Dataset-Adaptive Dimensionality Reduction

Hyeon Jeon, Jeongin Park, Soohyun Lee, Dae Hyun Kim, Sungbok Shin, and Jinwook Seo

Abstract—Selecting the appropriate dimensionality reduction (DR) technique and determining its optimal hyperparameter settings that maximize the accuracy of the output projections typically involves extensive trial and error, often resulting in unnecessary computational overhead. To address this challenge, we propose a *dataset-adaptive* approach to DR optimization guided by *structural complexity metrics*. These metrics quantify the intrinsic complexity of a dataset, predicting whether higher-dimensional spaces are necessary to represent it accurately. Since complex datasets are often inaccurately represented in two-dimensional projections, leveraging these metrics enables us to predict the maximum achievable accuracy of DR techniques for a given dataset, eliminating redundant trials in optimizing DR. We introduce the design and theoretical foundations of these structural complexity metrics. We quantitatively verify that our metrics effectively approximate the ground truth complexity of datasets and confirm their suitability for guiding dataset-adaptive DR workflow. Finally, we empirically show that our dataset-adaptive workflow significantly enhances the efficiency of DR optimization without compromising accuracy.

Index Terms—Dimensionality reduction, Structural complexity, High-dimensional data, Optimization, Dataset-adaptive workflow

---

◆

---

1 INTRODUCTION

Dimensionality reduction (DR) is an effective tool for visualizing and analyzing high-dimensional (HD) data [8,9,23,38,55,61]. However, DR projection of HD data inherently results in distortions leading to inaccurate representations of the original structure of the HD data (e.g., local neighborhoods or clusters) [48,49]. As a result, data analysis based on DR may occasionally lead to inaccurate findings or cause analysts to overlook important structural characteristics [3,38]; thus, analysts should *optimize* DR projections to minimize distortions of key structural characteristics [37].

Optimizing the hyperparameters of DR techniques to minimize distortions is typically a computationally expensive process (Sect. 3). Different hyperparameter settings must be tested iteratively, but the optimal number of iterations is rarely clear. Practitioners often run more iterations than necessary, observing minimal or no reduction in distortion after a certain point. This is because, unlike typical machine learning optimization such as gradient descent, DR hyperparameter optimization does not have explicit convergence criteria. In visual analytics, a desirable DR hyperparameter optimization involves more than just minimizing a loss function—additional constraints, such as the preservation of neighborhood [11,74] and cluster structures [34,56], must also be considered. Representing these diverse criteria and DR techniques together within a single mathematical formulation is challenging, making the validation of convergence also challenging. Furthermore, identifying the optimal DR technique requires comparing multiple techniques, further complicating the optimization process.

We propose a *dataset-adaptive* workflow that improves the efficiency of DR optimization. Building upon previous findings [46,47] that certain patterns are more prominent in HD data, our approach quantifies the prominence of these patterns to estimate the difficulty of accurately projecting the data into lower-dimensional spaces. We introduce *structural complexity metrics* to measure these patterns, and use these scores to predict the maximum accuracy achievable by DR techniques. The metrics thus enhance the efficiency of DR optimization by (1) guiding the selection of an appropriate DR technique for a

given dataset and (2) enabling early termination of optimization once near-optimal hyperparameters have been reached, avoiding unnecessary computations.

While existing metrics, such as intrinsic dimensionality metrics (Sect. 2.3), can potentially serve as structural complexity metrics, they lack the desired characteristics necessary for effective integration into our dataset-adaptive workflow (Sect. 4.1). We thus introduce two novel structural complexity metrics—*Pairwise Distance Shift* (PDS) and *Mutual Neighbor Consistency* (MNC)—tailored to the dataset-adaptive workflow. PDS characterizes the complexity of a dataset's global structure by quantifying the shift in pairwise distances [45,47], a well-established indicator associated with the curse of dimensionality. MNC, in contrast, captures the complexity of a dataset's local structure by measuring the inconsistency between two neighborhood-based similarity functions: k-Nearest Neighbors (kNN) and Shared Nearest Neighbors (SNN) [13]. By jointly characterizing global and local structural complexity, these metrics effectively guide the optimization of DR techniques. We theoretically and empirically verify that our metrics capture prominently observable patterns in higher-dimensional space, ensuring reliable guidance when identifying optimal DR techniques and hyperparameter settings.

A series of experiments with real-world datasets confirm the effectiveness of our structural complexity metrics and the dataset-adaptive workflow. First, we verify that PDS, MNC, and their ensemble produce scores that highly correlate with ground truth structural complexity approximated by an ensemble of multiple state-of-the-art DR techniques, significantly outperforming baselines such as intrinsic dimensionality metrics (Sect. 2.3). Second, we verify our metrics' utility in supporting the dataset-adaptive workflow of finding optimal DR projections. Finally, we show that the dataset-adaptive workflow significantly reduces the computational time required for DR optimization without compromising projection accuracy.

2 BACKGROUND AND RELATED WORK

Our work is relevant to three areas of previous work: (1) DR and its distortions, (2) dataset-adaptive approaches in machine learning, and (3) intrinsic dimensionality metrics.

2.1 Dimensionality Reduction and Distortions

DR techniques abstract HD data in a 2D space while preserving important structural characteristics. Formally, for a given HD dataset $X = \{x_i \in \mathbb{R}^D, i = 1, 2, \ldots, N\}$, DR aims to produce $Y = \{x_i \in \mathbb{R}^2, i = 1, 2, \ldots, N\}$ that minimize the structural difference between X and Y. For example, while PCA [63] seeks a 2D projection that maximizes the explainability of the original variance in HD data, UMAP [59] aims to preserve neighborhood structures, revealing local manifolds. However,

- *Hyeon Jeon, Soohyun Lee, Jeongin Park, and Jinwook Seo (corresponding author), are with Seoul National University. E-mail: {hj, shlee}@hcil.snu.ac.kr, {parkjeong02, jseo}@snu.ac.kr.*
- *Dae Hyun Kim is with Yonsei University. E-mail: dhkim16@yonsei.ac.kr.*
- *Sungbok Shin is with Aviz at Inria-Saclay and Université Paris-Saclay, Saclay, France. E-mail: sungbok.shin@inria.fr.*

DR projections inherently suffer from distortions [34, 49, 61]. Recent works [23, 34, 36, 37, 71] show the importance of mitigating distortions in achieving reliable data analysis.

A common approach to minimize distortions is to find an effective DR technique. The literature has proposed various benchmark studies that compare the performance of DR techniques in producing accurate projections [14, 15, 23, 35, 60, 77]. These benchmark studies provide guides in selecting DR techniques. However, they rely on limited datasets, raising concerns about generalizability. There is no guarantee that the technique winning the benchmark will also outperform others on a new, unseen dataset.

Another strategy to reduce distortions is to search for optimal DR projections for a given dataset. This is done by repeatedly evaluating the accuracy of projections while testing various DR techniques and hyperparameter settings. Various DR evaluation metrics are proposed for the purpose [33]. Local metrics (e.g., Trustworthiness & Continuity [74]) evaluate how well neighborhood structure is preserved, while global metrics (e.g., KL divergence [29]) focus on the preservation of the global distances between points. Cluster-level metrics (e.g., Label-Trustworthiness & Continuity [36]) focus on cluster structure. However, testing diverse DR techniques and hyperparameter settings is time-consuming [50, 69], as we require several iterations to reach local optima [54]. Furthermore, as we lack a definitive way to confirm that the optimum has been reached, iterations might carry on beyond their necessary point, leading to inefficiency.

Our contribution. We propose a dataset-adaptive workflow that makes DR optimization efficient. Our approach measures how difficult it is to accurately project a given dataset in the 2D space and uses this value to predict the maximum accuracy achievable by each DR technique. This helps analysts remove ineffective techniques from the optimization process, reducing computational demand (Fig. 1 DW1). This approach also makes hyperparameter optimization faster by setting the predicted maximum accuracy as a stopping criterion (Fig. 1 DW2).

2.2 Dataset-Adaptive Machine Learning

Making machine learning models dataset-adaptive, i.e., measuring dataset properties to guide the use of models, has been considered an effective strategy to improve their reliability and efficiency. We discuss how clustering, natural language processing, and computer vision fields benefit from dataset-adaptive approaches.

In the clustering field, *Clusterability* metrics [1, 2, 42, 58] are proposed to quantify the extent to which clusters are clearly structured in datasets [1, 2]. Adolfsson et al. [2] show that analysts can enhance the efficiency of data analysis by using clusterability to decide whether they should apply clustering techniques or not. The degree of alignment between clusters and class labels, i.e., *Cluster-Label Matching* (CLM) [4, 32], is also studied to guide the clustering benchmark. Jeon et al. [32] show that using high-CLM datasets makes the benchmark of clustering techniques produce more generalizable results.

The natural language processing field proposes metrics to predict the difficulty of datasets to be learned by machine learning models, i.e., *dataset difficulty*, highlighting their importance in building the models with more reliable outputs [7, 16, 76]. The literature shows that dataset difficulty can be used to predict model accuracy and overfitting before the training. Dataset difficulty is also verified to be effective in improving dataset quality, especially by identifying mislabeled [16] or ambiguously-labeled [76] data points.

Measuring dataset difficulty is also studied in computer vision [52, 57, 68]. Liu et al. [52] develop a dataset difficulty metric that balances the model complexity and accuracy of image segmentation models. Meanwhile, Mayo et al. [57] propose to use human viewing time as a proxy for dataset difficulty. These methods support scaling up the model architecture search or reducing the gap between benchmark results and real-world performance [68].

Our contribution. All these works validate the effectiveness of the dataset-adaptive approach. However, we lack discussion and validation on how DR—one of the most widely studied unsupervised machine learning domains—can be aligned with dataset properties.

This research bridges this gap (1) by demonstrating the importance of measuring the property (which is structural complexity) of HD datasets before optimizing DR projections and (2) by introducing accurate and fast structural complexity metrics.

2.3 Intrinsic Dimensionality Metrics

To the best of our knowledge, intrinsic dimensionality metrics [5, 24, 41] are currently the only available options for measuring how intricate the structure of a given dataset is. Projection-based intrinsic dimensionality metrics identify the dimensionality in which further increments of dimensionality may slightly enhance the accuracy of a particular DR technique. PCA is widely used for this purpose [14, 24]. On the other hand, geometric metrics (e.g., fractal dimension [17, 43, 73]) evaluate how detailed the geometric structure of datasets is. Both align well with our definition of structural complexity (Sect. 4.1); intuitively, if a dataset needs more dimensionality to be accurately projected, the dataset can be considered to be more complex and tends to be less accurately projected in low-dimensional spaces. Recent works show that intrinsic dimensionality provides a grounded basis in selecting the appropriate machine learning model to train on the datasets [5] and improving their efficiency [25, 27].

Our contribution. Intrinsic dimensionality metrics may correlate with structural complexity by intuition. However, they do not satisfy the desired properties of structural complexity metrics to support the dataset-adaptive workflow. For example, projection-based metrics suffer from low generalizability across multiple DR techniques as they depend on a specific DR technique (Sect. 4.1 P1). Also, geometric metrics produce scores that vary with the global scaling of the dataset, compromising their applicability to real-world datasets with diverse scales (Sect. 4.1 P2). We thus propose novel structural complexity metrics (Sect. 5) that satisfy these desired properties. Our metrics outperform intrinsic dimensionality metrics in terms of accurately predicting the ground truth structural complexity (Sect. 6) and in properly guiding the dataset-adaptive workflow (Sect. 7).

3 CONVENTIONAL WORKFLOW FOR FINDING OPTIMAL DR PROJECTIONS

We detail the conventional workflow (Fig. 1CW) to find optimal DR projections with high accuracy. We derive the workflow by reviewing prior work that optimizes hyperparameters of DR techniques to minimize distortions [14, 35, 39, 60, 78].

The conventional workflow aims to find a DR technique $T \in \mathbf{T}$ and hyperparameter setting $H \in \mathbf{H}_T$ that maximizes $C(X, T_H(X))$, where C is a DR evaluation metric. $T_H(X)$ denotes the projection generated using T and H, and $\mathbf{H}_T$ indicates the hyperparameter domain of T. Here, hyperparameter optimization often yields only marginal improvements over well-chosen default parameters [14]. This means that users can naively use default hyperparameters when efficiency is crucial (e.g., when interactivity is critical). Still, optimization remains essential as it consistently ensures improved reliability in visual analytics using DR, and thus cannot be entirely avoided.

3.1 Workflow

(Step 1) Optimizing hyperparameters of DR technique. In the conventional workflow, we first optimize hyperparameters H for individual DR techniques $T \in \mathbf{T}$ while using $C(X, T_H(X))$ as the target function (Fig. 1 CW1). We do so by repeatedly testing various hyperparameter settings for a fixed number of iterations. Here, detecting convergence in DR hyperparameter optimization is nontrivial, given the difficulty of unifying the DR technique and execution metric into one differentiable expression. Therefore, hyperparameter search methods for non-differentiable functions, such as grid search, random search [6], or Bayesian optimization [70], are commonly used.

(Step 2) Selecting the projection with the best accuracy. We compare the accuracy of the optimal projection achieved for each DR technique. We select the one with the highest accuracy as the optimal projection (Fig. 1 CW2).
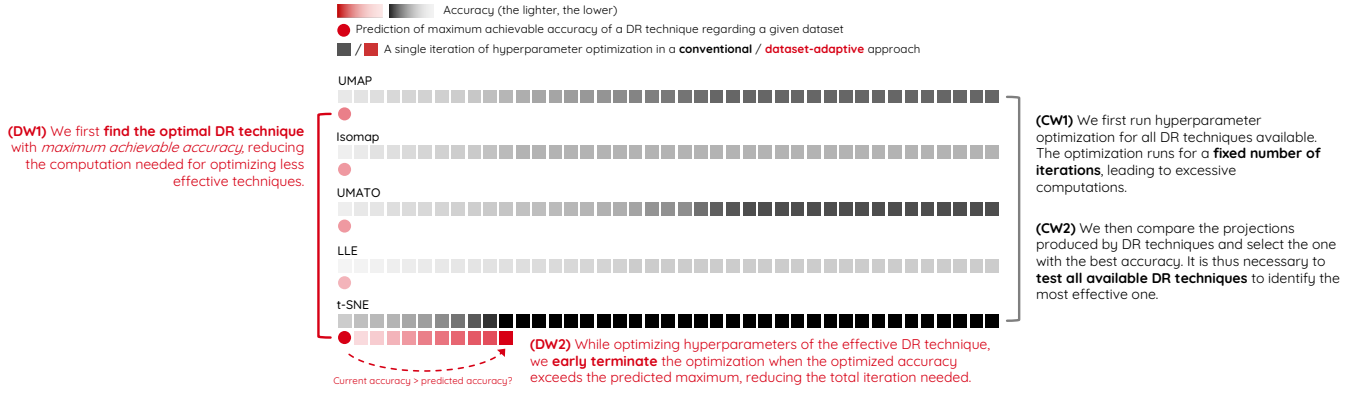
Fig. 1: Illustrations of our *dataset-adaptive* workflow (DW1, DW2) and conventional workflow (CW1, CW2) of finding an optimal DR projection. Each square depicts individual iterations of optimizing DR hyperparameters, where the opacity represents the maximum accuracy achieved by the current and previous iterations. The dataset-adaptive workflow reduces the number of iterations required to discover the optimal projections (red squares) compared to the conventional approach (gray squares).

3.2 Problems

This conventional workflow is time-consuming for two reasons. First, the workflow should test all DR techniques available (Fig. 1 CW2). Second, as detecting convergence in DR hyperparameter optimization is nontrivial, the workflow relies on a fixed number of iterations. If the number of iterations is set too low, the optimization may fail to reach an optimum; if it is set too high, computation is wasted after an optimum has been reached (Fig. 1 CW1). We detail how our dataset-adaptive workflow prevents these redundant computations in Sect. 4.

4 DATASET-ADAPTIVE WORKFLOW

We introduce the dataset-adaptive workflow that reduces the redundant computations in the conventional workflow (Sect. 3) while still achieving an accuracy close to what is achievable in the conventional workflow. The central idea is to quantify how structurally complex a given dataset is, i.e., structural complexity, and then use this value to predict the maximum achievable accuracy of DR techniques in representing the original HD dataset. Here, we can reduce the redundant computation in the conventional workflow by (1) focusing on DR techniques with high maximum predicted accuracy (Fig. 1 DW1), and (2) early terminating hyperparameter optimization when the predicted maximum accuracy is reached (Fig. 1 DW2).

In the following sections, we first define structural complexity and structural complexity metrics. We then detail our dataset-adaptive workflow for optimizing DR projection.

4.1 Structural Complexity and Complexity Metrics

If HD data is structurally complex, DR techniques will inevitably distort these structures, leading to less accurate low-dimensional representations.. Accordingly, we define the ***structural complexity*** of a dataset X as the degree to which even the best possible 2D representation having the same cardinality as X falls short in accurately capturing its structure. Formally, for $X \in \mathbb{R}^{N \times D}$ consisting of N points in D-dimensional space, structural complexity of X is defined as $S_C(X) = -\max_{Y \in \mathbb{R}^{N \times 2}} C(X, Y)$, where C is a DR evaluation metric that measures a projection Y's accuracy in representing X's structural characteristics. Note that we use an additive inverse to align the formula with our definition. We set the output dimension to two since DR is commonly visualized in 2D space in visual analytics. Note that the structural complexity of X can be explained in various aspects, depending on the target structural characteristics of DR evaluation metrics (e.g., local neighborhood structure). This feature is crucial in leveraging structural complexity scores to guide the dataset-adaptive workflow that comprises multiple DR techniques emphasizing distinct structural characteristics.

A ***structural complexity metric*** is a function $f : \mathbb{X} \to \mathbb{R}$ that is desired to have high predictive power with the ground truth structural complexity, where for any HD data X, $X \in \mathbb{X}$. Formally, given a set of datasets $\{X_1, X_2, \cdots, X_N\}$, the predictive power of f with respective to S_C is defined as the degree to which $\{S_C(X_1), S_C(X_2), \cdots, S_C(X_N)\}$ can be accurately predicted by $\{f(X_1), f(X_2), \cdots, f(X_N)\}$. To properly support the dataset-adaptive workflow, we want our metrics to be **(P1) independent of any DR technique**. This is because depending on a certain DR technique may make metrics work properly in predicting maximum achievable accuracy for some DR techniques while working poorly for others. Our metrics should also be **(P2) scale-invariant**, i.e., be independent of the global scale of datasets. Here, global scaling denotes the operation that multiplies an arbitrary positive real number $\alpha > 0$ to all values in the dataset. Global scaling is a characteristic unrelated to data pattern or distribution, thus also independent of the accuracy of projections in representing HD data. Failing to be scale-invariant thus makes the metrics hardly support the dataset-adaptive workflow, as scores can be artificially manipulated through global scaling. Finally, we want our metrics to be **(P3) computationally beneficial**, meaning that computing the metrics should be faster than a single run of the DR technique applied in the dataset-adaptive workflow. This requirement ensures that the workflow can achieve substantial efficiency gains when optimizing DR hyperparameters.

Here, it is worth noting that these requirements may be incomplete. Although satisfying these requirements results in effective structural complexity metrics (Sect. 6, 7, and 8), it remains uncertain whether they are the only requirements needed to characterize the metrics. Identifying new requirements that contribute to designing improved structural complexity metrics would be an interesting avenue to explore.

Validity of our desired properties in the lens of existing dataset property metrics. The fact that P1–P3 are established as the requirements for existing dataset property metrics (Sect. 2.2) further confirms their validity [1, 2, 31]. Adolfsson et al. [2] state that P1, P2, and P3 as requirements for clusterability metrics. Jeon et al. [31] state that P2 and P3 is an important requirement for CLM metrics.

Violations of the desired properties by widely used intrinsic dimensionality metrics. Intrinsic dimensionality metrics are natural candidates for being structural complexity metrics (Sect. 2.3). However, commonly used intrinsic dimensionality metrics do not meet our desired properties (P1 and P2) and thus have limited capability in supporting the dataset-adaptive workflow, justifying the need to design new metrics. For example, two existing methods to compute geometric metrics (correlation method [26] and box-counting method [51]) are not scale-invariant (P2; see Appendix A for proof). In terms of projection-based metrics, they are dependent on DR techniques by design (P1), making them weakly correlate with the ground truth and less applicable to DR optimization involving multiple techniques. For example, if DR techniques that focus on global structures like PCA are used, projection-based metrics may inaccurately predict the maximum achievable accuracy of local techniques like *t*-SNE or UMAP. We empirically

show that these metrics fall behind our proposed metrics (Sect. 5) in correlating with ground truth structural complexity (Sect. 6) and also show that they are not suitable for the dataset-adaptive workflow (Sect. 7).

4.2 Workflow

We detail our dataset-adaptive workflow.

Pretraining. The workflow requires a pretraining of regression models that predict the maximum achievable accuracy of DR techniques from structural complexity scores. We first prepare a set of HD datasets $\mathbf{X}$, a set of structural complexity metrics $\mathbf{f}$, a set of DR techniques $\mathbf{T}$, and a DR evaluation metric C. Then, for each dataset $X \in \mathbf{X}$, we compute $f(X)$ for all structural complexity metrics $f \in \mathbf{f}$, and also compute the optimal accuracy achievable by each technique $T \in \mathbf{T}$ by optimizing hyperparameters H while using $C(X, T_H(X))$ as target function, where we denote this value as $\hat{S}_{C,T}(X)$. Finally, for each $T \in \mathbf{T}$, we train a regression model that predicts $\hat{S}_{C,T}(X)$ from $\{f(X)|f \in \mathbf{f}\}$.

(Step 1) Finding effective DR techniques. Given an unseen dataset X', we start by predicting $\hat{S}_{C,T}(X')$ for each technique $T \in \mathbf{T}$ from $\{f(X')|f \in \mathbf{f}\}$ (Fig. 1 DW1). We then optimize the hyperparameters of only the techniques that have high predicted $\hat{S}_{C,T}(X')$, eliminating the redundant computations in running hyperparameter optimization on potentially less effective DR techniques.

(Step 2) Early terminating hyperparameter optimization. We then optimize the hyperparameters for a selected method T, where we halt the iteration early if the accuracy reaches $\hat{S}_{C,T}(X)$. This early termination prevents unnecessary optimization iterations that would likely result in a minimal gain in improving accuracy (Fig. 1 DW2). Note that we set the threshold exactly at $\hat{S}_{C,T}(X)$ to minimize the risk of sacrificing accuracy; choosing a lower threshold (e.g., 95% of $\hat{S}_{C,T}(X)$) will increase the efficiency but at the cost of further accuracy loss.

5 STRUCTURAL COMPLEXITY METRICS FOR DATASET-ADAPTIVE WORKFLOW

We introduce two structural complexity metrics—*Pairwise Distance Shift* (PDS), *Mutual Neighbor Consistency* (MNC)—that are tailored for the dataset-adaptive workflow. These metrics quantify phenomena that become more pronounced in higher-dimensional spaces as a proxy for structural complexity. PDS estimates the complexity of global structure such as pairwise distances between data points, and MNC focuses on local neighborhood structures. PDS and MNC thus complement each other in estimating ground truth structural complexity (Sect. 6) and guiding dataset-adaptive workflow (Sect. 7), which motivates us to propose their ensemble (PDS+MNC) as an additional metric.

Note that these two metrics depend solely on the distance matrix of the input data. This means that any HD data, regardless of its type (e.g., image, tabular), can be applied as long as the distances between pairs of points can be defined.

5.1 Pairwise Distance Shift (PDS)

PDS estimates the degree of shift made by pairwise distances between data points [45, 47] (also known as distance concentration [22]) as a proxy for complexity regarding global structure. The shift refers to a common phenomenon of HD space in which pairwise distances tend to have a large average and small deviation, shifting its distribution to the positive side [22, 46]. Since it is naturally easier to create complex, irreducible patterns with more dimensions (Theorem 4 in Appendix A), we design PDS to use the degree of shift as a target measurand.

Theoretical and empirical verification. Theorem 1 (Appendix A) confirms that distance shift happens more in HD spaces when the dataset is independent and identically distributed (i.i.d.); thus, measuring the shift in pairwise distances provides an effective proxy for structural complexity. We also empirically verify the existence of distance shift in Appendix A.2.

Note that as we rely on the assumption that the data follows i.i.d. density functions, the phenomenon may not be generalized to real-world datasets. We thus conduct experiments with real-world datasets (Sect. 6, 7) to verify the practical applicability of our metrics.
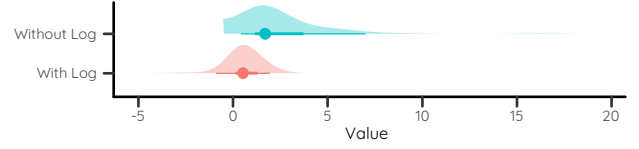

Fig. 2: The distribution of PDS scores across 96 datasets with and without log transformation. PDS is highly skewed without log transformation.

5.1.1 Algorithm

The detailed steps to compute PDS are as follows.

(Step 1) Compute pairwise distances. For a given HD dataset X, we first compute the set of pairwise distances $D(X)$.

(Step 2) Compute pairwise distance shift. We then compute the final score representing the degree of shift as:

$$\text{PDS} = \log \sigma(D(X))/\mathbf{E}(D(X)),$$

where $\sigma(D(X))$ and $\mathbf{E}(D(X))$ represent the standard deviation and the average of $D(X)$, respectively. Without log transformation, the distribution of PDS scores is highly skewed (Fig. 2), which can harm its applicability [20]. As $0 < \sigma(D(X)) < \infty$ and $0 < \mathbf{E}(D(X)) < \infty$, $-\infty < \text{PDS} < \infty$. However, as $\sigma(D(X))$ is typically less than $\mathbf{E}(D(X))$, PDS mostly ranges from $-\infty$ to 0. Lower scores indicate more distance shift, i.e., higher structural complexity.

5.1.2 Compliance with the Desired Properties

We confirm that PDS complies with our desired properties for structural complexity metrics (P1–P3).

(P1) Independence to any DR techniques. The metric only relies on Euclidean distance, thus independent of a specific DR technique.

(P2) Scale-invariance. For any scaling factor $\alpha > 0$ and dataset X, we have $D(\alpha X) = \alpha D(X)$. Here, we obtain:

$$\text{PDS}(\alpha X) = \frac{\sigma(D(\alpha X))}{E(D(\alpha X))} = \frac{\sigma(\alpha D(X))}{E(\alpha D(X))}$$
$$= \frac{\alpha \sigma(D(X))}{\alpha E(D(X))} = \frac{\sigma(D(X))}{E(D(X))} = \text{PDS}(X).$$

Therefore, PDS is scale-invariant.

(P3) Computational benefit. The computations of $D(X)$, $E(D(X))$, and $\sigma(D(X))$ are $O(N^2)$. However, each step can be fully parallelized, making PDS to highly scalable (see Sect. 5.4, 6). We empirically verify that PDS is computationally beneficial, i.e., faster than a single run of typical DR techniques like t-SNE and UMAP, in Appendix D.

5.2 Mutual Neighbor Consistency (MNC)

MNC computes the consistency between k-Nearest Neighbors (kNN) similarity [12] and Shared Nearest Neighbors (SNN) similarity [30, 53] as a proxy for the structural complexity of local structure. While kNN similarity considers a point and its kNN to be similar, SNN similarity regards the pair of points that share more kNN to be more similar.

MNC relies on the phenomenon that higher-dimensional spaces tend to make datasets exhibit greater inconsistency between kNN and SNN. This phenomenon originates from the pairwise distance shift. For any point p in a HD dataset, the distance shift makes all other points equidistantly far from p, distributed on a thin hypersphere C_p centered on p. Consequently, all other points are equally likely to be a kNN of p with probability $k/(N-1)$, where N denotes the number of points in the dataset. Now, consider p and its kNN q. To make p and q to become SNN, a point r that (1) stays within the intersection of C_p and C_q, and (2) is a kNN of both p and q should exist. The first condition has a low probability of occurring as C_p and C_q are very thin. The second condition is also unlikely to be satisfied since the probability is $k^2/(N-1)^2$, where $N \gg k$. Therefore, lower kNN-SNN consistency

likely indicates high dimensionality, which increases the likelihood of complex or irreducible patterns. We thus design MNC to target this consistency as a key measurand.

Theoretical and empirical verification. We theoretically and empirically verify that the inconsistency between kNN and SNN becomes higher in HD spaces when the dataset is i.i.d. (Theorem 2 in Appendix A, Appendix A.2), implying that measuring such inconsistency provides a proxy for structural complexity.

5.2.1 Algorithm

We detail the procedure of computing MNC with a hyperparameter k.

(Step 1) Compute the kNN similarity matrix. For a given dataset $X = \{p_1, p_2, \cdots, p_N\}$, we first compute kNN matrix M^{kNN}, which is an adjacency matrix of X that regarding that two points are connected if they are in kNN relationship. Formally,

$$M_{i,j}^{kNN} = \max(0, k - r + 1) \text{ if } i \neq j \text{ else } 0,$$

where p_j is p_i's r-th NN determined by Euclidean distance.

(Step 2) Compute the SNN similarity matrix. We compute SNN similarity matrix M^{SNN} by setting diagonal elements as 0 and every (i, j)-th non-diagonal elements as:

$$M_{i,j}^{SNN} = \sum_{(m,n) \in N_{p_i, p_j}} (k + 1 - m) \cdot (k + 1 - n),$$

where $(m, n) \in N_{p_i, p_j}$ if m-th NN of p_i is identical to n-th NN of p_j.

(Step 3) Compute the discrepancy between kNN and SNN matrices. At last, the final MNC score of X is computed as:

$$\text{MNC}(X) = \sum_{i \in \{1, \cdots, N\}} \cos(M_{i,*}^{kNN}, M_{i,*}^{SNN}) / N,$$

where $M_{i,*}$ denotes the i-th row of M, and cos designates cosine similarity. We use cosine similarity as it is invariant to the scaling of similarity matrices (P2) and works robustly regardless of N. As cosine similarity ranges from 0 to 1, MNC also ranges from 0 to 1 ($0 \leq \text{MNC} \leq 1$). Lower scores indicates more inconsistency, i.e., higher structural complexity.

5.2.2 Compliance with the Desired Properties

We discuss how MNC meets our desired properties for structural complexity metrics (Sect. 4.1).

(P1) Independence to any DR techniques. MNC relies only on distance functions (kNN and SNN) and is independent of any DR technique.

(P2) Scale-invariance. MNC scores do not change due to the global scaling of datasets as both kNN and SNN are scale-invariant. kNN is scale-invariant as the ranking of NNs is not affected by scaling. SNN is also scale-invariant as it only relies on kNN similarity.

(P3) Computational benefit. kNN and SNN matrix construction requires $O(kN \log N)$ and $O(kN^2)$, respectively. However, kNN can be accelerated up to $O(\log^2 |k|)$ [40], and for the SNN matrix, we can compute every matrix cell in parallel. Computing cosine similarity over rows can also be fully parallelized as every pair of rows can be treated individually. Such parallelization makes the run of MNC to be significantly faster than typical DR techniques (empirical validation in Appendix D). We empirically demonstrate the practical efficiency of MNC in our quantitative analysis (Sect. 6).

5.3 PDS+MNC

We propose to leverage the ensemble of PDS and MNC, which we call PDS+MNC. This can be one by using both PDS and MNC scores as independent variables of regression models in predicting structural complexity or the maximum achievable accuracy of DR techniques. As PDS and MNC focus on global and local structure, respectively, we can expect PDS+MNC to generally work well in practice. We recommend to use multiple MNC with different k values as regression models theoretically work better when with more variables.

5.4 Implementation

We develop our metrics using Python. We use `CUDA` provided by `numba` [44] to optimize and parallelize the algorithms. We also exploit `faiss` [40] to parallelize kNN computation. One limitation here is that we store distance matrices within the memory of a single GPU; thus, we cannot deal with large datasets (see Appendix D for details). Improving the implementation to handle larger data will be a critical future work to increase the practical applicability of the metrics.

6 EXPERIMENT 1: VALIDITY OF PDS+MNC

We evaluate how accurate our structural complexity metrics can predict the ground truth structural complexity.

6.1 Objectives and Study Design

We want to verify that PDS, MNC, and PDS+MNC have high predictive power towards ground truth structural complexity, comparing them against baselines (intrinsic dimensionality metrics). We also aim to provide practical guidelines for selecting structural complexity metrics for different DR evaluation metrics C. We first approximate the ground truth structural complexity of HD datasets following its definition (Sect. 4.1). We then evaluate the accuracy of our structural complexity metrics and baselines in predicting the approximated ground truth.

Approximating ground truth structural complexity. The procedure of approximating ground truth structural complexity aligns with its definition (Sect. 4.1). For a given evaluation metric C, we approximate ground truth $S_C(X)$ by identifying a projection Y that has maximal accuracy $C(X, Y)$, and using this accuracy as the ground truth structural complexity. Exhaustively testing every possible 2D projection guarantees finding the global optimum but requires testing an infinite number of projections. Instead, we test projections with a high probability of being local optima to approximate Y. We do so by leveraging the ensemble of multiple DR techniques. We prepare DR techniques and identify the optimal accuracy of each technique using Bayesian optimization [70]. Then, the maximum optimal accuracy obtained by DR techniques is set as an approximated ground truth structural complexity.

For DR techniques, we first use t-SNE, UMAP, LLE [66], and Isomap [72], which are currently (Mar 2025) the most widely referenced DR techniques according to their citation numbers in Google Scholar. We also use PCA as a representative global DR technique. Finally, we use UMATO [35] to add more diversity, as the technique aims to balance the preservation of local and global structures. Refer to Appendix B for the technical details, e.g., hyperparameter settings.

For C, we use five representative DR evaluation metrics [33] that quantify the distortions focusing on diverse structural characteristics. We pick two local metrics (T&C, MRRE), two global metrics (Spearman's ρ and Pearson's r), and one cluster-level metric (Label-T&C). For T&C, MRREs, and Label-T&C, we use F1 score of two scores produced by the metric. Note that we constrain C to satisfy P2 to make their scores comparable across datasets. For example, Steadiness & Cohesiveness [34], KL divergence [29], and DTM [10] are excluded as they are not independent of global scaling (proofs in Appendix A). Please also refer to Appendix B for the technical details.

HD datasets. We use 96 real-world HD datasets [32] having diverse characteristics (e.g., number of points, dimensionality). For the datasets exceeding 3,000 points, we randomly select a sample of 3,000 points. This is because computing approximated ground truth requires too much time (e.g., more than a week) for some datasets.

Baselines. We use both projection-based and geometric intrinsic dimensionality metrics (Sect. 2.3) as baselines. We use the PCA-based method for the projection-based metric due to its efficiency (P3) and popularity in literature [14, 18, 25]. We compute the number of principal components required to explain more than 95% of data variance, following Espadoto et al. [14]. For the geometric metric, we use the correlation method due to its robustness in HD compared to the box-counting method [26] (Appendix C).

Measurement. We evaluate how well our metrics and baselines correlate with the approximated ground truth structural complexity by

Table 1: Results of our analysis on the accuracy of structural complexity metrics in predicting the ground truth (Sect. 6). Each cell depicts the correlation between structural complexity metrics or intrinsic dimensionality metrics (columns) and ground truth structural complexity approximated using five DR evaluation metrics (rows). PDS and MNC show high correlations with global and local structural complexity, respectively, outperforming intrinsic dimensionality metrics. Their ensemble (PDS+MNC) achieves the best correlation for all cases.

		Local		Cluster	Global	
		T&C	MRREs	L-T&C	S-ρ	P-r
Int. Dim. (Projection)	LR	.5574	.5133	.3364	.3629	.3576
	PR	.5334	.4505	.2369	.4231	.3769
	kNN	.6353	.5537	.3770	.6495	.6281
	RF	.5815	.4710	.3013	.6309	.5996
	GB	.6132	.4021	.3183	.6278	.5652
Int. Dim. (Geometric)	LR	< 0	.0142	.0065	.0689	.0478
	PR	.0123	0.2351	.0045	.0716	.0566
	kNN	.5214	.3656	.3755	.5826	.5967
	RF	.5054	.3807	.3958	.6010	.5881
	GB	.5587	.4430	.3822	.5711	.5730
PDS	LR	.2781	.0673	.3404	.5152	.5332
	PR	.3440	.0866	.4032	.7304	.7164
	kNN	.4133	.3331	.4998	.8003	.8180
	RF	.4119	.3509	.5010	.8282	.8217
	GB	.4523	.3182	.4681	.7959	.8075
MNC	LR	.8454	.6784	.3692	.5677	.5241
	PR	.8807	.7244	.3174	.5525	.5140
	kNN	.8780	.7007	.4020	.5962	.5814
	RF	.8706	.7302	.4189	.6166	.5734
	GB	.8666	.7207	.2827	.5992	.5741
PDS+MNC	LR	.8513	.7484	.5104	.6772	.7474
	PR	***.8984***	***.8423***	.0015	.5954	.6572
	kNN	.7472	.6109	.4971	***.8290***	***.8401***
	RF	.8881	.7506	.5823	.8273	.8092
	GB	.8694	.7636	***.6067***	.8280	.8079

1. ■ / ■ / ■: very strong ($R^2 \geq 0.8$) / strong ($0.6 \leq R^2 < 0.8$) / moderate ($0.4 \leq R^2 < 0.6$) correlation [67]

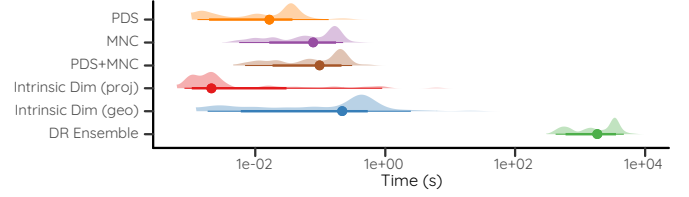2. ***Bold and italic*** refers to the top score of the column

Fig. 3: The runtime of our structural complexity metrics, intrinsic dimensionality metrics, and DR ensemble to produce ground truth structural complexity. PDS, MNC, and PDS+MNC are faster than the geometric intrinsic dimensionality metric and the ensemble. The thick and thin lines indicate the 66% and 99% interval of probability mass.

computing the capacity of regression models to predict the approximated ground truths from metric scores. This is done by quantifying average R^2 correlation scores obtained by five-fold cross-validation. As we only have 96 datasets, the results may depend on how datasets are split for cross-validation. Therefore, for each metric, we repeat the measurement 100 times with different splits and report the maximum score, i.e., the ideal correlation.

To avoid bias, we use five regression models: linear regression (LR), Polynomial regression (PR), kNN regression (kNN), Random forest regression (RF), and Gradient boosting regression (GB). Please refer to Appendix B for technical details.

Hyperparameters. We set $k = 50$ for MNC; we show the robustness of MNC against varying k in Appendix E. In terms of PDS+MNC, theoretically, adding more variables always increases the predictive power of regression models. However, as we have a small number of datasets, adding more variables can make the models suffer from data sparsity. Our ensemble thus consists of PDS and three MNCs ($k = 25, 50,$ and 75). All other competitors do not have hyperparameters.

Apparatus. We execute the experiment using a Linux server with 40-core Intel Xeon Silver 4210 CPUs, TITAN RTX, and 224GB RAM. We use a single GPU to compute complexity metrics. We use this machine also for the following experiments.

6.2 Results and Discussions

The followings are the findings of our experiment.

Validity of the design of PDS and MNC. The results (Table 1) verify that PDS and MNC are appropriately designed and work as intended. While PDS shows at most a very strong or strong correlation with global structural complexity (S-ρ and P-r), MNC achieves at most a very strong correlation with local structural complexity (T&C and MR-REs). Meanwhile, both metrics show relatively weak correlations in the opposite cases; showing at most strong or moderate correlations. PDS also shows at most strong correlation with linear and polynomial regression models (LR, PR). As these models do not work well with non-Gaussian data, such results verify the validity of log transformation (Step 2). In contrast, intrinsic dimensionality metrics correlate weakly with the ground truth. For all types of ground truth structural complexity, the best correlation achieved by these metrics is substantially lower than that achieved by PDS, MNC, and PDS+MNC.

Superior predictive power of PDS and MNC. The results moreover confirm the effectiveness of PDS+MNC. PDS+MNC achieves the best correlation regardless of DR evaluation metrics used to approximate ground truth structural complexity, making substantial performance gains compared to using MNC and PDS individually. For example, PDS+MNC has at most a strong correlation even for cluster-level structural complexity. PDS+MNC is also the sole subject that achieves a very strong correlation for the structural complexity approximated by MRREs. Such results underscore the importance of testing alternative strategies to ensemble complexity metrics or identifying the most effective combination. Aligned with this finding, we use PDS+MNC to demonstrate the applicability of our metrics in supporting our dataset-adaptive workflow (Sect. 7).

Weaknesses of our structural complexity metrics. We also reveal the weaknesses of our complexity metrics. PDS and MNC show at most a weak correlation with cluster-level structural complexity (L-T&C). Although PDS+MNC shows at most a strong correlation, the best R^2 value obtained is substantially lower compared to the ones from global and local structural complexity. Also, MNC has a relatively weak correlation with local structural complexity approximated by MRREs compared to the one approximated by T&C. We recommend not using the structural complexity metrics for such cases. For example, we recommend using PDS and PDS+MNC when S-ρ or P-r is used, but not MNC. This indicates that we currently lack structural complexity metrics that can be reliably used for cluster-level evaluation metrics (L-T&C). Such results underscore the need to design new, advanced structural complexity metrics that can complement PDS, MNC, and PDS+MNC.

Computational efficiency of the metrics. We also investigate the execution time of PDS, MNC, PDS+MNC, and baselines in being applied to 96 real-world datasets we use. We also examine the execution time for approximating ground truth structural complexity using the ensemble of DR techniques. Note that we rely on GPU-based parallelization and highly optimized libraries like `scikit-learn` [64] and PyTorch [62] while implementing intrinsic dimensionality metrics to ensure fairness.

As a result (Fig. 3), we find that our structural complexity metrics

Table 2: Accuracy of baseline metrics and PDS+MNC on predicting maximum accuracy achievable by different DR techniques (Sect. 7.1). PDS+MNC shows strong predictability for the majority of cases, outperforming baseline in the majority of cases.

		T&C	MRREs	L-T&C	S-ρ	P-r
Int. Dim. (Projection)	UMAP	.6540	.6449	.2656	.6389	.3690
	t-SNE	.7615	.6643	.0212	.5746	.1674
	PCA	.8027	.8430	.6524	.7112	.6445
	Isomap	.7998	.7924	.6094	.7326	.6557
	LLE	.7747	.7846	.2987	.6737	.7473
	UMATO	.7207	.6906	.3222	.7109	.2838
Int. Dim. (Geometric)	UMAP	.7146	.6519	.4204	.6550	.0405
	t-SNE	.5097	.5767	.0029	.3303	.1257
	PCA	.8588	.8417	.1720	.7989	.7160
	Isomap	.8253	.7069	.5629	.8544	.8855
	LLE	.8496	.4476	.7058	.7123	.7539
	UMATO	.8438	.7038	.2275	.5660	.1383
PDS+MNC	UMAP	.8955	.8183	.3944	.7978	.7578
	t-SNE	.9471	.7848	.4303	.8600	.6569
	PCA	.9150	.8446	.5784	.9063	.8836
	Isomap	.8514	.9123	.7962	.9454	.8935
	LLE	.8781	.8501	.6986	.7645	.7739
	UMATO	.9223	.7823	.5733	.8392	.6880

■ / ■ / ■: very strong ($R^2 \geq 0.8$) / strong ($0.6 \leq R^2 < 0.8$) / moderate ($0.4 \leq R^2 < 0.6$) predictive power [67]
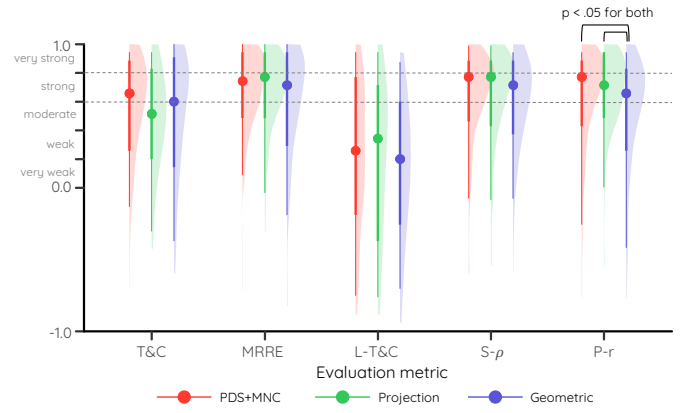

Fig. 4: The distribution of the correlation between the true accuracy ranking of DR techniques and the predicted ranking estimated based on PDS+MNC and the baselines (Sect. 7.2). We use the criteria of Prion and Hearing [65] to explain the strength of the correlation (very strong to very weak). Overall, PDS+MNC demonstrates strong predictive power for the majority of cases. We also find that all three metrics have similar performance in predicting the ranking. The thick and thin lines indicate the 66% and 99% interval of probability mass.

are slower than the projection-based intrinsic dimensionality metric but are faster than other baselines. All three metrics required less than one second for all datasets, substantially scaling up the DR ensemble method. Regarding the high accuracy of PDS and MNC in predicting the ground truth structural complexity produced by the DR ensemble (Sect. 6), the result indicates that we achieve a favorable balance between efficiency and accuracy. We further examine the efficiency of our metrics using larger datasets in Appendix D.

7 EXPERIMENT 2: SUITABILITY OF PDS+MNC FOR THE DATASET-ADAPTIVE WORKFLOW

We want to examine the utility of PDS+MNC, the most advanced structural complexity metrics, in properly guiding the dataset-adaptive workflow (Sect. 4.2). As with a previous experiment (Sect. 6), the experiment also aims to provide guidelines to select structural complexity metrics in practice. We first evaluate the accuracy of the pretrained regression models in predicting the maximum achievable accuracy of DR techniques from the scores from PDS+MNC (Sect. 7.1). We then evaluate the effectiveness of PDS+MNC in guiding each step of our workflow (Sect. 7.2 for Step 1 and Sect. 7.3 for Step 2).

7.1 Evaluation on Pretraining Regression Models

We detail our evaluation on the suitability of PDS+MNC in pretraining regression models for the dataset-adaptive workflow.

7.1.1 Objectives and Study Design

We want to evaluate the utility of PDS+MNC in training regression models that predict the maximum achievable accuracy of DR techniques. We assess the performance of regression models that predict the maximum achievable accuracy by DR techniques from the metric scores, comparing them to the ones that utilize baselines (intrinsic dimensionality metrics). For 96 datasets that we have (Sect. 6), we first obtain maximum scores of DR techniques by optimizing them using Bayesian optimization (detailed settings in Appendix B). We then compute the R^2 scores of regression models predicting the maximum accuracy from PDS+MNC and baselines. This is done by (1) splitting 96 datasets into 80 training datasets and 16 test datasets, (2) training the regression model with a training dataset with five-fold cross-validation, and (3)

assessing the performance of the model for unseen test datasets. To make scores comparable with previous experiments (Sect. 6), we repeat the measurement 10 times with different splits and report the ideal performance. For the structural complexity metric, we use PDS+MNC as it shows the best correlation with approximated ground truth structural complexity (Sect. 6).

Regression models. To examine ideal prediction, we use AutoML [28] based on `auto-sklearn` [21] to train the regression model. We train the model for 60 seconds. We do not test the five regression models used in our correlation analysis as they hardly achieve ideal predictions.

DR techniques and metrics. We use the same set of DR techniques and evaluation metrics with the correlation analysis (Sect. 6).

7.1.2 Results and discussions

Table 2 depicts the results. Overall, PDS+MNC has strong predictability with maximum accuracy achievable by DR techniques, outperforming baseline metrics. For example, while PDS+MNC shows at least a strong predictive power for S-ρ and P-r, the baselines fail to do so. PDS+MNC also has at least strong predictability for most combinations of DR techniques and evaluation metrics (26 out of 30; 87%), verifying that the metric can be generally applied to execute the dataset-adaptive workflow in practice. In contrast, projection-based and geometric projection metrics show at least strong predictability for 21 and 17 combinations, respectively.

However, PDS+MNC relatively works poorly for L-T&C, showing moderate or weak correlations for four cases. Such results align with the results from our correlation analysis (Sect. 6), clarifying the need for further development of complexity metrics that complement PDS and MNC. PDS+MNC still outperforms baselines also for L-T&C.

7.2 Evaluation on Predicting Effective DR Techniques

We investigate the effectiveness of PDS+MNC in predicting the accuracy ranking of DR techniques (Step 1; Fig. 1 DW1).

7.2.1 Objectives and Study Design

We examine whether the pretrained regression models using PDS+MNC can effectively distinguish DR techniques with high and low maximum achievable accuracy in the testing phase. This capability is important in properly guiding dataset-adaptive workflow to avoid the optimization of suboptimal DR techniques (Sect. 4, Fig. 1 DW1).

We reuse the maximum achievable accuracy of DR techniques computed previously (Sect. 7.1). Leveraging this score, we identify the

Table 3: The effectiveness of structural complexity metrics (PDS+MNC) and baselines (projection-based on geometric intrinsic dimensionality metrics) in guiding the early termination of hyperparameter optimization (Table 7.3). We report the average error across all trials, the worst-case errors at the 10%, 5%, and 1% percentiles, and the relative time required compared to full optimization. Each blue bar within a table cell represents the cell's relative value compared to the maximum value for each combination of DR evaluation metrics (rows) and measurands (columns). The bold denotes the best performance for each combination. The results verify that the approximation using PDS+MNC consistently achieves desirable accuracy while maintaining efficiency.

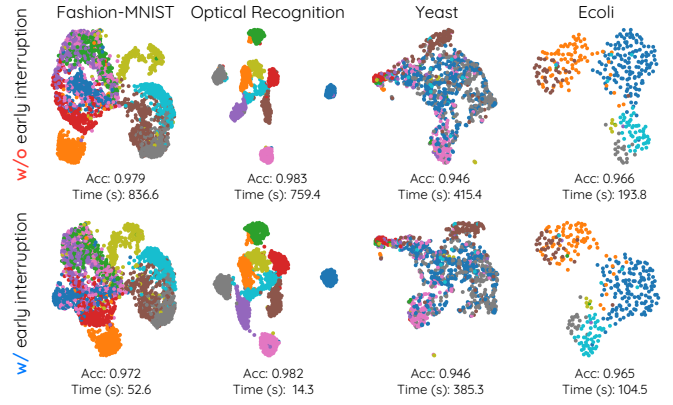| | | Error | | | | Rel. Time |
		All	10%	5%	1%	
UMAP (w T&C)	Projection	.0005	**.0112**	.0179	.0261	40.6%
	Geometric	.0005	.0157	.0192	.0285	**29.0%**
	PDS+MNC	**.0004**	.0118	**.0159**	**.0246**	35.7%
UMAP (w P-r)	Projection	.0395	.1287	.1560	.1913	50.3%
	Geometric	.0380	.1473	.1758	.2052	**44.7%**
	PDS+MNC	**.0173**	**.0639**	**.0861**	**.1005**	51.0%

Fig. 5: The projections made by optimizing UMAP with and without interruption based on predicted maximum accuracy (Sect. 7.1). The interruption substantially reduce runtime while maintaining accuracy.

ground truth accuracy ranking of DR techniques. We then train the regression models with 80 training datasets as we do in the previous experiment (Sect. 7.1) using PDS+MNC and baselines (intrinsic dimensionality metrics). Based on the models' approximated accuracy predictions, we derive the predicted accuracy ranking of DR techniques and compute the rank correlation between the ground truth and predicted rankings. We run the same experiment 10 times with different splits of datasets and report the average results.

DR techniques and metrics. We use the same set of DR techniques and evaluation metrics with the previous experiments (Sect. 6, 7.1).

7.2.2 Results and discussions

We find that PDS+MNC achieve strong accuracy in predicting the ground truth ranking of DR techniques computed by T&C, MRRE, S-ρ, and P-r (Fig. 4 red marks). The result reaffirms the capability of structural complexity metrics in guiding the dataset-adaptive workflow. In contrast, the rankings predicted by PDS+MNC show weak correlation with the ground truth for L-T&C case, aligning with our previous experiments (Sect. 6, 7.1).

We also find that baselines show good performance in predicting rankings that correlate well with ground truth rankings. We examine the difference in correlations between PDS and two baselines for each evaluation metric using ANOVA. As a result, we find no significant difference for T&C ($F_{2,477} = 2.26$, $p = .104$), MRRE ($F_{2,477} = 2.71$, $p = .067$), L-T&C ($F_{2,477} = 1.48$, $p = .226$), and S-ρ ($F_{2,477} = 0.91$, $p = .403$). We find a significant difference for P-r case ($F_{2,477} = 5.37$, $p < .01$), and post-hoc test using Tukey's HSD confirms that PDS+MNC and projection-based intrinsic dimensionality metric significantly outperform geometric intrinsic dimensionality metric ($p < .05$ for both). Still, the geometric intrinsic dimensionality metric shows strong predictive power. These results indicate that distinguishing between effective and ineffective techniques does not require precise maximum accuracy predictions. We discuss the takeaways of this phenomenon on the practical use of dataset-adaptive workflow in Sect. 9.1.

7.3 Evaluation on Early Terminating Optimization

We assess the utility of PDS+MNC and regression models in accelerating hyperparameter optimization by early terminating iterations.

7.3.1 Objectives and Study Design

We investigate the effectiveness of PDS+MNC in improving the efficiency of DR hyperparameter optimization by reducing redundant iterations (Fig. 1 DW2), comparing with baseline metrics (intrinsic dimensionality metrics). We simulate the optimization of hyperparameters for UMAP projections. We evaluate using T&C and P-r to compare two scenarios: one in which PDS+MNC and baseline metrics exhibit similar performance (T&C) and one in which PDS+MNC substantially outperforms the baselines in predicting the maximum achievable accuracy of DR techniques (P-r). This selection is based on our evaluation of pretraining regression models (Sect. 7.1). As with previous experiments, we divide the 96 HD datasets into 80 training datasets and 16 test datasets and train the AutoML regression model to predict the maximum accuracy from the training dataset. We then predict the maximum accuracy of 16 unseen test datasets. Note that for T&C, we use the F1 score of Trustworthiness and Continuity, following the convention of interpreting T&C as precision and recall of DR [75].

Finally, we optimize UMAP on 16 datasets with and without interruption based on the predicted maximum accuracy. For the former (with), we run Bayesian optimization with 50 iterations and halt the process upon reaching the optimal score. We set the default iteration number as 50, a default value recommended by `scikit-optimize` [54] library. For the latter condition (without), we run the optimization with 50 iterations without halting the process. We compare the two settings by evaluating how much error is introduced and assessing the relative running time compared to the optimizations executed without early termination. We run the same experiment 10 times with different splits of datasets. Note that we report the average error over all trials and specifically examine errors from the worst 10%, 5%, and 1% trials to detail the robustness of the early termination.

7.3.2 Results and discussions

Table 3 depicts the results, and Fig. 5 depicts the subset of projections generated in our experiment. The results demonstrate the efficacy of PDS+MNC in guiding the early termination of hyperparameter optimization, verifying its usefulness in executing the dataset-adaptive workflow in practice. Specifically, early termination using PDS+MNC substantially reduces runtime while incurring only minimal errors. While reducing runtime by more than half, PDS+MNC outperforms the baselines in terms of error in the majority of cases. In the P-r scenario, the error is approximately halved compared to the baselines.

8 EXPERIMENT 3: EFFECTIVENESS OF THE DATASET-ADAPTIVE WORKFLOW

We evaluate how well our dataset-adaptive workflow (Sect. 4) accelerates the process of finding optimal DR projection without compromising accuracy, comparing it with conventional workflow (Sect. 3).

8.1 Objectives and Study Design

We aim to verify two hypotheses:

H1 The dataset-adaptive workflow significantly accelerates the DR optimization process compared to the conventional workflow.

A. Trustworthiness & Continuity (T&C)
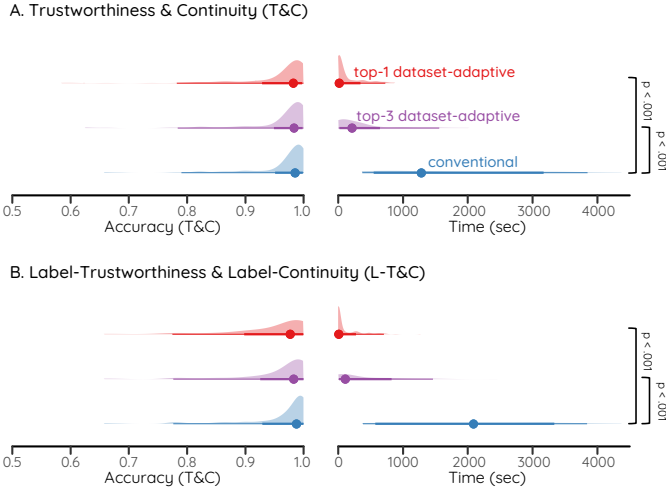
B. Label-Trustworthiness & Label-Continuity (L-T&C)

Fig. 6: Comparison of the performance of three different workflows in optimizing DR projections. Dataset-adaptive workflows achieve a significant gain in execution time with negligible accuracy loss. The thick and thin lines indicate the 66% and 99% interval of probability mass.

H2 The dataset-adaptive workflow finds DR projections with negligible accuracy loss compared to the conventional workflow.

We first split 96 HD datasets into 80 training datasets and 16 test datasets. Then, we train regression models for all DR techniques we use previously (Sect. 7.1). We then execute conventional (Sect. 3) and dataset-adaptive (Sect. 4) workflows for optimizing DR projection for test datasets, where we set the default iteration number as 50.

For our dataset-adaptive workflow, we test two variants: one optimizes the hyperparameters of the top-1 DR technique, and the other optimizes those of the top-3 techniques. This is because we want to examine the tradeoff between the execution time and the accuracy. We record the total execution time and the final accuracy obtained by each workflow (conventional, top-1 dataset-adaptive, top-3 dataset-adaptive). We run the same experiment 10 times with diverse dataset splits.

Evaluation metrics. We aim to examine both the full potential of the dataset-adaptive workflow and its effectiveness under worst-case scenarios. We measure the accuracy of DR projections using T&C and L-T&C, two metrics that PDS+MNC exhibits the best and worst predictive power in our previous experiment (Sect. 7.1), respectively.

8.1.1 Results and Discussions

Fig. 6 depicts the results. By running one-way ANOVA on the accuracy scores, we find no significant difference between workflows for both T&C ($F_{2,477} = 1.35$, $p = .259$) and L-T&C ($F_{2,477} = 2.49$, $p = .084$) cases. The result confirms **H2**. In contrast, ANOVA on the execution time indicates the statistically significant difference between the workflows (T&C: $F_{2,477} = 194.18$, $p < .001$; L-T&C: $F_{2,477} = 190.97$, $p < .001$). We find that top-1 and top-3 dataset-adaptive workflows are $\times 5.3$ and $\times 13$ faster than the conventional workflow, respectively. Post-hoc analysis using Tukey's HSD also confirms that top-1 and top-3 dataset-adaptive workflows require significantly shorter execution time than the conventional workflow ($p < .001$). This result confirms **H1**. Meanwhile, we observe no significant difference between the top-1 and top-3 dataset-adaptive workflows (T&C: $p = .056$, L-T&C: $p = .077$). In summary, our experiment verifies both hypotheses, confirming the benefit of using the dataset-adaptive workflow in practice.

The results also reveal that the dataset-adaptive workflow does not always perform ideally. Although the accuracy differences are not statistically significant, the L-T&C case yields a relatively low p-value (0.084), indicating a slight accuracy difference in the conventional and the dataset-adaptive workflow. Such results suggest that more efforts should be invested in designing advanced structural complexity metrics and further refining the dataset-adaptive workflow.

9 DISCUSSIONS

We discuss future research directions for dataset-adaptive workflow and structural complexity metrics.

9.1 Leveraging the Tradeoffs between Predictive Power and Efficiency

We find that PDS+MNC outperforms intrinsic dimensionality metrics in early terminating hyperparameter optimization (Sect. 7.3) but is similarly effective in identifying optimal DR techniques (Sect. 7.2). This indicates that the predictive power of structural complexity metrics is crucial for the success of the former but not for the latter.

This finding offers a new perspective on designing structural complexity metrics: leveraging the tradeoff between predictive power and efficiency. Although PDS and MNC run within a reasonable time frame for our dataset, they may be too slow for enormous datasets with millions of data points. In such cases, we can use fast but less accurate structural complexity metrics to select the DR technique in Step 1 and then apply accurate metrics like PDS and MNC only to the selected technique in Step 2. We can also dynamically adjust the optimization procedure for multiple DR techniques. For example, we can run multiple DR techniques in parallel and drop those predicted to produce low values. Applying a progressive visual analytics paradigm [19] will be effective here. For instance, we can begin optimization with rough estimates of promising DR techniques and eliminate those that prove to be ineffective later in the process.

To pursue this direction, it will also be crucial to investigate the utility of structural complexity metrics in depth. For example, we may examine the gap between the achieved predictive power and the theoretical optimum, and whether this gap is bounded. When metrics are computed progressively, it will also be important to investigate how the error bounds evolve over time. These endeavors will contribute to making DR-based visual analytics both more responsive and accurate.

9.2 Complementing Our Structural Complexity Metrics

Our experiments reveal the necessity of developing new structural complexity metrics that complement PDS and MNC, especially the ones that focus on cluster-level structure. This is because PDS, MNC, and even PDS+MNC fall short in estimating the ground truth structural complexity and the maximum accuracy of DR techniques computed with L-T&C, a cluster-level evaluation metric. One idea for developing cluster-level structural complexity metrics is to extend MNC to the class level. Instead of considering neighbors of "data points," we could consider those of "classes" or "clusters."

9.3 Exploring Additional Use Cases

We hypothesize that structural complexity metrics can improve the *replicability* of benchmarking DR techniques. DR benchmarks may produce different conclusions about the accuracy of DR techniques depending on the datasets used, i.e., they may have low replicability. Adding more benchmark datasets may enhance replicability, but this increases the computational burden. Here, complexity metrics may contribute to achieving high replicability with fewer datasets. For example, excluding datasets with simple patterns that are accurately reducible by any DR technique will improve replicability, as the rankings of DR techniques determined based on these datasets may noise the evaluation. Validating whether this hypothesis confirms or not will be a worthwhile future research avenue to explore.

10 CONCLUSION

It is important to find optimal DR projections to ensure reliable visual analytics. However, optimization processes are computationally demanding. We propose the dataset-adaptive workflow for accelerating DR optimization while maintaining accuracy. We introduce two structural complexity metrics, PDS and MNC, and verify their effectiveness in terms of precision and efficiency. We also demonstrate the utility of the dataset-adaptive workflow guided by these two metrics. Overall, our proposal opens up the discussion towards achieving a more reliable and efficient analysis of HD data.

REFERENCES

[1] M. Ackerman and S. Ben-David. Clusterability: A theoretical study. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, vol. 5 of *Proceedings of Machine Learning Research*, pp. 1–8. PMLR, 16–18 Apr 2009. 2, 3

[2] A. Adolfsson, M. Ackerman, and N. C. Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, 2019. doi: 10.1016/j.patcog.2018.10.026 2, 3

[3] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7):1304–1330, 2007. doi: 10.1016/j.neucom.2006.11.018 1

[4] M. Aupetit. Sanity check for class-coloring-based evaluation of dimension reduction techniques. In *Proc. of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, 8 pages, p. 134–141, 2014. doi: 10.1145/2669557.2669578 2

[5] J. Bac, E. M. Mirkes, A. N. Gorban, I. Tyukin, and A. Zinovyev. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10), 2021. doi: 10.3390/e23101368 2

[6] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. 2

[7] M. Byrd and S. Srivastava. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 119–130, May 2022. doi: 10.18653/v1/2022.acl-short.15 2

[8] D. Cashman, M. Keller, H. Jeon, B. C. Kwon, and Q. Wang. A critical analysis of the usage of dimensionality reduction in four domains. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–20, 2025. doi: 10.1109/TVCG.2025.3567989 1

[9] A. Chatzimparmpas, R. M. Martins, and A. Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2696–2714, 2020. doi: 10.1109/TVCG.2020.2986996 1

[10] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011. doi: 10.1007/s10208-011-9098-0 5

[11] B. Colange, L. Vuillon, S. Lespinats, and D. Dutykh. Interpreting distortions in dimensionality reduction by superimposing neighbourhood graphs. In *IEEE Visualization Conference*, pp. 211–215, 2019. doi: 10.1109/VISUAL.2019.8933568 1

[12] W. Dong, C. Moses, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, 10 pages, p. 577–586. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1963405.1963487 4

[13] L. Ertöz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High dimensional Data and its Applications at 2nd SIAM International Conference on Data mining*, pp. 105–115, 2002. 1

[14] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182 2, 5

[15] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):81–94, 2015. doi: 10.1109/TVCG.2014.2330617 2

[16] K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathscr{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. 2

[17] K. Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004. 2

[18] M. Fan, N. Gu, H. Qiao, and B. Zhang. Intrinsic dimension estimation of data by principal component analysis, 2010. doi: 10.48550/arXiv.1002.2050 5

[19] J.-D. Fekete, D. Fisher, and M. Sedlmair. Progressive data analysis. Eurographics, 2024. doi: 10.2312/pda.20242707 9

[20] C. Fernández and M. F. J. Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998. doi: 10.1080/01621459.1998.10474117 4

[21] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28, 2015. 7

[22] D. Francois, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007. doi: 10.1109/TKDE.2007.1037 4

[23] C. Fu, Y. Zhang, D. Cai, and X. Ren. Atsne: Efficient and robust visualization on gpu through hierarchical optimization. In *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 11 pages, p. 176–186, 2019. doi: 10.1145/3292500.3330834 1, 2

[24] K. Fukunaga and D. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, C-20(2):176–183, 1971. doi: 10.1109/T-C.1971.223208 2

[25] S. Gong, V. N. Boddeti, and A. K. Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5

[26] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1):189–208, 1983. doi: 10.1016/0167-2789(83)90298-1 3, 5

[27] M. Hasanlou and F. Samadzadegan. Comparative study of intrinsic dimensionality estimation and dimension reduction techniques on hyperspectral images using k-nn classifier. *IEEE Geoscience and Remote Sensing Letters*, 9(6):1046–1050, 2012. doi: 10.1109/LGRS.2012.2189547 2

[28] X. He, K. Zhao, and X. Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021. doi: 10.1016/j.knosys.2020.106622 7

[29] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Proc. of the 15th International Conference on Neural Information Processing Systems*, 8 pages, p. 857–864. MIT Press, Cambridge, MA, USA, 2002. 2, 5

[30] R. Jarvis and E. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, C-22(11):1025–1034, 1973. doi: 10.1109/T-C.1973.223640 4

[31] H. Jeon, M. Aupetit, D. Shin, A. Cho, S. Park, and J. Seo. Sanity check for external clustering validation benchmarks using internal validation measures, 2022. doi: 10.48550/ARXIV.2209.10042 3

[32] H. Jeon, M. Aupetit, D. Shin, A. Cho, S. Park, and J. Seo. Measuring the validity of clustering validation datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2025. doi: 10.1109/TPAMI.2025.3548011 2, 5

[33] H. Jeon, A. Cho, J. Jang, S. Lee, J. Hyun, H.-K. Ko, J. Jo, and J. Seo. Zadu: A python library for evaluating the reliability of dimensionality reduction embeddings. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 196–200, 2023. doi: 10.1109/VIS54172.2023.00048 2, 5

[34] H. Jeon, H.-K. Ko, J. Jo, Y. Kim, and J. Seo. Measuring and explaining the inter-cluster reliability of multidimensional projections. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):551–561, 2021. doi: 10.1109/TVCG.2021.3114833 1, 2, 5

[35] H. Jeon, H.-K. Ko, S. Lee, J. Jo, and J. Seo. Uniform manifold approximation with two-phase optimization, 2022. doi: 10.1109/VIS54862.2022.00025 2, 5

[36] H. Jeon, Y.-H. Kuo, M. Aupetit, K.-L. Ma, and J. Seo. Classes are not clusters: Improving label-based evaluation of dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):781–791, 2024. doi: 10.1109/TVCG.2023.3327187 2

[37] H. Jeon, H. Lee, Y.-H. Kuo, T. Yang, D. Archambault, S. Ko, T. Fujiwara, K.-L. Ma, and J. Seo. Unveiling high-dimensional backstage: A survey for reliable visual analytics with dimensionality reduction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, article no. 394, 24 pages, 2025. doi: 10.1145/3706598.3713551 1, 2

[38] H. Jeon, J. Park, S. Shin, and J. Seo. Stop misusing t-sne and umap for visual analytics, 2025. doi: 10.48550/arXiv.2506.0872 1

[39] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafir, and J. Seo. Clams: A cluster ambiguity measure for estimating perceptual variability in visual clustering. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):770–780, 2024. doi: 10.1109/TVCG.2023.3327201 2

[40] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572 5

[41] S. Kak. Information theory and dimensionality of space. *Scientific Reports*, 10(1):20733, 2020. doi: 10.1038/s41598-020-77855-9 2

[42] A. Kalogeratos and A. Likas. Dip-means: an incremental clustering method for estimating the number of clusters. In *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. 2

[43] R. Karbauskaitė and G. Dzemyda. Fractal-based methods as a technique for estimating the intrinsic dimensionality of high-dimensional data: a survey. *Informatica*, 27(2):257–281, 2016. 2

[44] S. K. Lam, A. Pitrou, and S. Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015. doi: 10.1145/2833157.2833162 5

[45] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer-Verlag New York, 2007. doi: 10.1007/978-0-387-39351-3 1, 4

[46] J. A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 4:538–547, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011. doi: 10.1016/j.procs.2011.04.056 1, 4

[47] J. A. Lee and M. Verleysen. Two key properties of dimensionality reduction methods. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 163–170, 2014. doi: 10.1109/CIDM.2014.7008663 1, 4

[48] S. Lespinats and M. Aupetit. Checkviz: Sanity check and topological clues for linear and non-linear mappings. *Computer Graphics Forum*, 30(1):113–125, 2011. doi: 10.1111/j.1467-8659.2010.01835.x 1

[49] S. Lespinats, M. Verleysen, A. Giron, and B. Fertil. Dd-hds: A method for visualization and exploration of high-dimensional data. *IEEE Transactions on Neural Networks*, 18(5):1265–1279, 2007. doi: 10.1109/TNN.2007.891682 1, 2

[50] Y.-T. Liao, H. Luo, and A. Ma. Efficient and robust bayesian selection of hyperparameters in dimension reduction for visualization, 2023. doi: 10.48550/arXiv.2306.00357 2

[51] L. S. Liebovitch and T. Toth. A fast algorithm to determine fractal dimensions by box counting. *Physics Letters A*, 141(8):386–390, 1989. doi: 10.1016/0375-9601(89)90854-2 3

[52] D. Liu, Y. Xiong, K. Pulli, and L. Shapiro. Estimating image segmentation difficulty. In P. Perner, ed., *Machine Learning and Data Mining in Pattern Recognition*, pp. 484–495. Springer Berlin Heidelberg, 2011. 2

[53] R. Liu, H. Wang, and X. Yu. Shared-nearest-neighbor-based clustering by fast search and find of density peaks. *Information Sciences*, 450:200–226, 2018. doi: 10.1016/j.ins.2018.03.031 4

[54] G. Louppe. Bayesian optimisation with scikit-optimize. In *PyData Amsterdam*, 2017. 2, 8

[55] Q. Mao, L. Wang, S. Goodison, and Y. Sun. Dimensionality reduction via graph structure learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 10 pages, p. 765–774, 2015. doi: 10.1145/2783258.2783309 1

[56] R. M. Martins, D. B. Coimbra, R. Minghim, and A. Telea. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics*, 41:26–42, 2014. doi: 10.1016/j.cag.2014.01.006 1

[57] D. Mayo, J. Cummings, X. Lin, D. Gutfreund, B. Katz, and A. Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. In *Advances in Neural Information Processing Systems*, vol. 36, pp. 11008–11036. Curran Associates, Inc., 2023. 2

[58] D. McCarthy, M. Apidianaki, and K. Erk. Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245–275, 06 2016. doi: 10.1162/COLI_a_00247 2

[59] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approxi-

mation and projection for dimension reduction, 2020. doi: 10.48550/arXiv.1802.03426 1

[60] M. Moor, M. Horn, B. Rieck, and K. Borgwardt. Topological autoencoders. In H. D. III and A. Singh, eds., *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 7045–7054. PMLR, 13–18 Jul 2020. 2

[61] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735 1, 2

[62] A. Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. 6

[63] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720 1

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 6

[65] S. Prion and K. A. Haerling. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10(10):535–536, 2014. doi: 10.1016/j.ecns.2014.07.005 7

[66] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323 5

[67] K. Sarjana, L. Hayati, and W. Wahidaturrahmi. Mathematical modelling and verbal abilities: How they determine students' ability to solve mathematical word problems? *Beta: Jurnal Tadris Matematika*, 13(2):117–129, Nov. 2020. doi: 10.20414/betajtm.v13i2.390 6, 7

[68] F. Scheidegger, R. Istrate, G. Mariani, L. Benini, C. Bekas, and C. Malossi. Efficient image dataset classification difficulty estimation for predicting deep-learning accuracy. *The Visual Computer*, 37(6):1593–1610, 2021. doi: 10.1007/s00371-020-01922-5 2

[69] N. Sharma and S. Sharma. Optimization of t-sne by tuning perplexity for dimensionality reduction in nlp. In *International Conference on Communication and Computational Technologies*, pp. 519–528. Springer, 2023. doi: 10.1007/978-981-99-3485-0_41 2

[70] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. 2, 5

[71] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):629–638, 2016. doi: 10.1109/TVCG.2015.2467717 2

[72] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319 5

[73] J. Theiler. Estimating fractal dimension. *J. Opt. Soc. Am. A*, 7(6):1055–1073, Jun 1990. doi: 10.1364/JOSAA.7.001055 2

[74] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889–899, 2006. doi: 10.1016/j.neunet.2006.05.014 1, 2

[75] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(13):451–490, 2010. doi: 10.5555/1756006.1756019 8

[76] C. Wang, Q. Dong, X. Wang, H. Wang, and Z. Sui. Statistical dataset evaluation: Reliability, difficulty, and validity, 2022. doi: 10.48550/arXiv.2212.09272 2

[77] J. Xia, L. Huang, W. Lin, X. Zhao, J. Wu, Y. Chen, Y. Zhao, and W. Chen. Interactive visual cluster analysis by contrastive dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2022. doi: 10.1109/TVCG.2022.3209423 2

[78] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3114694 2