Astrophysics
Quantum (von Neumann) entropy $S(\rho) = -\operatorname{Tr}(\rho \log \rho)$	$q = 1$	No	Quantum information, Entanglement, Quantum thermodynamics
LCR $D_q^Z(\mathbf{p}; Z) =$ $\left(\sum_i p_i (Z\mathbf{p})_i^{q-1}\right)^{1/(1-q)}$ for $q \neq 1$ $\exp\left(-\sum_i p_i \log(Z\mathbf{p})_i\right)$ for $q = 1$	Any q	Yes	Phylogenetic, attribute, and functional diversity are special cases. Biodiversity, microbiomes, language...
Vendi score $\mathrm{VS}(\mathbf{p}_n; Z_n) = \exp\left\{\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right)\right\}$	$q = 1$	Yes	λ_i are eigenvalues of $Z_n = Z/N$ (positive semi-definite). ML diversity, Generative model evaluation, Ecology
Cousins of the Vendi score $\mathrm{VS}_q(\mathbf{x}, \mathbf{k}) = \left(\sum_i \lambda_i^q\right)^{\frac{1}{1-q}}$	Any q	Yes	Quantum-inspired stats, Class diversity profiling
Other variants Burg: $\sum_i \log p_i$ KL: $D_{KL}(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$ CRE: $-\int_0^\infty \bar{F}(x) \log \bar{F}(x)\, dx$	Partly	Partly	Spectral estimation (Burg); Bayesian inference (KL); Survival analysis (CRE)

Table 1: Summary of major entropy families, their frequency-weighting forms, similarity sensitivity, and main applications.

Z_n are the same.) We can define Z_n in terms of Z as

$$(Z_n)_{i,j} = Z_{s(i),s(j)}. \tag{3}$$

where $s(i)$ is the unique element of which the i^{th} overall element is an instance or example.

The original VS is defined as the exponential of the Shannon entropy of the eigenvalues λ_i of Z_n/n. The division by n normalizes Z_n to unit trace (because as in Z, the diagonals of Z_n equal 1):

$$\mathrm{VS}(\mathbf{p}_n; Z_n) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right) \tag{4}$$

VS is to this similarity matrix what the von Neumann entropy is to the quantum density matrix. Replacing the Shannon term in Eq. 4 with a Rényi entropy of order q yields the so-called "cousins" [19] of the Vendi score:

$$\mathrm{VS}_q(\mathbf{p}_n; Z_n) = \left(\sum_{i=1}^{n} \lambda_i^q\right)^{\frac{1}{1-q}} \tag{5}$$

VS_q is a 1-parameter family of scores with q as in D_q^Z.

Note that LCR can also be written as a function of Z_n if desired:

$$D_q^Z(\mathbf{p}; Z) = D_q^Z(\tilde{\mathbf{p}} = \frac{1}{n}; Z_n) \tag{6}$$

where $\tilde{\mathbf{p}}$ is the uniform distribution on the n elements.

3 Methods

3.1 Datasets and definitions

The following imaging and tabular datasets were used for benchmarking LCR and VS:

- **Image datasets.** 22 large ($\geq 10,000$-image) medical and non-medical vision benchmarks were used: MNIST, Fashion-MNIST, CIFAR-10 and CIFAR-100 (from the `torchvision` Python package); BloodMNIST, ChestMNIST, OctMNIST, OrganAMNIST, OrganCMNIST, OrganSMNIST,

PathMNIST, and TissueMNIST from the MedMNIST collection [24]; the Amphibia, Insecta, Mammalia, Plantae, and Reptilia subsets from iNaturalist [25]; the computed tomography (CT) and ultrasound (US) subsets from RadImageNet [26], the NIH ChestXRay dataset [27]; COCO (Common Objects in Context) [28]; and MIDRC COVIDx CXR-4 [29]. For computational convenience while still preserving content, images in the iNaturalist collection as well as the MIDRC COVIDx CXR-4 dataset were downscaled to 250×250 pixels, and images in the COCO datast were scaled to 100×100 pixels; also for convenience, for each dataset, 10,000 images were selected uniformly at random for LCR and VS calculations.

- **Tabular datasets.** After selecting for a maximum of 30 columns and 200,000 rows, the 31 most frequently downloaded datasets from the University of California Irvine Machine Learning Repository (UCIML) were retrieved [30]. For each dataset, identifier columns and any features containing non-numeric strings or time-series data were removed.

Definitions (Fig. 1). The terms "system" and "dataset" are used interchangeably. The elements of image and tabular datasets are single images and single rows, respectively. Each element appears only once in these datasets; i.e. each element is also a unique element.

3.2 Similarity matrices

Similarity between each two (unique) elements $\mathbf{x}_i$ and $\mathbf{x}_j$ was calculated from the Euclidean distance (L2 norm) as follows:

$$z_{ij} = e^{-k||\mathbf{x}_i - \mathbf{x}_j||_2} \tag{7}$$

where k is a parameter that was varied during experiments to test for effects on results. This similarity measure is provably positive-definite (PSD), guaranteeing valid similarity matrices for VS (see Section 6); LCR has no such restriction [10]. Imaging datasets were first embedded into a two-dimensional space using UMAP [31] and the default k taken to be $2^{-1/2}$, making the norm in Eq. 7 equivalent to the root-mean-square distance (RMSD). The default k for tabular datasets was 1.

3.3 Entropy calculations

The `greylock` Python package [11] was used to compute LCR as the γ diversity; this is the effective-number form of the corresponding entropy (see Section 2.1). The package accepts a user-provided similarity matrix and returns results for any q; we primarily report results for $q = 1$ (Shannon-type LCR) for direct comparison with the default VS (see Section 2.2.2). VS was calculated using the `vendi-score` package [32]. This implementation computes the eigenvalues of the similarity matrix, retains the positive spectrum, and evaluates the exponential of the Shannon entropy of those eigenvalues, yielding a scalar appropriate for comparison to LCR. Similarity matrices were calculated on a GPU-enabled workstation equipped with multi-core CPUs. LCR and VS computations were performed on a 2024 M4 Apple Mac mini with 64GB unified memory.

3.4 Clustering

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [33] was used to obtain a non-entropic estimate of the effective number of unique elements as the number of clusters present in each dataset. HDBSCAN adapts to varying densities and does not require *a priori* specification of the cluster count, making it well-suited for heterogeneous data.

3.5 Numerical experiments

Following a previous approach [34], we searched for counterexamples to Conjecture 1 (Section 4.3) by numerically minimizing $VS_q(Z) - D_q(Z, \frac{1}{n})$ for uniform $\mathbf{p}$ using `scipy.optimize`. A negative value indicates a counterexample. To meet the assumptions of this conjecture, Z was constructed as a gram matrix (which are always PSD), finding an $r \times n$ matrix w, for some integer r, such that $Z = w^\intercal w$. To ensure that the diagonal elements of Z are equal to 1, the rows of w were r-dimensional unit vectors; to ensure the entries of Z are non-negative, we took the entries of w to be non-negative. We assumed a uniform distribution of unique elements. Thus, we parameterized Z by a $r \times n$ matrix x, from which we

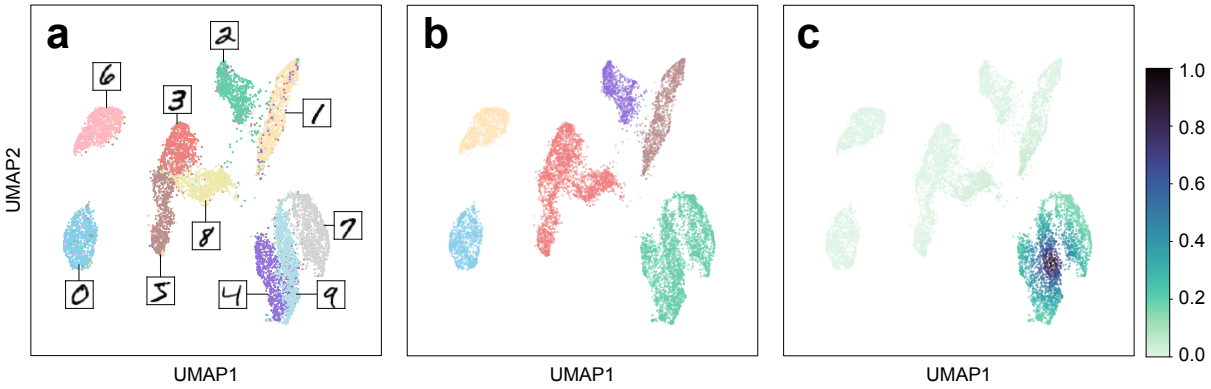

Figure 2: The first 10,000 images of the MNIST digits dataset colored (a) by digit (labeled using representative images), (b) by HDBSCAN cluster, and (c) by similarity to one of the images (a "9"). The effective number of images in this dataset is 12.5 by LCR and 95.9 by VS.

defined

$$Z = w^{\intercal}w; w_{ij} = e^{x_{ij}}(\sum_{k} e^{2x_{kj}})^{-1/2} \tag{8}$$

We minimized the difference $VS_q(Z) - D_q(Z, \frac{1}{n})$ for each combination of $q \in \{-10, -2.7, -2, -1, -0.3, 0, 0.3, 1, 2.7, 3.4, 4, 5, 2\pi, 7, 8, 9.1, 10\}$, $n \in \{3, 5, 10, 20\}$, $r \in \{2, \lfloor n/2 \rfloor, n\}$ if $n > 3$ and $r \in \{2, 3\}$ for $n = 3$ (to avoid $r = 1$). We also varied the distribution from which the initial guess fed to `scipy.optimize` was drawn. In total, 4420 minimization were done. The full Python code for performing these minimizations can be found in section A.1

3.6 Speed tests

Run times were calculated using the Python `timeit` package with 5-fold replications and measured on a 2024 M4 Apple Mac mini with 64GB unified memory.

4 Results

4.1 LCR and VS of major imaging and tabular ML datasets

We began by comparing LCR to VS, both at $q = 1$, on a simple, well-known imaging dataset: MNIST handwritten digits; the results below are for a random 10,000-image subset, taken for computational convenience. Both quantities were expressed in their effective-number forms to enable comparison with an independent benchmark: the number of clusters obtained by the state-of-the-art clustering algorithm HDBSCAN.

Fig. 2a displays a UMAP embedding of the data, coloring each digit by its label. HDBSCAN detects six clusters, grouping together digits that appear visually similar (Fig. 2b). LCR yields the effective number of unique images after accounting for pairwise image similarity; it allows visually similar images to each contribute less to the entropy than if they were completely distinct (Fig. 2c). Consequently, points that lie close together count less toward the total, whereas points separated by some distance contribute more independently, even for points that happen to be in the same cluster. Taken together, these considerations suggest that LCR will approximate the number of clusters—potentially exceeding it slightly because the clusters are somewhat loose. Indeed, the similarities among the images in this 10,000-image subset result in it having an effective number of just 12.5 images according to LCR: roughly twice the number of clusters and 25% greater than the number of classes (10). Mathematically, this indicates that this MNIST subset possesses the same entropy as a hypothetical dataset containing 12-13 completely dissimilar elements (i.e. with zero pairwise similarity). This result can be rationalized as follows: the many stylistic variations of certain digits—e.g., the different ways to write a 1, 4, or 7—increase the S-entropy, while visual similarities among digits—such as the similarity between many 4s, 7s, and 9s, which cluster together per HDBSCAN—decrease the entropy relative to a scenario in which all 10 digit classes were entirely distinct.

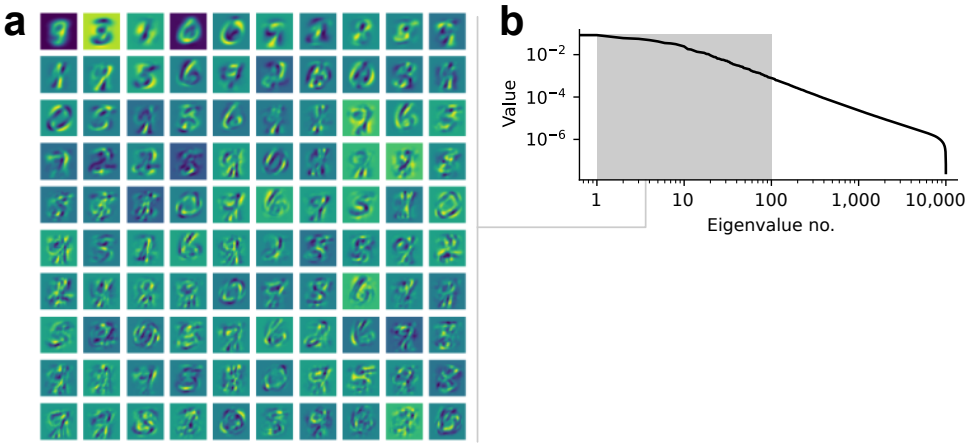

Figure 3: (a) Top 100 eigenimages for the MNIST digits dataset and (b) their eigenvalues (in the gray region), alongside the full eigenvalue spectrum. Note the log axes.

In contrast, VS for this dataset was substantially higher, yielding an effective number of 95.9 elements. However, its interpretation differs from that of LCR: the VS value indicates that the images possess the same *traditional* entropy as a collection of roughly 96 mutually orthogonal eigen- or "ur"-images, each actual image in the dataset being a linear combination of these. Fig. 3a displays the 100 eigen-images with the largest eigenvalues, which dominate the total entropy (Fig. 3b). Some of these eigen-images resemble recognizable digits, whereas others resemble superpositions of multiple digits. VS is precisely the entropy of the eigenvalue spectrum, which can be interpreted as the frequency distribution of a hypothetical dataset composed of eigen-images at frequencies given by the eigenvalues. The total number of eigen-images equals the total number of images in the subset (10,000); the fact that the VS effective number is much smaller than this implies that only a handful of eigenvalues are large, while most are small.

Fig. 4 shows LCR and VS values for 53 well known machine-learning datasets, including medical and non-medical imaging and tabular datasets.

4.2 Trends with LCR and VS

Fig. 5 presents comparisons among LCR, VS, and the number of HDBSCAN clusters across all datasets (for default k at $q = 1$; Figs. 6 and 7 investigate other ks and qs), with imaging and tabular datasets evaluated separately. As observed for MNIST, both LCR and VS were generally larger than the cluster count. The principal exceptions were the CIFAR image datasets, where the absence of meaningful substructure produced hundreds of very small clusters. Correlations between the number of clusters and LCR, and between the number of clusters and VS, were both low ($R^2 = 0.05$ and 0.06, respectively), indicating that these entropy measures capture information not captured by clustering. In contrast, LCR and VS correlated strongly with each other for both imaging and tabular datasets ($R^2 = 1.00$ and 0.90, respectively). Fig. 7 shows results for $q = 0$ and ∞.

The picture changed when the similarity matrix Z was scaled by varying the parameter k in Eq. 7. Fig. 6 shows results for tabular datasets for three k values spanning 15 orders of magnitude—$2.6 \times 10^{-8}, 2.0 \times 10^{-1}$, and 8.4×10^{7}—selected to bracket the reciprocals of the minimal and maximal entries of the distance matrices. We refer to k as the half-distance (by analogy with half-life): consider elements i and j separated by a distance d_{ij} with similarity z_{ij}; increasing their separation by an additional k halves the similarity $z_{ij}/2$.

When k is very small, all z_{ij} approach 1, collapsing the system to a single individual for both LCR and VS (Fig. 6a and c, left), yielding a perfect correlation ($R^2 = 1.00$). When k is very large, all off-diagonal z_{ij} approach 0, making every element essentially unique; consequently, both LCR and VS converge to the number of elements in the dataset, and the correlation again is perfect ($R^2 = 1.00$) (Fig. 6a and c, right). Only at intermediate k values do LCR and VS provide independent information (Fig. 6a and c, middle).

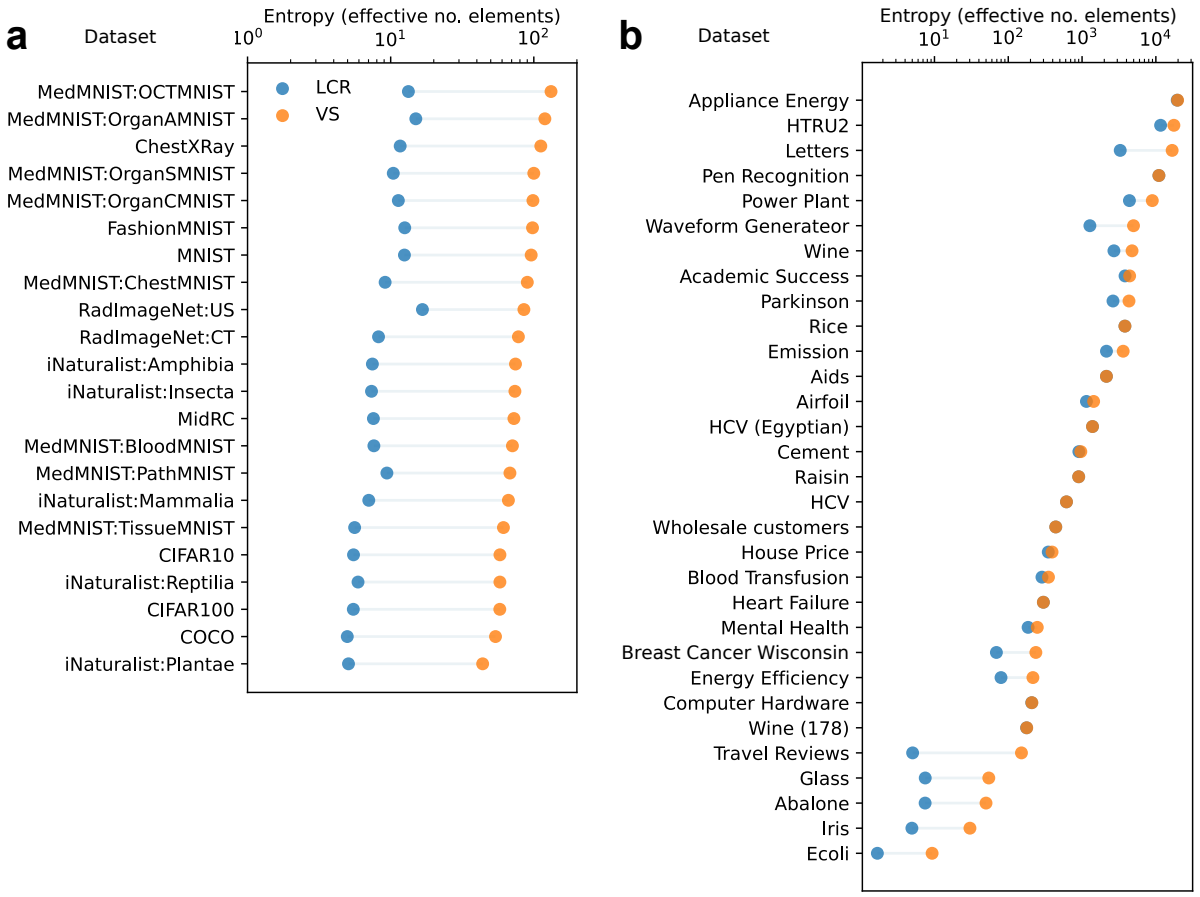

Figure 4: LCR and VS for (a) imaging and (b) tabular datasets, sorted by VS ($q = 1$; default values of k for each dataset type).

To further illustrate this pattern, Fig. 6b and d display R^2, the Pearson correlation, and the corresponding covariance as functions of k for imaging and tabular datasets. For tabular datasets at $q = 1$, the R^2 and correlation curves exhibit several relative minima; the absolute minima—$R^2 = 0.68$ and correlation $= 0.83$—did not occur at the same k value that yields the maximal covariance, near $k = 0.2$. This k value for the middle panels of Fig. 6a and c was chosen for this reason. We observed different-shaped curves for $q = 0$ and $q = \infty$, and the values of k at the extrema had no obvious connection to each other or to the extreme regimes and also differed between the tabular and imaging datasets; for a generic similarity matrix Z, it is unclear how to predict their location.

4.3 Bounds on LCR and VS

In Figs. 5 and 6, VS values were consistently higher than LCR, and both measures converge to 1 as the half-distance decreases. It is therefore reasonable to ask whether the inequality $VS \geq LCR$ holds universally. We formalized this question as the following conjecture:

Conjecture 1. *Let Z be an $n \times n$ $(n > 0)$ PSD similarity matrix, i.e. a PSD symmetric matrix with entries between 0 and 1 whose diagonal entries equal 1. Denote LCR and VS at order q as $D_q(Z, p = \frac{1}{n})$ and $VS_q(Z)$, respectively, on the uniform distribution. Then for all $q \in [-\infty, +\infty]$*

$$VS_q(Z) \geq D_q(Z, \frac{1}{n}) \tag{9}$$

Note that in the context of this conjecture, since we are talking about the LCR diversity on a set with the uniform distribution, there is no distinction between Z and Z_n (like that made in section 2.2.2), as these are equal. We prove this for $q = 2, 3$, and ∞ in Appendix A.2. We consider these proofs strong evidence that Conjecture 1 holds for all $q \geq 2$, since it would be surprising for VS to bound LCR at

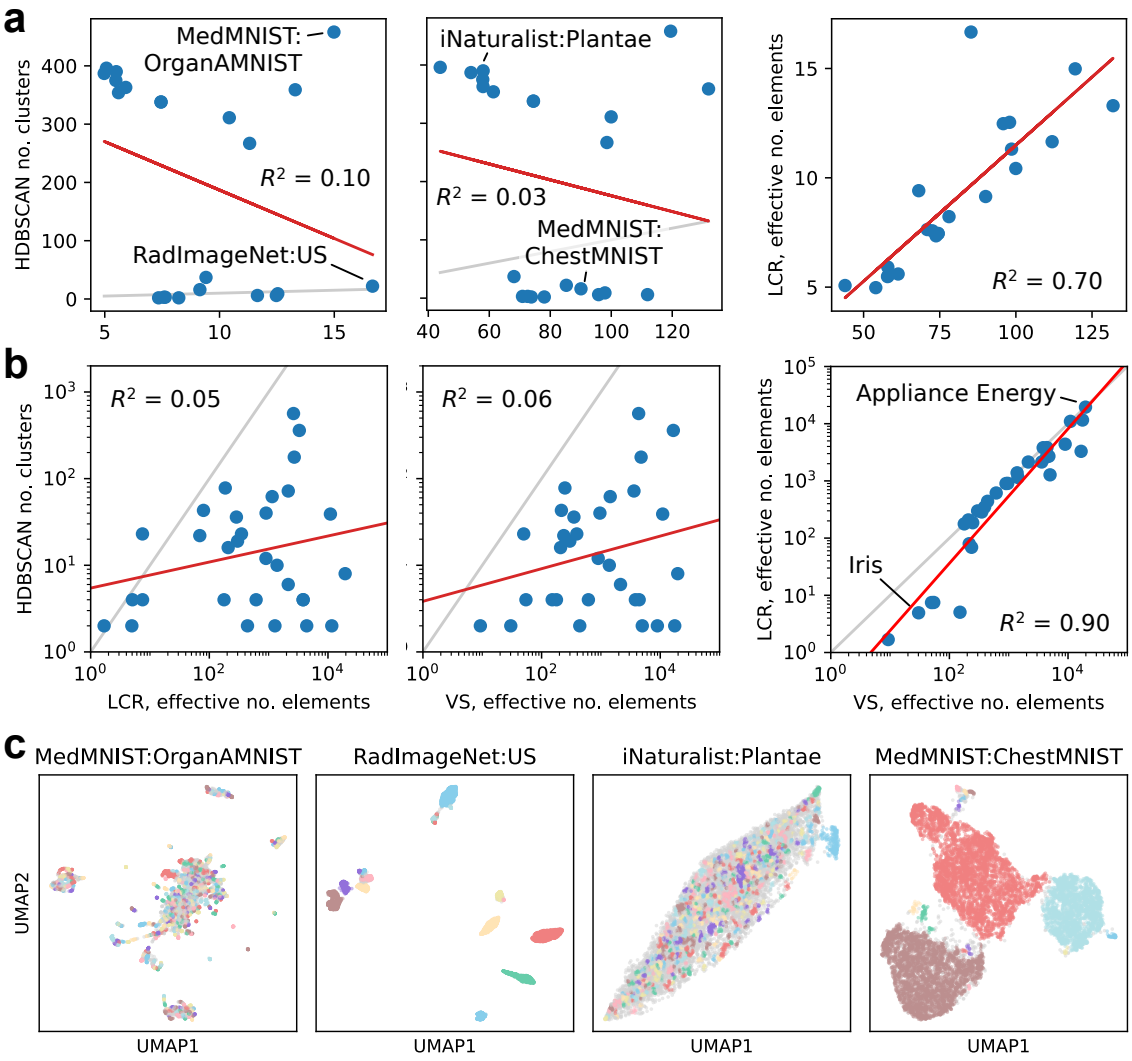

Figure 5: Correlations among LCR, VS, and HDBSCAN for (a) imaging and (b) tabular datasets. Each point represents a dataset. All entropies are effective-number forms at $q = 1$. Red line = linear regression fit; gray line = 1:1 (off the scale to the left for (a), far right). Select datasets are labeled. (c) UMAPs for labeled imaging datasets. Gray points = unclustered elements.

$q = 2, 3$, and ∞ but not $q = 2.5, 4$, or 27.3. The following lemma supports this conjecture holding for all q:

Lemma 1. *When Z is full rank but otherwise as in conjecture 1, for all $q \in [-\infty, 0]$*

$$VS_q(Z) \geq n \geq D_q(Z, \frac{1}{n}) \tag{10}$$

Proof. By definition, $VS_0(Z) = \text{rank}(Z) = n$. Rényi entropies are known to decrease monotonically with q [10]. Generalized VS is simply the Rényi entropy of the eigenvalues of Z/n. This implies

$$D_{-\infty}(Z, \frac{1}{n}) \geq D_q(Z, \frac{1}{n}) \tag{11}$$

Define $r_i = \sum_j Z_{ij}$. Then $r_i \geq 1$, since the entries of Z are non-negative, and the diagonal entries (one of which contributes to r_i) are all 1. Thus

$$D_{-\infty}(Z, \frac{1}{n}) = \frac{n}{\min r_i} \leq n \tag{12}$$

□

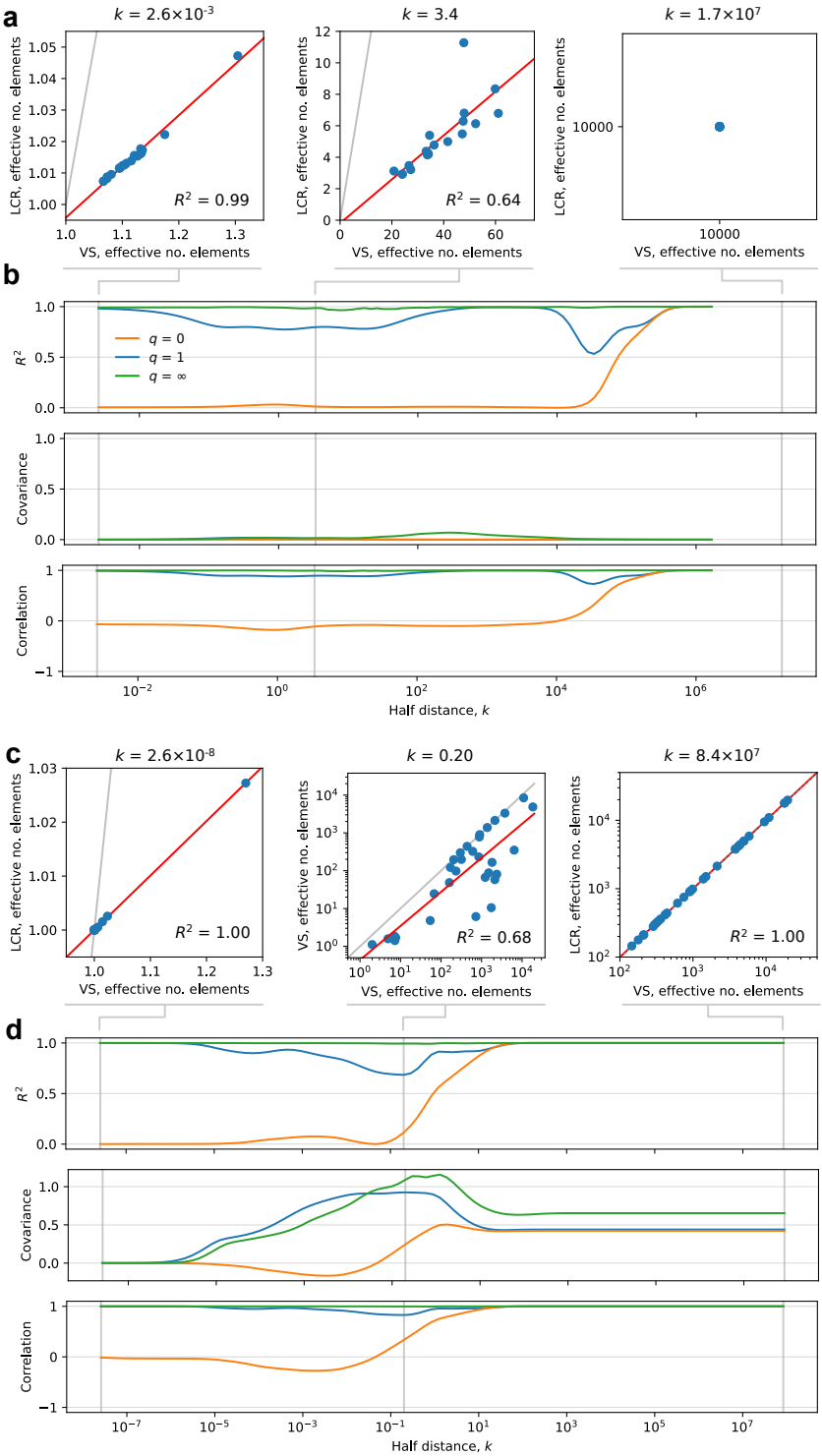

Figure 6: LCR vs. VS at $q = 1$ at a very small half distance, an intermediate half-distance for which the correlation is low, and a very large half distance, together with the R^2, covariance, and correlation for $q = 0, 1$, and ∞ for (a-b) imaging and (c-d) tabular datasets. For tabular datasets, the intermediate half-distance is the one simultaneously minimizing the R^2, the correlation, and maximizing the covariance at $q = 1$. For image datasets, the intermediate half-distance does not coincide with the global minimum of the correlation at $q = 1$, because VSs and LCRs already come very close to 10,000 at that global minimum, so we chose a smaller value. (a) and (c): red line = linear regression fit; gray line = 1:1. Each point represents a dataset. (b) and (d): vertical gray lines = ks for (a) and (c).

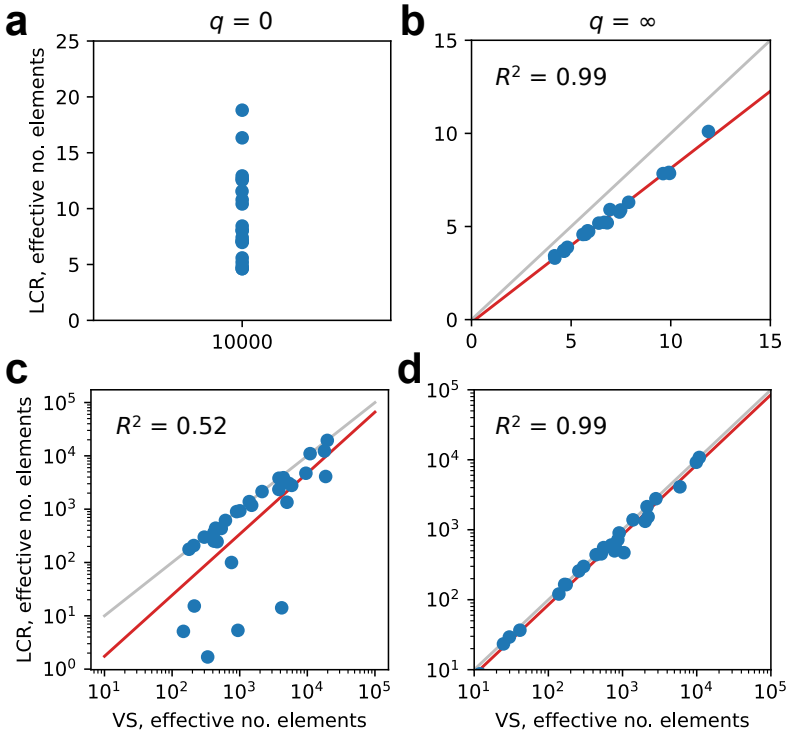

Figure 7: LCR vs. VS for $q = 0$ and ∞ for (a-b) imaging and (c-d) tabular datasets at their default k values (1 and $\approx 2^{-1/2}$, respectively). Each point represents a dataset.

A computational search of $\approx 4,000$ examples involving similarity matrices of up to size 50×50 failed to find counterexamples for any of several values of q, including $q = 0$ and $q = 1$. This search used a method that successfully found a counterexample to a different inequality previously [34], adding further support for Conjecture 1 being true. We present an approach to a proof for all q in Appendix A.3 and present an additional proof on bounds in Appendix A.4.

5 Discussion

Quantifying the information contained in a system is essential for virtually any complex system. This includes ML datasets, where the fact that most elements are generally unique—no repeated images (in image datasets) or rows (in tabular datasets)—renders traditional entropy (Shannon entropy and the other Rényi entropies) uninformative. This is because traditional entropy measures information encoded by the shape of the elements' frequency distribution, but when the elements each appear only once, this distribution is flat (Fig. 1): there is no additional information there. Traditional entropy *is* used to measure the distribution of class sizes, which often vary; this is class balance. However, there is no obvious way to use class balance to measure the the information contained within or between classes, which is a property of the elements themselves.

5.1 Richer measures of the information that ML datasets contain

One way to get at this missing information is to use clustering algorithms, to quantify dataset substructure as the number of clusters and possibly clusters' relative sizes (which can be measured by traditional entropy). Like LCR and VS, which require a choice of similarity measure and q, clustering algorithms require choices. For example, HDBSCAN requires choosing `min_cluster_size`, `min_samples`, etc. Conceptually, the main difference between clustering and S-entropy is that S-entropy accounts for similarities both within and between clusters, whereas most clustering algorithms (and classifiers) lack this ability: effectively, the within-cluster (or intra-class) similarity is always 1 and the between-cluster similarity is always 0.

The exception is fuzzy clustering, in which each element can belong to more than one cluster, with each

element being parameterized by its strength of membership for one or more clusters. One can view LCR as carrying fuzzy-clustering to its logical limit, with each unique element being its own cluster and the similarities giving other elements' strength of membership to that cluster and thereby contributing to the size of that cluster. This is a fuzzy-clustering interpretation of the ordinariness, $Z\mathbf{p}_i$: a big cluster that many unique elements belong to fairly strongly is "ordinary" (Fig. 2c). This perspective provides another way of seeing that LCR is the entropy of a system after accounting/adjusting/correcting for the similarity among its elements.

Traditional entropy accounts for frequency; LCR accounts for both frequency and similarity. In this way, LCR is a richer, more accurate measure of the information in a system. Note that, regardless of whether they are represented as effective numbers (as we [11] and others [21, 20, 10] have advocated) or as bits of information, LCR is always bounded above by the corresponding traditional entropy [10]; for example, at $q = 1$ LCR never exceeds Shannon entropy. Similarity saps individual elements of some of their uniqueness, but in doing so, it reveals differences between datasets that traditional entropy cannot (Fig. 1).

How do LCR and VS compare with the number of clusters identified by HDBSCAN? Predicting the relationship is difficult: large clusters whose peripheral elements are relatively distinct tend to increase the LCR/VS values, whereas strong similarities among elements—both within a cluster and across neighboring clusters—tend to reduce them (see Figs. 1, 5c). Indeed, many datasets examined exhibited a pronounced difference in magnitude between tabular and imaging datasets when comparing the HDBSCAN cluster count with LCR and VS. For tabular data sets, both LCR and VS almost always surpassed the HDBSCAN cluster count, often by one to two orders of magnitude, reflecting that many rows remain unclustered by hard clustering.

In contrast, for imaging datasets, LCR and VS were typically one to two orders of magnitude *lower* than the number of clusters, indicating substantial similarity among clusters that hard clustering does not capture (Fig. 2c). We interpret these systematic differences as evidence of the advantages of similarity-sensitive entropy relative to hard clustering, for describing a system's structure. We do not regard this as a failure of HDBSCAN; we note as a limitation of our study that we did not explore a range of HDBSCAN hyperparameters, since the goal was simply to illustrate the general relationship between similarity-sensitive entropy and clustering. However, we note that by the "eye test," UMAP plots validate HDBSCAN's clustering, suggesting hyperparameter search would be unlikely to meaningfully affect our findings.

5.2 The effect of scaling the similarity measure/similarity matrix

We demonstrated that the relationship between LCR and VS depends strongly on how the similarity-matrix values are scaled, which we illustrated by introducing the half-distance parameter k. Recall that similarity between two elements ranges from 0 to 1; scaling modifies this relationship by raising the values to a power, thereby shifting the curve upward or downward in the middle, while maintaining the anchoring at each end (at 0 and 1).

Why should one need to scale the similarity between elements in the first place? Should scaling be permitted at all? When similarity is derived from external knowledge or some principle or outside requirement, we argue that scaling should *not* be applied: the similarity function is defined *a priori*, its output for any given pair is fixed, and the similarity matrix simply collects these values for all pairs. An example is binding similarity in immune repertoires, defined as the relative dissociation constants of two antibodies (or TCRs) for a given antigen (or set of antigens) [16]. This is the unique definition that preserves biophysical additivity; arbitrary scaling would break this relationship and thereby produce a similarity matrix that is objectively incorrect.

In contrast, when similarity is heuristic and not constrained by such principles, as with the Euclidean-distance-based similarity measures used here, there is no *a priori* reason to privilege the default similarity ($k = 1$) over other values of k. For instance, RMSD provides a perfectly valid similarity definition and corresponds to Euclidean distance with $k = 1/\sqrt{d}$, where d is the dimension of the underlying data (number of columns for tabular data, number of pixels × number of color channels for image data, 2 for the umap embedding, etc). Therefore, in heuristic cases, scaling can serve as a reasonable dial that investigators can adjust to obtain values that make the most sense given the context and/or are

maximally informative about the system.

Such choices arise often in measurement. For example, we note that the parameter q offers a loosely analogous scaling to that used in traditional (and similarity-sensitive) entropy, allowing investigators to adjust the emphasis on frequency. Lest choices feel too arbitrary, remember that specific integer values of q correspond to well-known statistics; for example, $q = 2$ corresponds to Simpson's index, which reflects the statistics of collision frequency. It is possible that alternative values of k could similarly be linked to other quantities of interest. Investigating this possibility is left for future work.

A key feature of scaling is that it interpolates similarity-sensitive entropy between 1 (for small k) and traditional entropy (for large k). When k is small, the off-diagonal entries of the similarity matrix approach 1, implying that all unique elements are mutually similar; in the limit the system behaves as if it contained a single unique element. Conversely, for large k the off-diagonal entries approach zero, yielding the identity matrix in the limit; the unique elements become completely dissimilar, and traditional entropy is recovered. In a system where the n elements are all unique—typically the case for ML data sets, including the imaging and tabular data sets examined here—the effective-number form of traditional entropy equals the dataset size. Consequently, for such uniform systems k interpolates both LCR and VS from 1 to n. Other scaling approaches may exhibit the same property.

So: for heuristic similarity functions, which k should one choose? Qualitatively, the answer is a value roughly midway between the extremes (Fig. 6). However, our quantitative findings indicate that the optimal choice is not obvious. We found that the correlation between LCR and VS is minimized at $k = 0.2$ for tabular data sets and 3.4 for imaging data sets—not at any obvious definition of the midpoint of the range (defined for example as the geometric mean of the k values where the correlation first falls to 0.99 on each side). Moreover, the shapes of these curves, with their multiple extrema, do not appear to follow a simple parametric form, making it difficult to predict precisely where the two metrics are maximally mutually informative. We also observed that even at the k that minimizes the LCR-VS correlation, the R^2 between LCR and VS remained relatively high in absolute terms, at 0.68. Perhaps this was to be expected, since both metrics quantify the effective number of elements using the same underlying similarity measure, albeit in different ways (see Section 6). The dependence on k, however, was quite strong.

Our results suggest first testing whether the entropy lies near 1 or n; if so, to explore, if possible, different k values to delineate a non-trivial range; and finally to select a k approximately in the middle of this range (on a logarithmic scale). Both the choice of similarity measure and the value of k have effects on the resulting entropy that cannot be ignored. Thus to some extent, the information in a system is in the eye of the beholder.

6 Conclusion

So: LCR or VS? Our results support defaulting to LCR when the goal is to measure the similarity-adjusted entropy of a system, and to use VS only in one of two special cases: (*i*) when the elements are usefully thought of as linear combinations of a set of mutually orthogonal ur-elements/eigen-elements, or (*ii*) when the system possesses a quantum-mechanical character (see below). If the aim is a more general multidimensional characterization, there is no clear reason to avoid employing both metrics; keeping in mind that they are likely to correlate, our findings recommend selecting k within the non-trivial range (see Section 5.2) to maximize their independence and thereby their combined informativeness.

We note that, quantitatively, VS will be larger than LCR for any given k and q. We observed this ordering empirically and proved it for a broad class of cases, outlining a proof for the remaining situations; a computational search uncovered no counter-examples. This relationship can be rationalized by recognizing that, although VS is a similarity-sensitive entropy, its computation includes a traditional entropy component: after incorporating similarity to derive the eigen-elements, VS is the traditional entropy of the resulting eigenvalues; traditional entropy is always greater than or equal to similarity-sensitive entropy.

From a practical perspective, LCR has three advantages. First, it is less computationally expensive whenever elements appear multiple times (i.e., when $\mathbf{p}$ is not flat/uniform), because the similarity matrix for LCR has order equal to the number of *unique* elements, whereas VS requires a matrix whose order

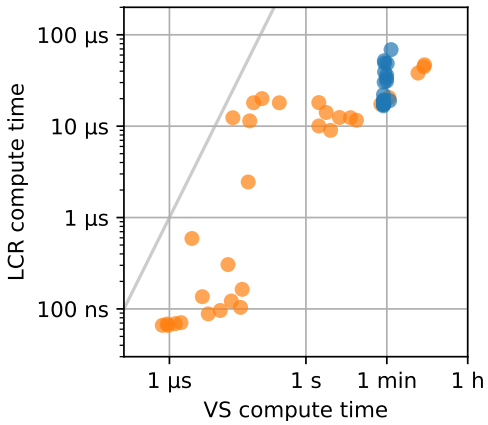

Figure 8: Compute times for all 53 datasets for LCR (using the `greylock` Python package [11]) vs. VS (using `vendi-score`). Imaging datasets were all 10,000 images. Smaller datasets were faster. Note the different timescales.

equals the *total* number of elements; both calculations scale as the square of the matrix order. Empirically, we observed a speed advantage of `greylock` over `vendi-score` even when all elements were unique, a situation that frequently occurs in ML datasets (Fig. 8). This performance gap may narrow or widen as codebases evolve.

Second, LCR does not require the similarity matrix to be PSD, whereas VS does. Similarity matrices need not be PSD; a perfectly reasonable set of similarity measurements can produce a matrix that fails the PSD condition, which would complicate the use of VS. Even symmetric similarity matrices—the most common and most natural type, satisfying $z_{ij} = z_{ji}$, so that similarity is reciprocal—are not guaranteed to be PSD, even when their entries obey the triangle inequality. The prevalence of such non-PSD cases in practice is unknown; we note that all similarity matrices examined for our datasets were PSD (a consequence of how we chose our similarity measures; this follows from Theorem 2.11 in [35]). Thus, PSD-ness is irrelevant for LCR but must be checked to apply VS.

And third, LCR in its various forms [22, 23] (Table 1) is part of an extensively characterized framework [10] that provides similarity-sensitive analogs of mutual information, relative entropy, and cross-entropy and offers methods for partitioning the overall similarity-sensitive entropy (γ) into within-class (α) and between-class (β) components. In entropic terms, α is the (effective-number form of the) joint entropy of the elements together with their class label; β is the mutual information between the elements and the class label; and γ is the marginal entropy of the elements irrespective of class. In this work we computed LCR in its γ form; when only a single class is present, γ coincides with α. Measuring the full α, β, and γ decomposition across ML classes is a promising direction for future research. (The `greylock` implementation computes α, β, and γ.)

The necessary mathematics for extending this decomposition to VS exists, because VS is based on von Neumann (quantum) entropy, for which quantum mutual information, quantum relative entropy, and related quantum quantities are well understood [36]. However, computing these quantities requires a matrix that simultaneously encodes each element and its class membership. If there are n total elements and k classes, the VS decomposition demands either an $nk \times nk$ matrix or an $n \times k \times n \times k$ 4-dimensional tensor (which can be reduced to $n \times n$ and $k \times k$ matrices by taking a partial trace). In contrast, LCR requires only the similarity matrix Z. Consequently, VS needs an explicit notion of similarity both between classes and between elements, whereas in LCR any class-level similarity arises purely from element-level similarities within the classes. In principle, the VS formulation would permit the confusing situation in which two classes contain exactly the same individuals but nonetheless have zero similarity—a scenario that cannot occur under LCR, in which the similarity between classes is purely a function of similarities among the elements they contain.

In sum, LCR and VS provide complementary descriptions of similarity-sensitive entropy, a unifying concept that is likely to find applications in many domains in which traditional entropy is currently a mainstay. We anticipate that the deeper understanding of their relationship and properties through work

such as that presented here will aid such investigations.

7 Acknowledgements

This work was supported by the Gordon and Betty Moore Foundation and by the NIH under grants R01HL150394, R01HL150394-SI, R01AI148747, and R01AI148747-SI.

A Appendix

A.1 Python code for finding counterexamples to $VS \geq LCR$

```
import numpy as np
import pandas as pd
from scipy.optimize import minimize
from tqdm import tqdm, trange
from multiprocessing import Pool

n_workers = 96

def uniform_div(Z,q):
    p = (1.0/Z.shape[0])*np.ones_like(Z[0])
    Zp = Z@p
    if q==1:
        h1 = -np.sum(p*np.log(Zp))
        return np.exp(h1)
    else:
        X = np.sum(p*(Zp**(q-1)))
        h = (1/(1-q)) * np.log(X)
        return np.exp(h)

def vendi_score(Z,q):
    evals = np.linalg.eigvalsh(Z/Z.shape[0])
    evals = np.array([e for e in evals if e>0]) # keep only the support
    if q==1:
        h1 = -np.sum(evals*np.log(evals))
        return np.exp(h1)
    else:
        X = np.sum(evals**q)
        h = (1/(1-q)) * np.log(X)
        return np.exp(h)

def func(x, n, r, q):
    Z = x_to_Z(x, n, r)
    return vendi_score(Z, q) - uniform_div(Z, q)

def x_to_Z(x, n, r):
    # map from reals to positive reals, reshape
    w = (x**2).reshape(r, n)
    # normalize rows
    w = w/np.sqrt((w**2).sum(0))
    Z = w.transpose()@w
    return Z

def find_min(n, r, q, scale):
    args = (n, r, q)
    x0 = scale*np.random.normal(0, 1, n*r)
```

```
    min_obj = minimize(func, x0, args)
    ret = dict(min_obj)
    ret.update({'x0': x0, 'Z0': x_to_Z(x0, n, r), 'Z': x_to_Z(min_obj.x, n, r)})
    return ret

qlist = [-10, -2.7, -2, -1, -0.3, 0, 0.3, 1, 2.7, 3.4, 4, 5, 2*np.pi, 7, 8, 9.1, 10]
sizes = [3, 5, 10, 20]
scales = [0.1, 1, 2, 10]
replicates = [i for i in range(1,6)]
combos = [{'q': q, 'n': n, 'r': r, 'rep': rep, 'scale': s} \
for q in qlist for n in sizes for s in scales \
for rep in replicates \
for r in set([x for x in [2, n//2, n] if x>1])]


def map_func(combo):
    q = combo['q']
    r = combo['r']
    n = combo['n']
    rep = combo['rep']
    scale = combo['scale']
    row = combo
    try:
        row.update(find_min(n, r, q, scale))
        return row
    except np.linalg.LinAlgError:
        return row

results = []
np.random.shuffle(combos)

n_batches = len(combos)//n_workers
if n_batches*n_workers< len(combos):
    n_batches +=1


pool = Pool(n_workers)

for batch in trange(n_batches):
    ix_start = batch*n_workers
    ix_end = ix_start + n_workers
    rows = pool.map(map_func, combos[ix_start:ix_end])
    results = results + rows
    results_df = pd.DataFrame(results)
    results_df.to_pickle('../data/vendi_inequality_counterexample_search.pkl')
```

A.2 Proof that $VS \geq LCR$ for $q \in \{2, 3, \infty\}$

Theorem 1. *Let Z, $VS_q(Z)$, and $D_q(Z, \frac{1}{n})$ be as in conjecture 1. Then when $q \in \{2, 3, \infty\}$,*

$$VS_q(Z) \geq D_q(Z, \frac{1}{n}) \tag{13}$$

Proof. We will consider the cases separately. For convenience, in what follows, denote by λ_i the eigenvalues of Z and r_i the row sums of Z, i.e. $r_i = \sum_j Z_{ij}$.

At $q = 2$,

$$\mathrm{VS}_2(Z) \quad = \quad e^{-\log \mathrm{tr}\left(\frac{Z}{n}\right)^2} = \frac{1}{\mathrm{tr}\left(\frac{Z}{n}\right)^2} = \frac{n^2}{\sum_{i,j} Z_{i,j}^2} \tag{14}$$

However, since $0 \leq Z_{ij} \leq 1 \forall (i,j)$, it follows that $Z_{i,j}^2 \leq Z_{ij}$ and so $\sum_{i,j} Z_{i,j}^2 \leq \sum_{i,j} Z_{ij}$. Therefore,

$$\frac{n^2}{\sum_{i,j} Z_{i,j}^2} \geq \frac{n^2}{\sum_{i,j} Z_{i,j}} \tag{15}$$

However, $\frac{n^2}{\sum_{i,j} Z_{i,j}}$ is simply $D_2(Z, \frac{1}{n})$, proving the $q = 2$ case.

For $q = 3$, we have that

$$\mathrm{tr}\left(Z^3\right) = \sum_{i,j} \left(Z^2\right)_{ij} Z_{ij} \leq \sum_{i,j} \left(Z^2\right)_{ij} \tag{16}$$

$$= \sum_{i,j,k} Z_{ik} Z_{kj} = \sum_k \left(\sum_i Z_{ki} \right) \left(\sum_j Z_{kj} \right) = \sum_i r_i^2. \tag{17}$$

This gives

$$\mathrm{VS}_3(Z) = e^{-\frac{1}{2}\log \mathrm{tr}\left(\frac{Z}{n}\right)^3} = \sqrt{\frac{n^3}{\mathrm{tr}\left(Z^3\right)}} \geq \sqrt{\frac{n^3}{\sum_i r_i^2}} = D_3(Z, \frac{1}{n}) \tag{18}$$

proving the $q = 3$ case.

For $q = \infty$, we invoke the Gershgorin circle inequality. We first note that

$$\mathrm{VS}_\infty(Z) \quad = \quad \frac{n}{\lambda_{\max}} \tag{19}$$

$$D_\infty(Z, \frac{1}{n}) \quad = \quad \frac{n}{r_{\max}} \tag{20}$$

where $\lambda_{\max}$ is the largest eigenvalue of Z, i.e. $\lambda_{\max} = \max_i \lambda_i$, and likewise $r_{\max} = \max_i r_i$. However, by the Gershgorin circle inequality, every eigenvalue of Z must fall within a Gershgorin interval, i.e. in one of the intervals $[Z_{ii} - \sum_{j \neq i} Z_{i,j}, Z_{ii} + \sum_{j \neq i} Z_{i,j}]$. As such, noting that $Z_{ii} + \sum_{j \neq i} Z_{i,j} = \sum_j Z_{i,j} = r_i$, we must have that every eigenvalue λ_μ of Z is bounded by $\lambda_\mu \leq r_{\max}$. But this means that

$$\mathrm{VS}_\infty(Z) = \frac{n}{\lambda_{\max}} \geq \frac{n}{r_{\max}} = D_\infty(Z, \frac{1}{n}) \tag{21}$$

□

A.3 Approach to a proof of $VS \geq LCR$ for all q

A proof for conjecture 1 at all q except for 1 (which could perhaps be filled in using continuity considerations) would follow from the following:

Conjecture 2. *Let Z be as in conjecture 1, and denote by $\mathbf{1}$ the n dimensional vector each of whose entries are 1. Then for $q > 1$,*

$$\mathrm{tr}\left(Z^q\right) \leq \sum_i \left(\sum_j Z_{ij} \right)^{q-1} = \mathbf{1}^\intercal \left(Z\mathbf{1}\right)^{q-1}. \tag{22}$$

Likewise, for $q < 1$,

$$\mathrm{tr}\left(Z^q\right) \geq \mathbf{1}^\intercal \left(Z\mathbf{1}\right)^{q-1}. \tag{23}$$

At $q > 1$, conjecture 2 would in turn be proven by the following conjecture:

Conjecture 3. *Let M be any $n \times n$ symmetric matrix whose entries obey $0 \leq M_{ij} \leq 1$, and denote by $\mathbf{1}$ the n dimensional vector each of whose entries are 1. Then for any $\alpha \geq 0$*

$$\mathbf{1}^\intercal M^\alpha \mathbf{1} \leq \mathbf{1}^\intercal \left(M\mathbf{1}\right)^\alpha \tag{24}$$

A.4 Other bounds

Beyond the conjectures upper bound of uniform diversity-with-similarity by Vendi score, we are able to prove the following bound in the opposite direction:

Theorem 2. *Let Z, $VS_q(Z)$, and $D_q(Z, \frac{1}{n})$ be as in conjecture 1. Then*

$$\mathrm{VS}_1(Z) \leq \frac{1}{n}\left(\sum_{i,j} Z_{ij}\right) D_1(Z, \frac{1}{n}) \tag{25}$$

Proof. We begin by constructing the diagonal matrix W according to

$$W_{ii} = \frac{1}{T}\sum_j Z_{ij} \tag{26}$$

$$T = \sum_{i,j} Z_{ij} \tag{27}$$

Notice that W is positive definite (since the entries of Z are non-negative and the diagonal entries are 1) and $\operatorname{tr} W = 1$. Treating W and Z/n as quantum density matrices, the quantum relative entropy between them should be positive, i.e.

$$\operatorname{tr}\left[\frac{Z}{n}\left(\log\frac{Z}{n} - \log W\right)\right] \geq 0 \tag{28}$$

$$\Rightarrow$$

$$\log \mathrm{VS}_{q=1}(Z) = -\operatorname{tr}\frac{Z}{n}\log\frac{Z}{n} \leq -\operatorname{tr}\frac{Z}{n}\log W. \tag{29}$$

However, since W is diagonal we have that

$$-\operatorname{tr}\frac{Z}{n}\log W = -\sum_i \frac{Z_{ii}}{n}\log W_{ii} \tag{30}$$

$$= -\sum_i \frac{1}{n}\log\frac{\sum_j Z_{ij}}{T} \tag{31}$$

$$= -\sum_i \frac{1}{n}\left(\log\frac{\sum_j Z_{ij}}{n} - \log\frac{T}{n}\right) \tag{32}$$

$$= \left(-\sum_i \frac{1}{n}\log\frac{\sum_j Z_{ij}}{n}\right) + \log\frac{T}{n} \tag{33}$$

As such, we have that

$$\log \mathrm{VS}_1(Z) \leq \left(-\sum_i \frac{1}{n}\log\frac{\sum_j Z_{ij}}{n}\right) + \log\frac{T}{n} \tag{34}$$

Exponentiating both sides, and recognizing that $e^{-\sum_i \frac{1}{n}\log\frac{\sum_j Z_{ij}}{n}} = D_1(Z, \frac{1}{n})$, we have that

$$\mathrm{VS}_1(Z) \leq \frac{T}{n} D_1(Z, \frac{1}{n}) \tag{35}$$

and so the theorem is proved.

□

References

[1] Tom Leinster and Mark W Meckes. Maximizing diversity in biology and beyond. Entropy, 18(3):88, 2016.

[2] Claude E. Shannon. Prediction and entropy of printed english. Bell System Technical Journal, 1951.

[3] M. A. Montemurro and D. H. Zanette. Entropy of natural languages: Theory and experiment. Scientific Programming, 1994.

[4] Michael Tawil and Anthony M. DiGiorgio. Competition in California's Medi-Cal Managed Care Market Assessed by Herfindahl-Hirschman Index. Inquiry: A Journal of Medical Care Organization, Provision and Financing, 59:469580221127063, 2022.

[5] Claude E Shannon. A mathematical theory of communication. The Bell system technical journal, 27(3):379–423, 1948.

[6] Alfréd Rényi. On Measures of Entropy and Information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, volume 4.1, pages 547–562. University of California Press, January 1961.

[7] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. Journal of statistical physics, 52(1):479–487, 1988.

[8] Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. Ecology, 93(3):477–489, 2012.

[9] Richard Reeve, Tom Leinster, Christina A Cobbold, Jill Thompson, Neil Brummitt, Sonia N Mitchell, and Louise Matthews. How to partition diversity. arXiv preprint arXiv:1404.6520, 2014.

[10] Tom Leinster. Entropy and Diversity: The Axiomatic Approach. arXiv preprint arXiv:2012.02113, 2020.

[11] Phuc Nguyen, Rohit Arora, Elliot D Hill, Jasper Braun, Alexandra Morgan, Liza M Quintana, Gabrielle Mazzoni, Ghee Rye Lee, Rima Arnaout, and Ramy Arnaout. greylock: A python package for measuring the composition of complex datasets. ArXiv, pages arXiv–2401, 2023.

[12] Josiah Couch, Rima Arnaout, and Ramy Arnaout. Beyond size and class balance: Alpha as a new dataset quality metric for deep learning. ArXiv, pages arXiv–2407, 2024.

[13] Josiah Couch, Miao Li, Rima Arnaout, and Ramy Arnaout. X-factor: Quality is a dataset-intrinsic property. arXiv preprint arXiv:2505.22813, 2025.

[14] Erin Chinn, Rohit Arora, Ramy Arnaout, and Rima Arnaout. ENRICHing medical imaging training sets enables more efficient machine learning. Journal of the American Medical Informatics Association: JAMIA, 30(6):1079–1090, May 2023.

[15] Ramy A. Arnaout, Eline T. Luning Prak, Nicholas Schwab, Florian Rubelt, and the Adaptive Immune Receptor Repertoire Community. The Future of Blood Testing Is the Immunome. Frontiers in Immunology, 12:626793, March 2021.

[16] Rohit Arora and Ramy Arnaout. Repertoire-scale measures of antigen binding. Proceedings of the National Academy of Sciences, 119(34):e2203505119, August 2022.

[17] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410, 2022.

[18] Amey P Pasarkar and Adji Bousso Dieng. The vendiscope: An algorithmic microscope for data collections. arXiv preprint arXiv:2502.10828, 2025.

[19] Amey P Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based

diversity metrics for science and machine learning. arXiv preprint arXiv:2310.12952, 2023.

[20] Lou Jost. Entropy and diversity. Oikos, 113(2):363–375, 2006.

[21] Mark O Hill. Diversity and evenness: a unifying notation and its consequences. Ecology, 54(2):427–432, 1973.

[22] Anne Chao, Chun-Huo Chiu, and Lou Jost. Phylogenetic diversity measures based on hill numbers. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1558):3599–3609, 2010.

[23] Anne Chao, Chun-Huo Chiu, Sébastien Villéger, I Fang Sun, Simon Thorn, Yiching Lin, Jyh-Min Chiang, and William B. Sherwin. An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis)similarity measures. Ecological Monographs, 2019.

[24] Jiancheng Yang, Rui Shi, and Bingbing Ni. MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 191–195, April 2021. arXiv:2010.14925 [cs].

[25] visipedia. inat_comp/2017/README.md at master · visipedia/inat_comp.

[26] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, and others. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence, 4(5):e210315, 2022. Publisher: Radiological Society of North America.

[27] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3462–3471, July 2017. ISSN: 1063-6919.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. arXiv:1405.0312 [cs].

[29] Yifan Wu, Hayden Gunraj, Chi-en Amy Tai, and Alexander Wong. COVIDx CXR-4: An Expanded Multi-Institutional Open-Source Benchmark Dataset for Chest X-ray Image-Based Computer-Aided COVID-19 Diagnostics, November 2023. arXiv:2311.17677 [eess].

[30] Kenneth Ge, Phuc Nguyen, and Ramy Arnaout. $\textit{lucie}$: An Improved Python Package for Loading Datasets from the UCI Machine Learning Repository, October 2024. arXiv:2410.09119 [cs].

[31] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. arXiv:1802.03426 [cs, stat].

[32] Dan Friedman. vendi-score: A diversity metric for machine learning (python package). `https://pypi.org/project/vendi-score/`, 2022. Accessed: 2025-10-13.

[33] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11):205, March 2017.

[34] Josiah Couch, Phuc Nguyen, Sarah Racz, Georgios Stratis, and Yuxuan Zhang. Possibility of entanglement of purification to be less than half of the reflected entropy. Phys. Rev. A, 109(2):022426, 2024.

[35] Tom Leinster. The magnitude of metric spaces. Documenta Mathematica, 18:857–905, 2013.

[36] M.A. Nielsen and I.L. Chuang. Quantum Computation and Quantum Information: 10th Anniversary Edition. Cambridge University Press, 2010.