

Bliss of dimensionality: How an unsupervised criterion identifies optimal low-resolution representations of high-dimensional datasets

Margherita Mele^{1,2}, Daniel Campos³, and Raffaello Potestio^{1,2,*}

¹Physics Department, *University of Trento*, via Sommarive 14, I-38123 Trento, Italy

²INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, I-38123 Trento, Italy

³Grup de Física Estadística, Departament de Física, Facultat de Ciències, *Universitat Autònoma de Barcelona*, 08193 Bellaterra, Barcelona, Spain



(Received 6 March 2026; accepted 29 September 2026; published 26 October 2026)

Selecting the optimal resolution for discretizing high-dimensional data is a central problem in physics and data analysis, particularly in unsupervised settings where the underlying distribution is unknown. The resolution-relevance (Res-Rel) framework addresses this issue through an information-theoretic trade-off between descriptive detail and statistical reliability. Here, we provide a systematic validation of this approach by comparing its characteristic optima—maximum relevance and the -1 slope point—with the discretization that minimizes the Kullback-Leibler (KL) divergence from a known or physically motivated ground truth distribution. Across unstructured and structured synthetic datasets, Gaussian clones of MNIST, and molecular dynamics simulations of the alanine dipeptide, we find that as the dimensionality or informative content increases the KL-optimal discretization consistently lies within the Res-Rel optimality region. Furthermore, in high-dimensional regimes the -1 slope criterion closely matches the KL divergence minimum. These results establish the quantitative consistency of unsupervised information-theoretic selection with distribution-based optimality.

DOI: [10.1103/3jcr-qy3d](https://doi.org/10.1103/3jcr-qy3d)

Introduction. Discrete representations of continuous data cover a wide range of situations, from building simple one-dimensional histograms to complex clustering techniques for high-dimensional datasets. Such methods are ubiquitous across physics [1], data science [2], and machine learning [3], and provide the bedrock for tasks such as coarse graining [4], density estimation [5], and model inference [6]. Here, as in many other related contexts, choosing the appropriate level of detail of the representation remains a fundamental open problem [7,8]: Indeed, overly coarse descriptions discard important structure, whereas excessively fine ones introduce sampling noise and statistical unreliability [9,10]. This trade-off is particularly acute in finite-sample regimes and high-dimensional systems, where empirical frequencies may poorly approximate the underlying generative process of the data [11–13]. Standard criteria developed to select an adequate level of detail of the data representation often rely on supervised information about the target distribution (e.g., likelihood optimization or divergence minimization) [14–16], but such an approach is naturally excluded in unsupervised settings. Consequently, intrinsically data-driven criteria are needed to identify informative representations without explicit knowledge of the true probability distribution [17–19]. The

resolution-relevance (Res-Rel) framework [20,21] addresses this problem by introducing an information-theoretic (IT) criterion that balances representational granularity (resolution) against statistical significance (relevance). In this approach, clusters of instances of a dataset are grouped together according to a given criterion, thereby inducing a spectrum of reduced representations of the data that are evaluated in terms of their resolution and relevance; by analyzing how empirical frequency statistics changes with the number of discrete states, the framework identifies optimal resolution regimes directly from the data, without external supervision [22]. The simplicity and the unsupervised character of the framework make it extremely attractive and potentially powerful, and its increasingly numerous applications provide concrete support to its effectiveness [23–29]; furthermore, an important body of work has been done to lay solid mathematical and statistical foundations to the approach [30–33]. However, a simple, bottom-up validation of its working and its range of validity is still missing. In this Letter, we address this problem by analyzing synthetic data points generated from distributions known *a priori*, as well as realistic datasets such as MNIST and molecular dynamics trajectories of the alanine dipeptide, systematically evaluating the capability of the resolution-relevance approach to provide physically informative low-resolution representations of high-dimensional data and setting boundaries for its applicability.

Methods. Consider an N -dimensional dataset of size P partitioned into n discrete states, or clusters, by a mapping that assigns each data point to a state. Let k_i denote the number of data points assigned to state i and m_k the number of states with occupancy exactly equal to k . Resolution and relevance

*Contact author: raffaello.potestio@unitn.it

are defined, respectively, as

$$H_{\text{res}} = - \sum_{i=1}^n \frac{k_i}{P} \ln \frac{k_i}{P}, \quad H_{\text{rel}} = - \sum_k \frac{km_k}{P} \ln \frac{km_k}{P}. \quad (1)$$

Resolution, H_{res} , is the Shannon entropy of the empirical distribution of frequencies and quantifies the level of detail of the representation. Relevance, H_{rel} , instead captures the heterogeneity of empirical frequencies through the occupancy distribution m_k , and so reflects the amount of statistically significant information entailed in the coarse representation of the data [22,23]. By varying the number of states, one can generate the *resolution-relevance* curve, which shows and quantifies the trade-off between descriptive detail and statistical reliability [30]. Resolution increases with the number of states, while relevance displays a non-monotonic behavior: It first increases as more and more informative structure is resolved, and then decreases when further refinement produces poorly sampled states dominated by noise. An optimal representation thus lies in the intermediate regime, between the maximum relevance and the point where the curve attains slope -1 , which marks the information-theoretical optimum [21,33]. We consider this interval comprised between the relevance maximum and the IT optimum to represent the *optimality region*: Beyond this point, information gains from increasing resolution are lower than the corresponding information losses by reducing statistical significance.

Each representation induces an empirical probability distribution: Given a partition $\{C_j\}_{j=1}^n$ with occupancies $k_j = |C_j|$, the empirical probability assigned to each data point is $\hat{p}(x_i) = k_j/P$, for $x_i \in C_j$. If the true generative distribution $p(x)$ were known, a representation at the optimal level of resolution would minimize the discrepancy between $p(x)$ and the empirical distribution $\hat{p}(x)$. A natural measure for this is the Kullback-Leibler (KL) divergence between both distributions,

$$D_{\text{KL}}(p \parallel \hat{p}) = \sum_i p(x_i) \ln \frac{p(x_i)}{\hat{p}(x_i)}. \quad (2)$$

When evaluating the KL divergence, the reference probability density is evaluated at the data points and normalized over the dataset; the empirical representation is treated consistently, so that both induced discrete distributions are properly normalized, such that $\sum_i p(x_i) = \sum_i \hat{p}(x_i) = 1$. In typical unsupervised settings, however, the true distribution $p(x)$ is unknown and cannot be used to guide the choice of representation. The problem therefore reduces to identifying informative discretizations directly from empirical frequency statistics, and the resolution-relevance framework provides a fully data-driven criterion to select such representations. To assess how the representations selected by this criterion relate to optimal ones reproducing the empirical distribution, we analyze datasets for which such empirical is available—either known by construction (synthetic data) or physically/statistically motivated (real systems). This controlled setting enables a direct comparison between the Res-Rel optima and the discretization that minimizes the KL divergence from the reference. The datasets analyzed and the associated analysis procedure are summarized below, with full technical details reported in the Supplemental Material [34].

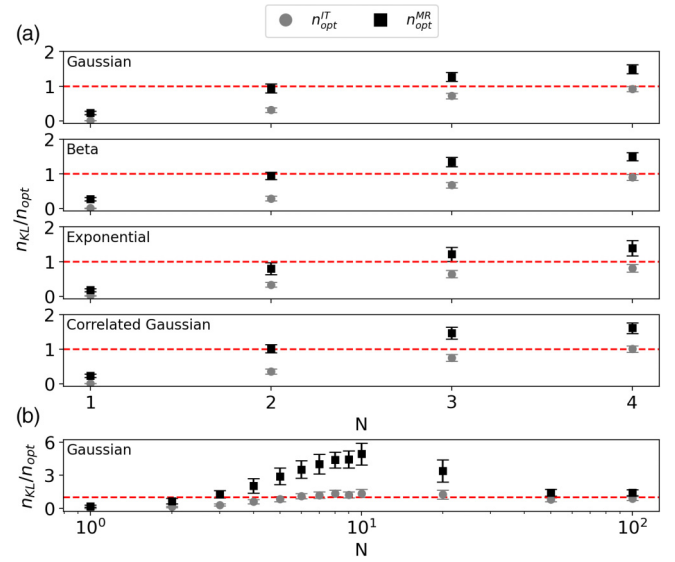


FIG. 1. Comparison between resolution-relevance and Kullback-Leibler optimal representations for unstructured synthetic data. Ratio $n_{\text{KL}}/n_{\text{opt}}$ as a function of the data dimensionality N for unstructured synthetic datasets. In each panel, black squares correspond to $n_{\text{opt}}^{\text{MR}}$ (maximum relevance), while gray circles correspond to $n_{\text{opt}}^{\text{IT}}$ (-1 slope point). The red dashed line indicates equality between the two estimates. Error bars represent the standard deviation over 50 independent realizations. Panel (a) shows low-dimensional datasets analyzed using N -dimensional histograms, generated from Gaussian, beta, exponential, and correlated Gaussian distributions. Panel (b) shows the corresponding analysis for i.i.d. Gaussian data spanning from low to high dimensions, where representations are constructed using UPGMA clustering.

Methodology. For each dataset realization, representations are constructed by partitioning the data into n discrete states. For a first test on low-dimensional ($N \leq 4$) datasets, we discretized the data space of all four distributions under examination using N -dimensional histograms. The direct discretization of spaces with $N > 4$ rapidly becomes numerically intractable (see below); hence, we focused on the case of Gaussian distributions in spaces of dimension 1–100, partitioned using agglomerative hierarchical clustering with unweighted pair group method with arithmetic mean (UPGMA) linkage [35]. The number of states n parametrizes the representation and generates a unique resolution-relevance curve for each realization. For every n , we compute $(H_{\text{res}}, H_{\text{rel}})$ and the empirical distribution induced by the partition. From each resolution-relevance curve, we identify two characteristic values of n : the point of maximum relevance (MR), defining $n_{\text{opt}}^{\text{MR}}$, and the point where the curve attains slope -1 , defining $n_{\text{opt}}^{\text{IT}}$. Independently, the KL divergence between the reference distribution and the empirical distribution is evaluated as a function of n , and the minimizer n_{KL} is identified. For synthetic and semi-real datasets, results are averaged over 50 independent realizations; for the alanine dipeptide [36] the analysis is performed over ten independent molecular dynamics (MD) trajectories. Implementation details and additional robustness analyses are reported in the Supplemental Material [34].

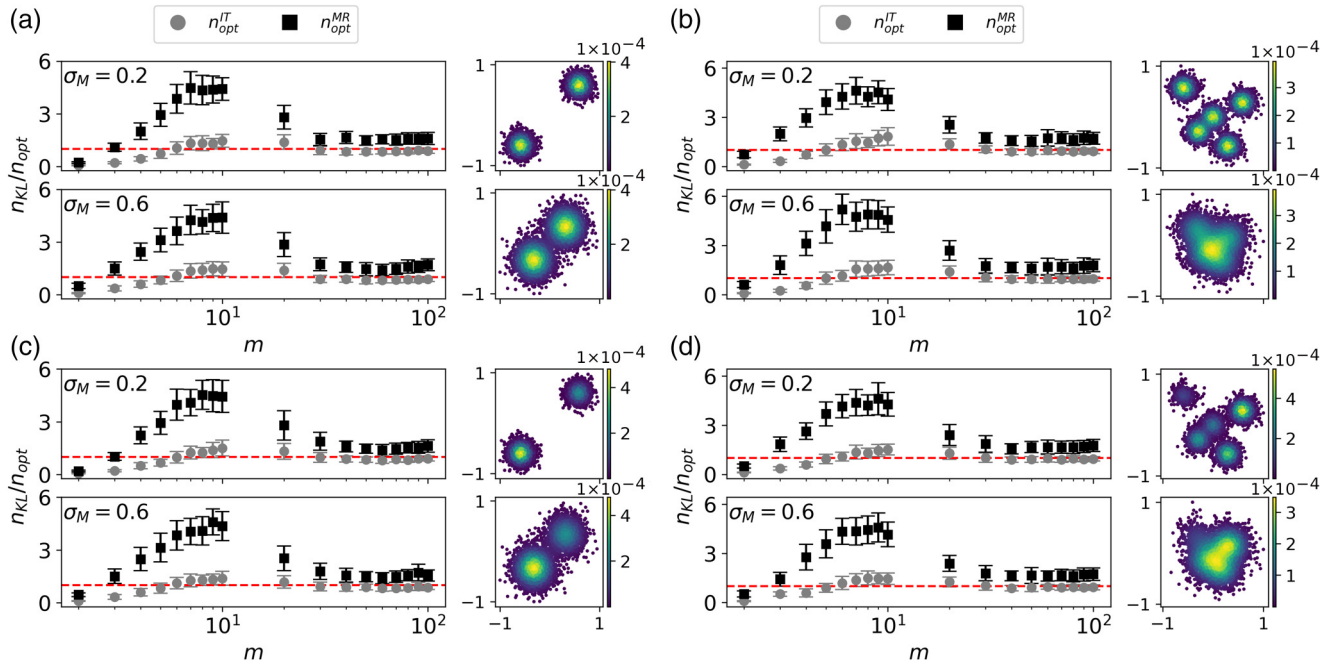


FIG. 2. Comparison between resolution-relevance and Kullback-Leibler optimal representations for structured synthetic data. Ratio $n_{\text{KL}}/n_{\text{opt}}$ as a function of the number of informative dimensions m for structured synthetic datasets with latent Gaussian mixture structure. In each panel, black squares correspond to $n_{\text{MR}}^{\text{opt}}$ (maximum relevance), while gray circles correspond to $n_{\text{IT}}^{\text{opt}}$ (-1 slope point). The red dashed line indicates equality between the two estimates. Error bars represent the standard deviation over 50 independent realizations. Panels (a) and (b) show equal-weight mixtures for $K = 2$ and $K = 5$, while panels (c) and (d) show the corresponding unequal-weight cases with weights $[0.66, 0.33]$ and $[0.34, 0.27, 0.19, 0.13, 0.07]$, respectively. Different rows correspond to increasing values of the mixture standard deviation σ_M , as indicated. For each row, the right-hand subpanels show two-dimensional projections of representative datasets generated with $m = 2$ for the corresponding parameter values. For each value of K , the random seed used to generate the data is fixed across different scenarios, ensuring that the relative positions of the mixture means are identical.

Results.

a. Unstructured synthetic data. To establish a baseline, we first consider synthetic datasets without latent discrete structure, where all components are sampled independently from a prescribed continuous distribution; this setting allows us to isolate purely dimensional effects. Datasets are generated from Gaussian, beta, exponential, and correlated Gaussian distributions (see the Supplemental Material [34] for technical details). For low-dimensional data ($N \leq 4$), representations are constructed via N -dimensional histograms, which provide a direct and controlled discretization of the data space and constitute the natural benchmark for estimating empirical probability densities. However, the number of bins grows exponentially with N , leading to severe sparsity and unstable frequency estimates at fixed sample size [11]. To probe higher-dimensional regimes, we therefore adopt agglomerative clustering (UPGMA), which defines discrete states directly from interpoint distances without relying on a regular grid. For each dimensionality N , we compare $n_{\text{MR}}^{\text{opt}}$ and $n_{\text{IT}}^{\text{opt}}$ with the KL-optimal value n_{KL} . Figure 1(a) shows that in one dimension the Res-Rel framework overestimates the KL-optimal number of states, with both criteria yielding $n_{\text{opt}} > n_{\text{KL}}$. The discrepancy rapidly decreases with increasing N : Already for $N = 2$, $n_{\text{opt}}^{\text{MR}}$ is close to n_{KL} , and for $N \geq 2$ the KL-optimal value systematically falls within the Res-Rel optimality region defined by $[n_{\text{opt}}^{\text{MR}}, n_{\text{opt}}^{\text{IT}}]$. This behavior is robust across all tested distributions. The analysis is extended to higher dimensions using i.i.d. Gaussian data

[Fig. 1(b)]. Also in this case, the low-dimensional overestimation progressively vanishes as N increases. For intermediate dimensions, the Res-Rel optimality region broadens, while for $N > 10$ the two criteria converge, yielding close values of n_{opt} . Overall, the discrepancy observed at low dimensionality progressively reduces as N increases, and the KL-optimal value falls within the Res-Rel optimality region for $N \geq 2$.

b. Structured synthetic data. Having characterized the dependence of the sheer dimensionality of the data, we now introduce latent discrete structure embedded in a high-dimensional space. The total number of dimensions is fixed to $N = 100$, of which only $m \leq N$ components carry informative signal. In the informative subspace, samples are drawn from mixtures of K Gaussian components with standard deviation σ_M , whereas the remaining dimensions consist of independent Gaussian noise with $\sigma \ll \sigma_M$ (see the Supplemental Material [34] for full construction details). Both binary ($K = 2$) and multicomponent ($K = 5$) mixtures, with equal and unequal weights, are analyzed. Figure 2 reports the ratio between the KL-optimal number of discrete elements n_{KL} and the Res-Rel optima as a function of the number of informative dimensions m . For very low informative dimensionality ($m = 2$), the Res-Rel criteria select values exceeding n_{KL} , consistent with the behavior observed in low-dimensional unstructured data. As m increases, the agreement improves: The KL-optimal value systematically falls within the Res-Rel optimality region and is typically located close to the -1 slope point (gray circles). The width of the optimality region initially increases with m

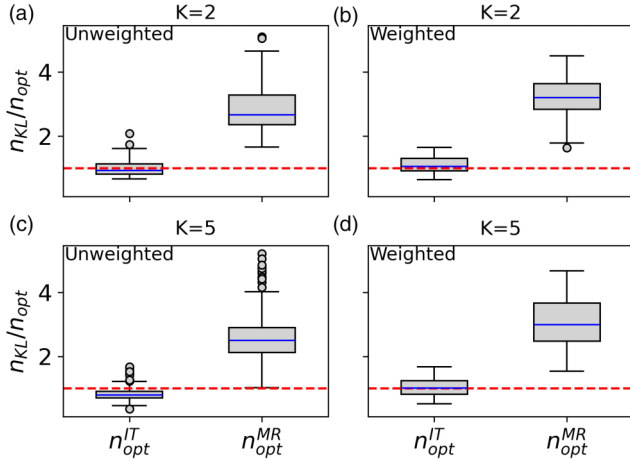


FIG. 3. Comparison between resolution-relevance and Kullback-Leibler optimal representations for Gaussian clones of MNIST. Boxplots show the ratio $n_{\text{KL}}/n_{\text{opt}}$ evaluated at two characteristic points of the resolution-relevance curve: the -1 slope point $n_{\text{opt}}^{\text{IT}}$ (information-theoretic optimum) and the maximum-relevance point $n_{\text{opt}}^{\text{MR}}$. The horizontal red dashed line marks the ideal value $n_{\text{KL}}/n_{\text{opt}} = 1$, corresponding to perfect agreement between the two criteria, while blue horizontal lines indicate the median of each distribution. Panels (a) and (b) correspond to mixtures with $K = 2$ components, with equal weights in panel (a) and unequal weights [0.66, 0.33] in panel (b). Panels (c) and (d) show mixtures with $K = 5$ components, with equal weights in panel (c) and unequal weights [0.34, 0.27, 0.19, 0.13, 0.07] in panel (d).

and subsequently decreases for $m \geq 10$, as the informative signal dominates the noisy background. These trends are robust across all investigated mixture configurations and signal widths (additional parameter combinations are reported in the Supplemental Material [34]).

c. Semi-real data: Gaussian clones of MNIST. To bridge fully synthetic models and real-world data, we analyze Gaussian clones derived from the MNIST handwritten digit database [37]. For each selected digit class, a class-conditional multivariate Gaussian distribution is estimated from the original data and used to generate synthetic samples (“Gaussian clones”). The reference distribution is thus a Gaussian mixture with K components corresponding to distinct digit classes (see the Supplemental Material [34] for estimation details). Digit 1 is excluded due to its anomalous Res-Rel behavior (see the Supplemental Material [34]). We analyze mixtures with $K = 2$ and $K = 5$, considering both equal and unequal class weights. For each configuration, all possible combinations of K digit classes are examined. Figure 3 reports the ratio $n_{\text{KL}}/n_{\text{opt}}$ evaluated at the two Res-Rel characteristic points. Across all scenarios, the -1 slope criterion yields values closely aligned with the KL-minimizing discretization, with distributions centered around $n_{\text{KL}}/n_{\text{opt}}^{\text{IT}} \simeq 1$. In contrast, the maximum-relevance criterion systematically selects smaller numbers of discrete elements, resulting in $n_{\text{KL}}/n_{\text{opt}}^{\text{MR}} > 1$, albeit with deviations remaining within a factor smaller than 4. The same qualitative behavior is observed for both equal- and unequal-weight mixtures and across different class combinations, and is consistent with the features of the unstructured and structured synthetic data discussed

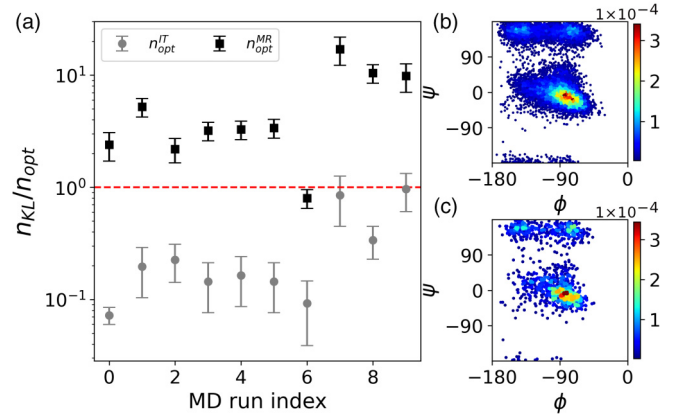


FIG. 4. Comparison between resolution-relevance and Kullback-Leibler optimal representations for alanine dipeptide. (a) Ratio $n_{\text{KL}}/n_{\text{opt}}$ computed over ten independent MD simulations. Gray circles denote the ratio evaluated at the maximum-relevance point $n_{\text{opt}}^{\text{MR}}$, while black squares correspond to the information-theoretic optimum $n_{\text{opt}}^{\text{IT}}$, defined as the -1 slope point of the resolution-relevance curve. The horizontal red dashed line marks the reference value $n_{\text{KL}}/n_{\text{opt}} = 1$, corresponding to perfect agreement between the Kullback-Leibler-optimal discretization and the resolution-relevance prediction. (b), (c) Two-dimensional projections in the space of backbone dihedral angles (ϕ , ψ) for a representative trajectory. In panel (b), colors represent a two-dimensional histogram of the raw MD frames, providing an empirical estimate of the reference probability distribution. In panel (c), only the centroids of the clusters obtained by RMSD-based clustering at $n_{\text{opt}}^{\text{IT}}$ are shown. Colors are assigned from a two-dimensional histogram in (ϕ , ψ) constructed over these centroids, each weighted by the multiplicity of its corresponding cluster.

in the previous sections, particularly in those cases where the effective dimensionality of the data is in the 4–10 range. Interestingly, this result is line with previous works [38–40] that estimated the effective dimension of the MNIST dataset in the range 12–14: Our data thus suggest that the pairs or 5-tuples of clone digits are embedded in a space of effective dimension $O(10)$.

d. Real data: alanine dipeptide. We finally consider a dataset consisting of MD simulations of the alanine dipeptide, a standard benchmark system whose equilibrium behavior is well described in terms of the backbone dihedral angles (ϕ , ψ) [36,41]. As commonly done, the reference probability distribution is defined in this reduced space and estimated from each trajectory via a two-dimensional histogram of the dihedral angles (see the Supplemental Material [34] for simulation details). Discrete representations are instead constructed by clustering molecular configurations in the full configurational space using the atomwise root mean square distance (RMSD) as a metric. This setup allows us to test whether the Res-Rel-selected discretizations recover the physically relevant probability landscape defined in dihedral angle space. Figure 4(a) reports, for ten independent trajectories, the ratio between the KL-optimal number of clusters n_{KL} and the Res-Rel optima. In all cases, n_{KL} lies within the optimality region. However, unlike the synthetic and semi-real datasets, no single Res-Rel criterion systematically coincides with the KL

minimum across trajectories. Figures 4(b) and 4(c) compare, for a representative trajectory, the empirical dihedral angle density and the probability distribution reconstructed from clustering at $n_{\text{opt}}^{\text{IT}}$. The two representations display consistent large-scale conformational features. Despite the absence of an exact generative distribution and the finite-sampling nature of the reference estimate, the Res-Rel framework restricts the optimal discretization to a narrow range of cluster numbers across independent trajectories, with a dispersion in line with what was observed in the previously discussed cases for a low number of informative features.

Discussion. In this work, we carried out a systematic investigation of low-resolution representations of synthetic, semi-real, and real datasets, comparing the optimal clusterings obtained through the application of the resolution-relevance framework against those that minimize the KL divergence from the underlying generative model. In low-dimensional regimes, the Res-Rel criteria tend to overestimate the KL minimum; as the effective dimensionality or informative content of the data increases, however, the KL-optimal number of discrete elements falls within the Res-Rel optimality region. In high-dimensional settings, the -1 slope criterion provides the closest agreement with the KL divergence minimum: Indeed, as already observed by Cubero *et al.* [32], this point strikes an optimal balance between reducing resolution and maximizing informativeness, and corresponds to the threshold between lossy and lossless compression; in contrast, the point of largest relevance is deep in the lossy compression regime, thereby leading to a potentially excessive coarsening. Crucially, also in a real molecular system, where the reference distribution is empirically estimated, the KL-based resolution level falls within the Res-Rel optimality region. However, even though the -1 slope point is closer to the reference than the relevance maximum, we observe an important dispersion among different simulation replicas; this shows that, in real case scenarios, it is advisable to consider the whole range of resolution in between the two optimal points $n_{\text{opt}}^{\text{IT}}$ and $n_{\text{opt}}^{\text{MR}}$. A central result of this work is that, for a given sample size, the larger the dimensionality of the embedding space of the data under examination, the easier it becomes for the Res-Rel analysis approach to identify the partition that most faithfully reproduces the key features of the underlying generative process.

This behavior is particularly relevant as the so-called *curse of dimensionality* [42] here becomes a “bliss”, reminiscent of the simplification ushered in by the thermodynamic limit in the context of statistical mechanics. In conclusion, the Res-Rel framework offers a principled, fully data-driven mechanism for selecting informative low-resolution representations of high-dimensional datasets without requiring explicit knowledge of the underlying generative distribution. Importantly, one of the most critical advantages of the approach, namely, that it relies only on the occurrence statistics, requires that one interprets the results in the light of the clustering framework employed [26,32]; this, however, does not represent a limitation, but rather an inherent necessity for context in the application of an unsupervised analysis approach. Taken together, this investigation shows that the Res-Rel criterion identifies discretizations that are statistically robust and probabilistically meaningful, effectively linking unsupervised information-theoretic selection to supervised distribution-based optimality.

Acknowledgments. R.P. acknowledges support from ICSC—Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing, funded by the European Union under NextGenerationEU. This work was conducted in the spirit of the Slow Science Manifesto, advocating for collaborative and sustainable research [43].

All authors conceived the study and proposed the method. M.M. produced the data and performed the analysis. R.P. and M.M. interpreted the results. All authors drafted the paper, reviewed the results, and approved the final version of the manuscript.

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

The authors declare no competing interests.

Data availability. The data that support the findings of this article are openly available [44].

- [1] L. P. Kadanoff, Scaling laws for Ising models near t_c , *Phys. Phys. Fiz.* **2**, 263 (1966).
- [2] A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering: A review, *ACM Comput. Surv.* **31**, 264 (1999).
- [3] V. N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
- [4] K. G. Wilson and J. Kogut, The renormalization group and the ϵ expansion, *Phys. Rep.* **12**, 75 (1974).
- [5] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Routledge, London, 2018).
- [6] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* **19**, 716 (1974).
- [7] J. Rissanen, Modeling by shortest data description, *Automatica* **14**, 465 (1978).
- [8] V. N. Vapnik and A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, in

Measures of Complexity: Festschrift for Alexey Chervonenkis, edited by V. Vovk, H. Papadopoulos, and A. Gammerman (Springer, Cham, 2015), pp. 11–30.

- [9] P. J. Bickel and E. Levina, Some theory for Fisher’s linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations, *Bernoulli* **10**, 989 (2004).
- [10] M. J. Wainwright, *High-Dimensional Statistics: A Non-asymptotic Viewpoint* (Cambridge University Press, Cambridge, 2019).
- [11] R. Bellman, Dynamic programming, *Science* **153**, 34 (1966).
- [12] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 2013).
- [13] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, *Proc. Natl. Acad. Sci. USA* **116**, 15849 (2019).

- [14] A. Gelman, J. Hwang, and A. Vehtari, Understanding predictive information criteria for Bayesian models, *Stat. Comput.* **24**, 997 (2014).
- [15] G. Schwarz, Estimating the dimension of a model, *Ann. Stat.* **6**, 461 (1978).
- [16] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, New York, 2004), 2nd ed.
- [17] N. Tishby and N. Zaslavsky, Deep learning and the information bottleneck principle, in *2015 IEEE Information Theory Workshop (ITW), Jerusalem* (IEEE, New York, 2015), pp. 1–5.
- [18] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, A survey on contrastive self-supervised learning, *Technologies* **9**, 2 (2021).
- [19] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).
- [20] M. Marsili, I. Mastromatteo, and Y. Roudi, On sampling and modeling complex systems, *J. Stat. Mech.* (2013) P09003.
- [21] A. Haimovici and M. Marsili, Criticality of mostly informative samples: A Bayesian model selection approach, *J. Stat. Mech.* (2015) P10013.
- [22] M. Marsili and Y. Roudi, Quantifying relevance in learning and inference, *Phys. Rep.* **963**, 1 (2022).
- [23] S. Grigolon, S. Franz, and M. Marsili, Identifying relevant positions in proteins by critical variable selection, *Mol. Biosyst.* **12**, 2147 (2016).
- [24] C. Battistini, B. Dunn, and Y. Roudi, Learning with unknowns: Analyzing biological data in the presence of hidden variables, *Curr. Opin. Syst. Biol.* **1**, 122 (2017).
- [25] R. Holtzman, M. Giuliani, and R. Potesio, Making sense of complex systems through resolution, relevance, and mapping entropy, *Phys. Rev. E* **106**, 044101 (2022).
- [26] M. Mele, R. Covino, and R. Potesio, Information-theoretical measures identify accurate low-resolution representations of protein configurational space, *Soft Matter* **18**, 7064 (2022).
- [27] J. Morand, S. Yip, Y. Velegrakis, G. Lattanzi, R. Potesio, and L. Tubiana, Quality assessment and community detection methods for anonymized mobility data in the Italian COVID context, *Sci. Rep.* **14**, 4636 (2024).
- [28] R. Aldrigo, R. Menichetti, and R. Potesio, Low-resolution descriptions of model neural activity reveal hidden features and underlying system properties, *Phys. Rev. E* **111**, 044315 (2025).
- [29] M. Mele, R. Fiorentini, T. Tarenzi, G. Mattiotti, and R. Potesio, Determining the optimal structural resolution of proteins through an information-theoretic analysis of their conformational ensemble, *J. Chem. Theory Comput.* **22**, 1244 (2026).
- [30] J. Song, M. Marsili, and J. Jo, Resolution and relevance trade-offs in deep learning, *J. Stat. Mech.* (2018) 123406.
- [31] R. J. Cubero, M. Marsili, and Y. Roudi, Minimum description length codes are critical, *Entropy* **20**, 755 (2018).
- [32] R. J. Cubero, J. Jo, M. Marsili, Y. Roudi, and J. Song, Statistical criticality arises in most informative representations, *J. Stat. Mech.* (2019) 063402.
- [33] O. Duranthon, M. Marsili, and R. Xie, Maximal relevance and optimal learning machines, *J. Stat. Mech.* (2021) 033409.
- [34] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/3jcr-qy3d> for methodological details, dataset specifications, numerical procedures, and additional analyses, including robustness checks, results for structured synthetic datasets, MNIST and alanine dipeptide, and KL-based analyses, which includes Ref. [45].
- [35] R. R. Sokal and C. D. Michener, A statistical method for evaluating systematic relationships, *Univ. Kans. Sci. Bull.* **38**, 1409 (1958).
- [36] P. E. Smith, The alanine dipeptide free energy surface in solution, *J. Chem. Phys.* **111**, 5568 (1999).
- [37] L. Deng, The MNIST database of handwritten digit images for machine learning research [best of the web], *IEEE Signal Process. Mag.* **29**, 141 (2012).
- [38] M. Hein and J.-Y. Audibert, Intrinsic dimensionality estimation of submanifolds in R_d , in *Proceedings of the 22nd International Conference on Machine Learning*, edited by S. Dzeroski (General chair), L. D. Raedt, and S. Wrobel (Program chairs) (2005), pp. 289–296.
- [39] E. Facco, M. d’Errico, A. Rodriguez, and A. Laio, Estimating the intrinsic dimension of datasets by a minimal neighborhood information, *Sci. Rep.* **7**, 12140 (2017).
- [40] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, The intrinsic dimension of images and its impact on learning, in *International Conference on Learning Representations*, (2021), <https://openreview.net/forum?id=XJk19XzGq2J>.
- [41] P. G. Bolhuis, C. Dellago, and D. Chandler, Reaction coordinates of biomolecular isomerization, *Proc. Natl. Acad. Sci. USA* **97**, 5877 (2000).
- [42] N. Altman and M. Krzywinski, The curse(s) of dimensionality, *Nat. Methods* **15**, 399 (2018).
- [43] Slow Science, slow-science.com.
- [44] R. Potesio and M. Mele, Project files provided as supporting information to the manuscript “Bliss of dimensionality: how an unsupervised criterion identifies optimal low-resolution representations of high-dimensional datasets” [Dataset], Zenodo, 2026, doi: [10.5281/zenodo.18860257](https://doi.org/10.5281/zenodo.18860257).
- [45] R. T. McGibbon, MD Trajectories of ALA2 (fileset), 2014, <https://doi.org/10.6084/m9.figshare.1026131>.