Journal of Data Analysis and Information Processing, 2026, 14(2), 189-214
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

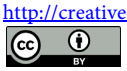
Dimensionality Reduction Evaluation for Multivariate Quantum Data

Cassio R. Cristani, Daniele Tessera

Department of Physics and Mathematics, Catholic University of Sacred Heart, Brescia, Italy
Email: cassiorodrigo.cristani@unicatt.it, tessera.daniele@unicatt.it

How to cite this paper: Cristani, C.R. and Tessera, D. (2026) Dimensionality Reduction Evaluation for Multivariate Quantum Data. *Journal of Data Analysis and Information Processing*, **14**, 189-214.
https://doi.org/10.4236/jdaip.2026.142010

Received: November 26, 2025
Accepted: April 4, 2026
Published: April 7, 2026



Abstract

Interdisciplinary research on high-dimensional quantum data faces significant challenges due to differing interpretive frameworks across physics, mathematics, and data science. Those differences in perspective create a communication gap that slows research progress and prevents deeper interdisciplinary discussions. This paper introduces a visual analytics framework to bridge these disciplinary divides through 2D embedding visualization, which facilitates collaborative discovery and incremental hypothesis testing. Our evaluation systematically compares PCA and UMAP using both absolute values—preserving quantum symmetries in femto scales—and standardized values. We assess the stability of UMAP in both data scenarios, which reveals corresponding class organizations equivalent to those produced by PCA. When applied to absolute values, UMAP is sensitive to the variable N, revealing intrinsic low-variance relationships within the data but sacrificing reproducibility. This often results in unique embeddings accompanied by noisy outliers. Conversely, scaling the data allows UMAP to eliminate fine-scale structures, ensuring more consistent structural convergence. Our framework formalizes these trade-offs, demonstrating how UMAP can function as a hybrid goal-driven tool. Additionally, it can benefit from PCA cross-validation to enhance interpretability, contributing to science and knowledge discovery in complex quantum data even when the ground truth is unknown.

Keywords

Visualization, Dimensionality Reduction, Reproducibility, Stability, Multivariate Quantum Data, Information Retrieval

1. Introduction

Analyzing high-dimensional quantum datasets poses significant challenges due to

extreme measurement ranges (from 10^{0} to 10^{-18}) and complex multivariate interactions. Interdisciplinary teams of physicists, mathematicians, and data scientists require a shared platform for interpretation, and visualization offers such a common ground. However, producing trustworthy low-dimensional embeddings demands careful consideration of *stability and reproducibility*, as distortions introduced by dimensionality reduction (DR) can compromise consistent interpretation.

Principal Component Analysis (PCA) provides a deterministic, variance-preserving reference, ensuring consistent embeddings for the same dataset [1]. In contrast, Uniform Manifold Approximation and Projection (UMAP) is stochastic, and its embeddings are sensitive to parameter choices, particularly the number of neighbors (N) [2]. Understanding how UMAP can reliably capture high-variance structure is essential to produce reproducible and scientifically valid visualizations.

Prior work has demonstrated that linear dimensional reduction methods, such as PCA, offer mathematical guarantees of convergence [3], while non-linear methods are often employed for complex, multiscale datasets in other domains, including gene expression [4]-[6] and medical diagnostics [7]. In quantum many-body physics, datasets often exhibit fast energy decay and intricate correlations, making non-linear DR methods potentially advantageous, but also raising concerns about distortions and reproducibility [8] [9].

In this work, we systematically compare PCA and UMAP on a complex multivariate quantum dataset, focusing exclusively on *high-variance structure consistency*. We introduce a rigorous reproducibility score and perform a batch of independent UMAP runs to evaluate key sources of randomness and identify parameter configurations that produce either *stable embeddings* (minor outlier fluctuations) or *fully reproducible embeddings* (no fluctuations). Using PCA as a deterministic baseline, our framework provides practical criteria for trustworthy embeddings, ensuring that visualization can serve as a reliable platform for interdisciplinary collaboration.

Our contributions are:

1. A systematic comparison of PCA and UMAP focused on high-variance structure preservation.

2. A reproducibility score to quantify stability and guide UMAP parameter selection.

3. Practical recommendations for producing stable and reproducible embeddings in high-dimensional datasets.

The following sections describe the dataset and its functional statistics in Section 2, present the dimensionality reduction methods in Section 3, explore preprocessing effects on PCA in Section 3.1, perform UMAP sensitivity analysis in Section 3.2, validate UMAP cross-embedding reproducibility against PCA in Section 5, discuss implications of stability and reproducibility in Section 6, and conclude with recommendations for trustworthy embeddings in Section 7.

2. Data Organization

Data used in this study consists of tabular datasets organized in a typical 2D structure, where sample space grows in one axis while feature space grows in another. Each dataset contains quantum observations (or samples) represented by multivariate features. The features measure energy levels of quantum particles immersed in a many-body system. Each observation is generated according to the Generalized Aubry-Andre (GAA) model [10] when fed by Entangled Spectrum interactions [11]. The features within the dataset are ordered from the highest to the lowest scales. For a better understanding of the framework, a single dataset was selected, which was the dataset with a perturbation level of $\lambda = 0.3$ in GAA model, due to the fact that it represents the most challenging scenario [9]. It has 23,277 quantum observations, each with 750 features, creating a high-dimensional structure that requires careful analysis. For intuition, our goal using DR is to reduce those 750 features of each sample into a point represented in a 2D or 3D space.

The complexity of the dataset arises from its multivariate nature, where each feature represents different energy scales at various points. Hypothetically, if two distinct features were planets, comparing them would be akin to comparing the size of Earth with that of the Milky Way galaxy—operating on vastly different scales. In contrast, comparing two observations within the same feature is more like comparing the volumes of a golf ball and a soccer ball—differing in magnitude but not in scale. Numerical methods interpret scale in different ways, potentially introducing intrinsic biases into the projections.

As a starting point for understanding the complex, high-dimensional quantum data, the dataset will be described using statistical analysis. **Figure 1** presents four functional statistics for the dataset with perturbation level of $\lambda = 0.3$ in GAA model, which is on a logarithmic scale through vertical axis. The whole feature space is shown on the horizontal axis. The gray shadow shows the range of observable values. The blue line represents the sample mean. The dashed green line represents the standard deviation. The whole sample space with all 23.277 observations is computed for each feature. It shows a rough overview of how the features are bounded within the dataset. The same dataset will be the focus of different analyses and DR scenarios in this work, allowing us to evaluate dimensionality reduction techniques from different angles within a challenging context. The goal is to retain meaningful patterns while simplifying the complexity of the data for analysis and interpretation.

The statistical functionals plot provides key insights: the minimum values (purple line) show a sharp decline, reaching an initial plateau around feature index 100. The maximum and mean values experience an even more abrupt drop next to the feature of index 250, transitioning to values between 10^{-17} and 10^{-16}, where they remain relatively constant for the remaining features. The standard deviation shows variable behavior throughout the feature range. It remains distinctly separated from the mean, the only region that keeps constant is after feature 550.

In subsequent sections, we will demonstrate how these minimum values are

captured distinctly by LDR and NLDR methods and how they might be intrinsically biasing decisions.

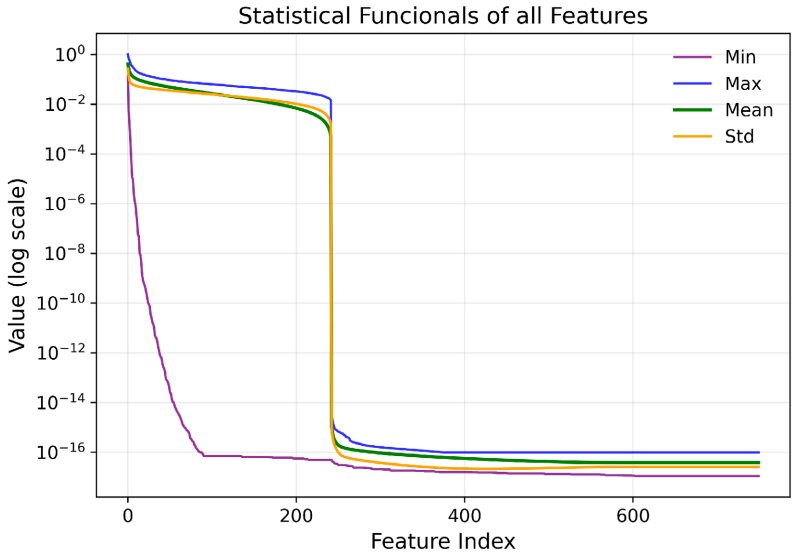


Figure 1. Statistical functionals for dataset GAA $\lambda = 0.3$.

3. Dimensional Reduction

The features represent different aspects of the quantum system's behavior and are ordered based on their scale, starting with those that have higher scales and transitioning to those with lower scales. To effectively formulate complex hypotheses and analyze this dataset, visualization of the data's behavior is essential. However, reducing the sample space would not be meaningful, and, yet, would significantly diminish the observable quantum behavior spectrum, potentially biasing results. Thus, we opt to follow a more detailed investigation on feature space through Dimensional Reduction.

DR was necessary to understand how different approaches perceive a large number of features and how preprocessing data would affect the methods, making intrinsic choices based on specific data traits.

The complexity of the datasets lies in their multivariate—or multiscale—nature, where each feature corresponds to different energy levels measured at varying points in the quantum system. We want to test LDR and NLDR isolately to perceive how the scales are interpreted by each method using original quantum data that preserves small relationships and standard scaled values. Later, we want to cross-validate both scenarios to check whether it is possible to perceive the differences by checking 2D projections.

3.1. Linear Dimensional Reduction Using PCA

Principal Component Analysis (PCA) [1] is a widely used linear dimensionality reduction technique that transforms high-dimensional data into a smaller set of orthogonal axes, known as principal components. The primary goal of PCA is to

retain the most significant features of the data by maximizing variance along the new orthogonal axes. This method is useful in quantum data analysis for identifying which features capture the most variance. However, PCA has limitations, particularly when the relationships between features exhibit non-linear behavior. This can lead to misrepresentations in its projections.

As described in [12], linear methods cannot unfold complex structures onto a visual space, making them prone to introducing considerable distortions when dealing with tangled manifolds. PCA assumes that the most important features are those carrying the highest variance. In other words, it tries to preserve the real distances in the data, greedily favouring the bigger scales. However, in quantum data, subtle interactions may happen in minimum bounds (on the order of 10^{-17}). In this case, PCA may fail to capture these relationships' small details, disregarding these critical structures in the data by considering them insignificant.

3.2. Non-Linear Dimensional Reduction Using UMAP

Uniform Manifold Approximation and Projection (UMAP) [2] is a non-linear dimensionality reduction technique designed to preserve the balance of local and global relationships within original data. UMAP operates on the principle that high-dimensional data often lies on a lower-dimensional manifold, aiming to approximate this manifold while maintaining the data's intrinsic geometry. This method constructs a graph of data points based on their local neighborhood relationships and then optimizes the geometric disposition of the points in a lower-dimensional space in successive stochastic iterations.

UMAP's key strengths include its ability to produce visualizations that explicitly represent local and global data structures. Local structure refers to the relationships between data points that are most correlated to each other, while global structure encompasses the connection of distant relationships—or non-correlated regions—within the entire dataset. By preserving both aspects, UMAP provides a more holistic view of the data compared to methods that prioritize preserving only the real distances, such as PCA.

In a simplified explanation of the UMAP algorithm it has two main steps: Initialization and Iterative Optimization. The intuition to initialization is processing all points to build a fuzzy topological matrix of symmetry. The full matrix containing the distances among all samples and all features is highly impractical for high-dimensional data. Due to the sparsity of the matrix, most positions are zero. For example, taking two different points of distinct features, it is likely that they do not correlate. Modern techniques use this characteristic to leverage performance by applying a graph-oriented approach to limit the size to the K-Nearest Neighbors (kNN), allowing visualization of high-dimensional spaces without losing sample representativeness [13].

The optimization stage employs a dual-kernel approach to bridge highly localized correlations (neighborhoods) and integrate them into a global structure, ensuring that similar neighborhoods remain close while distinct ones are pushed apart. In practice, UMAP applies attractive forces to positively correlated sam-

ples—those closely linked in the k-nearest neighbors (kNN) graph—while exerting repulsive forces on negatively correlated samples from low-correlation regions. A cross-entropy loss function then optimizes the low-dimensional embedding through stochastic gradient descent over a fixed number of iterations. Each iteration of this manifold learning process in UMAP is referred to as an epoch.

UMAP has two main parameters used to constrain the neighborhoods after the topological initialization: N and *min_dist*. The parameter N controls the minimum number of points within a given similarity radius (defined as *min_dist*), allowing practical control over the embedding. In general, when setting a low N, the algorithm detects smaller, less dense regions and reveals finer associations in the data. Increasing N for a fixed *min_dist* raises the density threshold for neighborhood formation, resulting in a visualization that emphasizes larger and more prominent groupings.

To tackle the challenges of parameter selection, robust validation methods and fair cross-validation comparisons are essential to ensure that UMAP's embeddings remain interpretable and aligned with the underlying physics. Experts in quantum physics, mathematics, and data science collaborate to identify potential weaknesses and reliability issues across different UMAP configurations. Through troubleshooting and discussions, new insights emerged on how to assess the stability and quality of the resulting embeddings.

A key challenge was that experts from different backgrounds approached data analysis with distinct perspectives, which could lead to inconsistencies in evaluations. However, implementing a map-based assessment to bridge trustworthiness paved a common ground for interpretation among the peers. Humans naturally analyze maps and develop a deeper 3D understanding from visual cues in everyday life [14]. This shared ability allowed experts to collaboratively assess UMAP embeddings, ensuring that stability and reproducibility were evaluated in an intuitive and scientifically rigorous manner.

4. Evaluation of Linear and Non-Linear Dimensionality Reduction

Since LDR and NLDR produce different coordinate systems, one based on distance and the other based on the balance of optimized manifold of local-global data's relationship, the comparison is not trivial. There is no obvious mapping function from one method to the other. Despite both methods having convergent solutions, there is no guarantee that a convergent mapping function exists between them. In case it exists, it might have a huge cost due to the complexity of each method and the large sample space, which would impose a giant combinatorial. Heuristically, some problems can offer correspondence, while others with intrincated homogeneity groups would not offer a perceptible equivalence. We are interested in offering a methodology to check whether any small evidence of correspondence between PCA and UMAP may appear, given some pre-established conditions.

In this section, we will test independently how the DR methods vary under those conditions. In the following section, we will cross-validate PCA and UMAP, trying to detect some correspondence or some gross distortion that invalidates its employment.

4.1. PCA Evaluation

Since PCA is a linear equation system, it is guaranteed that for the same data and the same number of coordinates, it will reach the same outcome embedding. This property gives it a mathematical guarantee for reproducibility, making it a good choice to be a baseline for evaluating DR.

An unavoidable challenge in quantum data manipulation is the trade-off between data precision and feature importance. The common practice is to standardize data before manipulation to guarantee that all the features have the same balance. However, doing that might give more weight to features that are not too important. The other option would be to treat the data in absolute values. As the data is tabular, it offers a way to keep the original relationship, treating the sample and features as independent, conserving representation of both magnitude and scale. Once we have this theoretical duality, we need to investigate how the methods has been influenced by that.

4.1.1. Absolute Values vs. Standardization

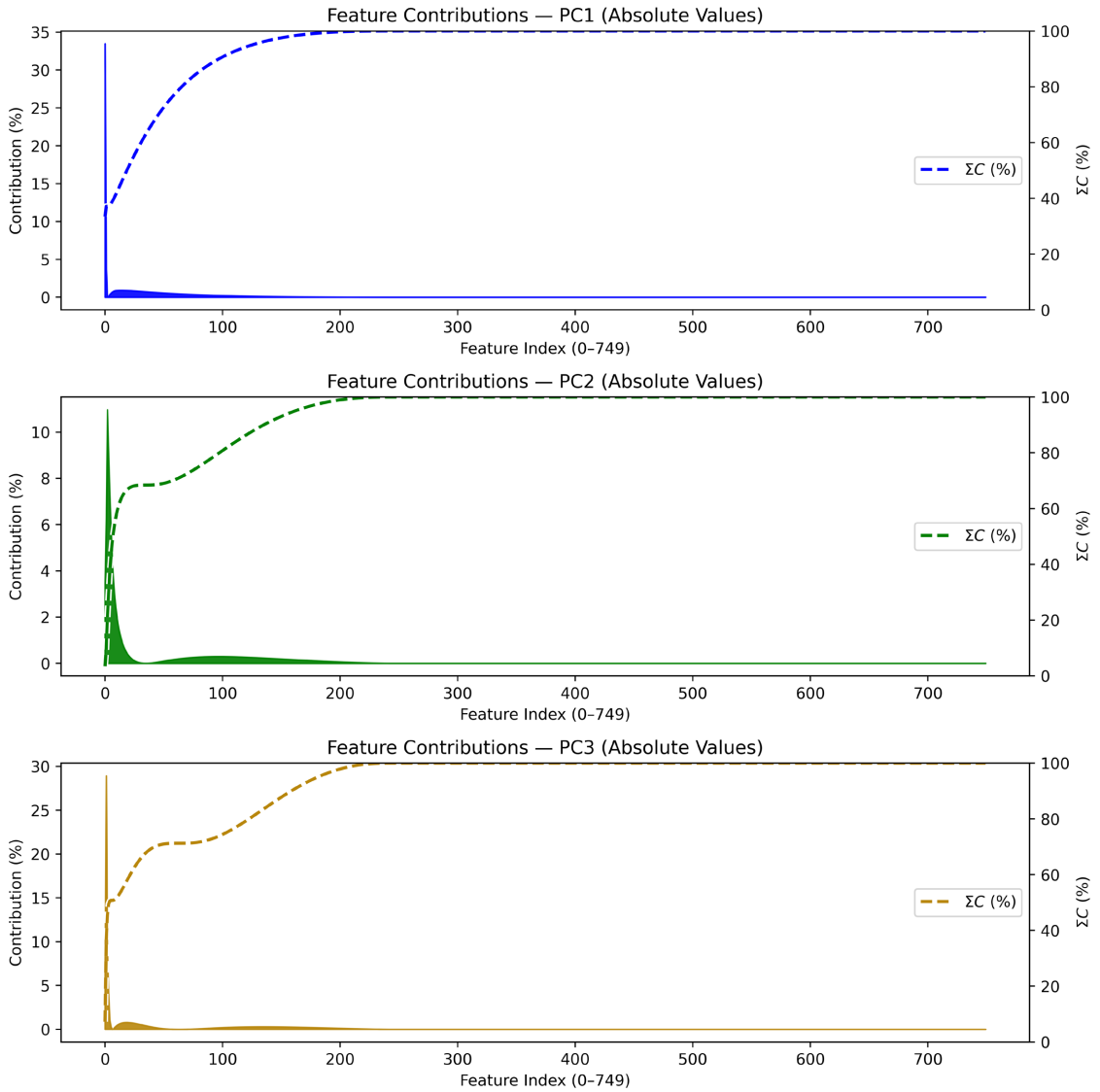
One significant challenge when applying PCA is addressing the feature space structure and the impact of non-standardized values. In datasets with features of varying scales, PCA naturally favors those with larger variance, potentially distorting the analysis. For instance, features with substantially higher value ranges may disproportionately influence the principal components, overshadowing smaller yet important features. This is particularly problematic in quantum data analysis, where subtle variations in low-variance features may trigger and be crucial for understanding underlying quantum phenomena.

To mitigate such distortions, standardizing the data before applying PCA has become common practice. Standardization ensures equal contribution from all features by removing scale effects that would otherwise allow higher-variance features to dominate the weights. This allows the method to capture a more accurate representation of the data's true structure and relationships.

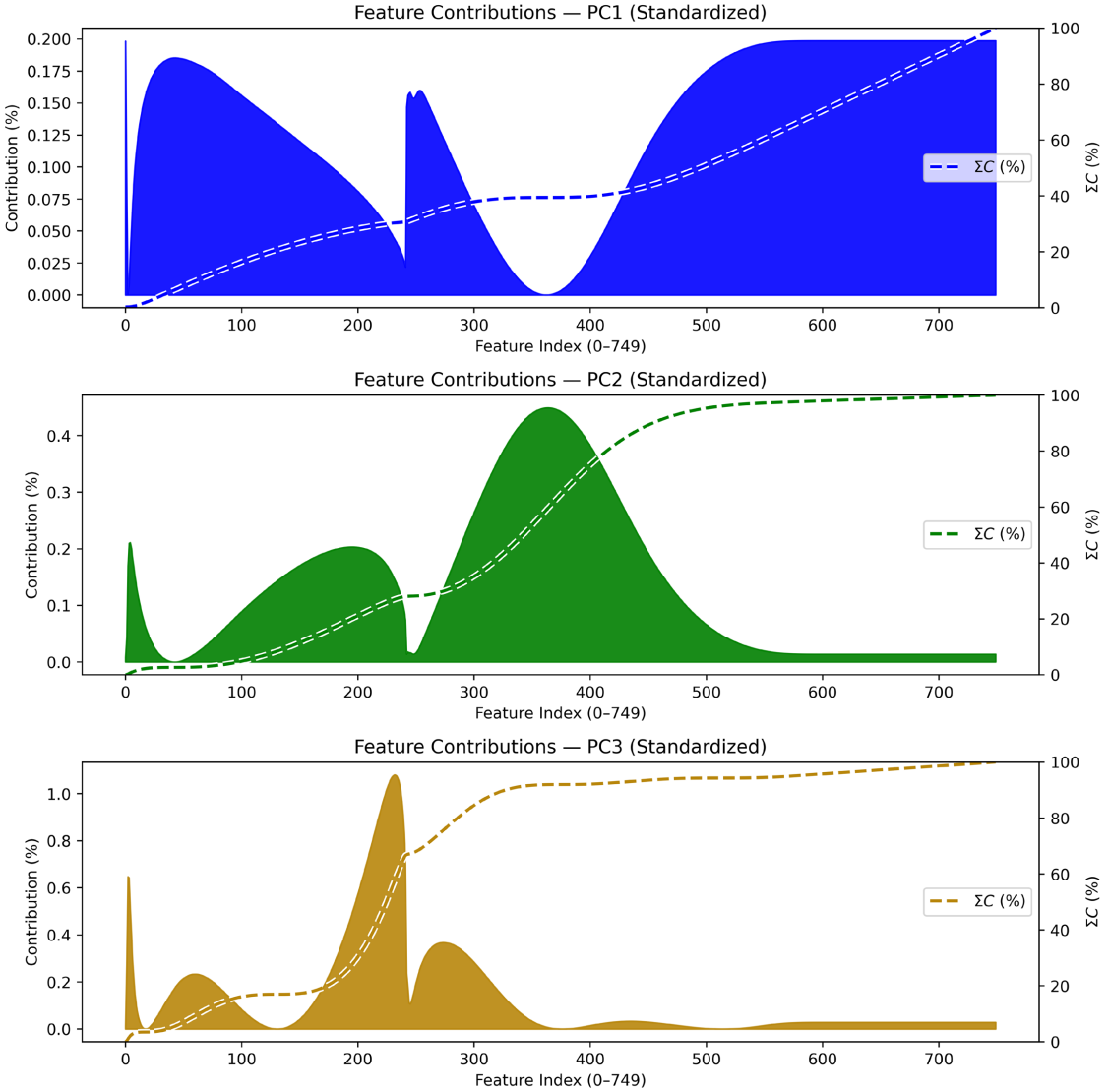
However, standardization introduces its own set of challenges, particularly in fields requiring high precision like quantum physics. The process can distort data representation, especially when dealing with features spanning multiple orders of magnitude. For example, if the original data ranges from 10^{-17} to 1, compressing it into an interval of $(-1, 1)$ with float precision may result in loss of magnitude and precision. The probability density may increase as values become closer to each other. Additionally, the choice of distribution curve to fit the data impacts the analysis; assuming normality may not always be appropriate. For these reasons, non-standardized approaches are often preferred when manipulating physics values that require maintaining original magnitude and precision.

We performed PCA using both absolute and standardized configurations to assess their impact on the principal components' index. Figure 2 illustrates the three principal components under these different configurations. Figure 2(a) shows how PCA extracts contributions from each feature to compose the three principal components using absolute values, preserving the original relationships in the data.

Notably, contributions drop to zero around the feature indexed at 250, coinciding with the abrupt scale fall of mean and maximum values observed in Figure 1. Beyond this point, PCA effectively disregards these features until the last one, which can be interpreted as a "tail cut". Specifically, the last contributing feature for all three principal components is feature 242, meaning 67.73% of the data is excluded in this configuration. Consequently, detecting quantum effects from features beyond 242 using absolute-value PCA becomes impossible, despite mathematical guarantees for capturing high-variance relationships in the included data.



(a)



(b)

Figure 2. Principal components 1, 2 and 3 using (a) absolute values in multivariate scale and (b) feature standardization with $N(0, 1)$.

The left features (higher energy levels) are more weighted due to the data's skewed distribution, which was previsible given the ordering from highest to minimum energy levels.

Figure 2(b) presents the principal components using standardized values with mean 0 and variance 1. Standardization distributes feature importance more evenly, ensuring all features contribute to the principal component index. This approach captures contributions from all features, with constant contributions appearing around feature 550 until the end for all three components, showing a similar behavior with the standard deviation presented in **Figure 1**.

4.1.2. Explained Variance in Principal Component Analysis

PCA transforms high-dimensional data into a lower-dimensional space while pre-

serving as much variance as possible. The amount of variance captured by each principal component is measured by the explained variance ratio [1], which is defined as:

$$EV_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{1}$$

where:

- EV_i is the explained variance ratio of the i-th principal component.
- λ_i is the eigenvalue associated with the i-th principal component.
- $\sum_{j=1}^{d} \lambda_j$ represents the total variance in the dataset.

To assess the structure of the data, we computed the cumulative explained variance of the first three principal components:

$$EV_{\text{sum}} = EV_1 + EV_2 + EV_3 \tag{2}$$

where EV_{sum} represents the fraction of variance retained in the first three components, indicating the extent to which the intrinsic structure of the data is preserved.

4.1.3. Results and Analysis

We applied PCA to two versions of the data: the original absolute values in multivariate scale and the standard scaled values. Figure 3 illustrates a bar plot for the explained variance of the first three components and their cumulative sum. The results emphasize the importance of preprocessing decisions, as scaling influences the interpretability of PCA results by altering the variance distribution.

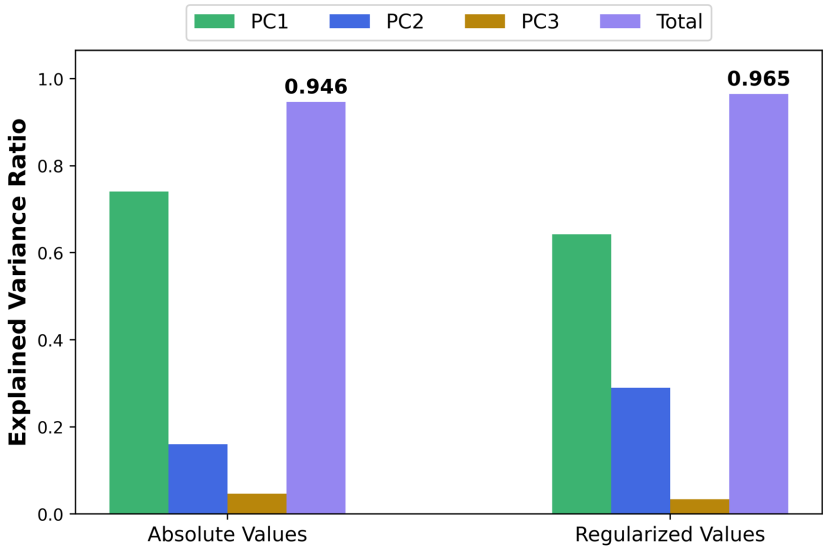


Figure 3. Explained variance ratios of the first three principal components for both datasets.

The results indicate that the standardized dataset has a higher cumulative explained variance compared to the original dataset. The three principal compo-

nents account for 96.5% of the total variance in the standardized dataset, while the original dataset reaches 94.6%. This implies that the effects of tail cutting result in nearly a 2% reduction in explained variance. The scaled dataset effectively distributes the weights across all features, contributing to the principal component index and demonstrating how standardization minimizes the dominance of certain features. This suggests that including all features in the principal components significantly influences the distribution of variance.

4.2. UMAP Evaluation

In this section, we aim to better understand the following aspects: a) Random effects; b) The impacts of preprocessing; c) The variable N as a control parameter; d) Access to reproducibility through convergence. We will keep all other parameters constant while testing the influence of the parameter N in UMAP. Our goal is to evaluate whether we can stabilize a reproducible and convergent embedding by adjusting the parameter N.

4.3. Experimental Setup

Due to UMAP's random steps and the nature of the stochastic gradient descent process, we designed an experiment comprising multiple simulations, executing them twelve times for each parameterization. This approach allows us to provide a reasonable estimation of probabilities and to visually represent the embeddings for inspection and cross-validation purposes.

For each evaluated component, we will assign a binary annotation, where 1 denotes good performance while 0 denotes bad performance. This evaluation will be repeated with varying values of the parameter N, and two data types, allowing us to assess its impact on the embedding's stability.

Through the analysis of these results, the study seeks to identify an optimal threshold for N and assess its impact on reproducibility. It further investigates whether specific values of N lead to systematic instability in visual interpretations. Additionally, a comparative evaluation of raw and standardized data provides insights into how preprocessing decisions influence the robustness of the embeddings.

Increasing N forces the algorithm to construct neighborhoods graph from denser regions of the data; as each neighborhood contains more samples, the number of distinct regions decreases. In the limit, a large enough N value drives the embedding toward a single connected neighborhood containing all samples. To balance representational fidelity with computational cost, we selected the smallest N that still yields a connected region and explored values up to $N = 25$. This upper bound was sufficient to observe the transition from unstable to reproducible behavior, and varying N in steps of five allows clear identification of the tipping points.

4.4. Reproducibility and Stability Trigger Scores

The assessment framework is structured around three fundamental indicators to

infer reproducibility: overall structure consistency, the presence of outliers and substructures, and the continuity of local relationships.

Overall structure consistency is examined to determine whether the general geometric arrangement of the embedded space remains stable across multiple executions. A reproducible embedding should preserve the relative positioning of major patterns and maintain coherent topological features. The reliability of the embedding can be called into question if the formation of substructures varies significantly between different runs. This variation may lead to unpredictable shifts in their relative positions or distortions in their spatial organization. Such instability indicates a sensitivity to initialization or parameter changes, which could undermine the method's usefulness in exploratory data analysis.

The presence of outliers and substructures offers important insights into the reproducibility of the embedding. Ideally, peripheral points that appear as outliers should consistently occupy the same locations across multiple runs. Additionally, substructures—whether they form distinct patterns or reveal finer organizational details—should emerge reliably. If outliers appear inconsistently or if substructures do not materialize consistently, it indicates that the embedding process is introducing artifacts that obscure the true relationships within the data, highlighting a susceptibility to local minima. This instability undermines reproducibility, limiting our ability to derive consistent insights from the low-dimensional representation.

These insights can be non-reproducible and still be meaningful since present associations of low-variance features and tail effects, which we will discuss further with the interpretability paradox.

A third critical reproducibility component is the continuity of points. The aim is to see a regular pattern in terms of continuity across several simulations. Disruptions in continuity will give a low score. Such discontinuities indicate a failure to maintain the intrinsic organization of the data, meaning more susceptibility to fall into local minima and, impose interpretation barriers.

To quantify reproducibility, we introduce a formal criterion based on three fundamental control components: global structure consistency (S), presence of outliers and substructures (O), and the continuity of local relationships (C). Each component is assigned a binary score, where 1 denotes likely reproducibility and 0 denotes unlikely reproducibility, based on a systematic visual evaluation of the embeddings. The complete reproducibility score R of an embedding is then determined by the logic product:

$$R = S \cdot O \cdot C \tag{3}$$

This formulation ensures that an embedding is considered fully reproducible ($R = 1$) only if all three conditions are met simultaneously. If any of the components indicate instability (that is, if S , or O , or C is 0), then $R = 0$, meaning that the embedding has failed to achieve reproducibility under the given configuration.

To address varying levels of outlier influence, we can create a discretization of

component O in Equation (3), which provides greater flexibility in handling their presence. On the other hand, excluding outliers can lead to a useful stability trigger score that focuses exclusively on the overall structure's stability and continuity. In this scenario, outliers are not taken into account. This approach streamlines the analysis by disregarding outliers, resulting in a simplified logic trigger for stability:

$$St = S \cdot C \tag{4}$$

Using this approach, we can categorize a configuration for more restrictive scenarios that prioritize precision and outlier detection, as outlined in Equation (3). Additionally, stable embeddings can serve as a reliable approximation of stability while minimizing computational effort, as indicated in Equation (4).

4.5. Sensitive Analysis of Number of Neighbors

The sensitivity analysis of N was conducted to assess the reproducibility of non-linear dimensionality reduction embeddings, considering two data representations: absolute values, which preserve the original measurement relationships, and standardized data, which regularizes scale variations. Given the high volume of embeddings analyzed, the results of 12 executions were condensed for three critical values of N (5, 10, and 15) and are summarized in Table 1, while the complete set of embeddings is included in Figure 4.

Table 1 shows annotations for inputs (S, O and C) and output trigger scores (R and St) computed based on embeddings presented on 10. Also, the outcomes for trigger scores of reproducibility and stability are shown for original raw data and scaled data.

Table 1. Annotated binary inputs and outputs for UMAP across N: Structure (S), Outliers (O), Continuity (C), and computed metrics for Reproducibility (R) and Stability (St) using raw data and scaled data.

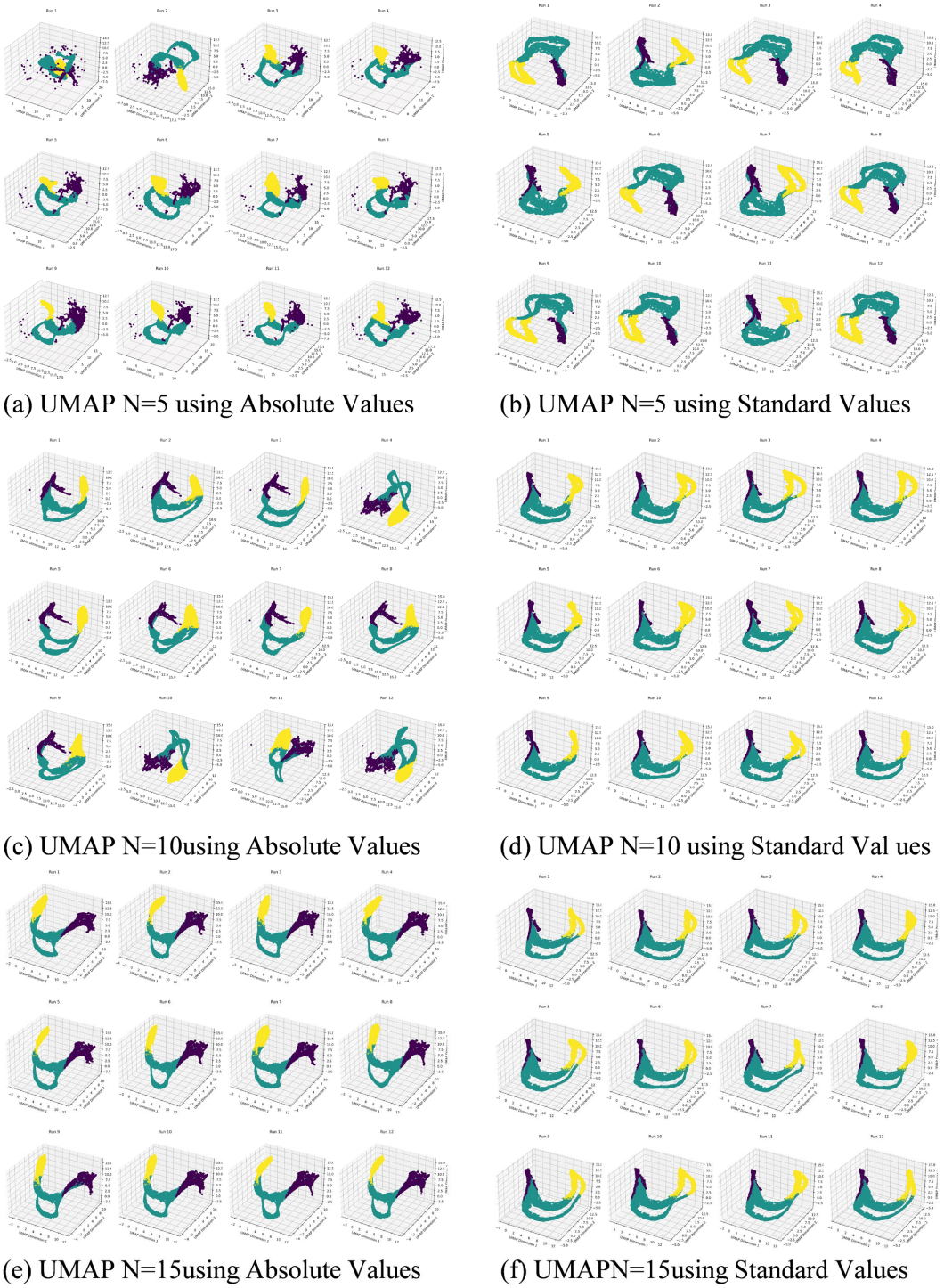
	RAW DATA					SCALED DATA				
N	S	O	C	R	St	S	O	C	R	St
5	1	0	0	0	0	1	0	1	0	1
10	1	0	1	0	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	1	1	1	1

In the supplementary material, you will find the complete set of embeddings for each configuration listed in Table 1, specifically for N values of 5, 10, and 15. However, due to space limitations, the embeddings for the convergent values of N = 20 and 25 are included only in the repository for verification purposes.

Visual Embeddings Validation

Figure 4 shows a set of 12 executions for each configuration. The left column runs

in absolute values while the right column presents executions based on standard scaled values. We will evaluate the columns individually and then draw conclusions regarding the effects of preprocessing.



(a) UMAP N=5 using Absolute Values (b) UMAP N=5 using Standard Values

(c) UMAP N=10using Absolute Values (d) UMAP N=10 using Standard Val ues

(e) UMAP N=15using Absolute Values (f) UMAPN=15using Standard Values

Figure 4. Multiple simulations for UMAP used to compose **Table 1**, with absolute values (left) and standard values (right) at different N values. Each sub-figure set corresponds to: (a)-(b) $N = 5$; (c)-(d) $N = 10$; (e)-(f) $N = 15$.

The number of observations N increases as we move down the figure. For the absolute values at $N = 5$, which includes 12 executions in the top left set, we observe noisy embeddings in all runs. Significant gaps of discontinuity appear in runs 1, 2, 4, 5, 6, 7, 8, and 9. Despite this noise, the overall structure remains consistent with two distinct branches, with the exception of runs 2 and 12, which are difficult to distinguish due to a high number of scattered points.

When examining $N = 10$, the overall structure is much more stable, with no discontinuities. Although outliers are still present, they are fewer in number and more consistently located. Since our reproducibility score is defined as binary, we pragmatically focus on the worst-case scenarios. The presence of outliers—despite the stability observed—results in a reproducibility score of $R = 0$. However, we can conclude that the embeddings are stable at $N = 10$.

At $N = 15$, outliers have fully integrated into the structure, which shows no discontinuities and remains consistent across different executions. The only noticeable difference is the rotation angle of the embedding, which does not affect the score; we can interactively navigate the embedding in any direction. Given that $S = O = C = 1$, our reproducibility score is $R = 1$, which can be visually validated by the convergence of the embeddings.

Now, turning to the right column, which represents standard scaled data, the overall structure and continuity remain constant across all N values. For $N = 5$ in the right column, we detected a few stable outliers in run 7, mirroring the observation at $N = 10$ in the previous column. Based on our definition of reproducibility, this leads to the presence of outliers ($O = 0$), resulting in $R = 0$. However, the overall stability score is 1 due to the few outliers combined with continuity and a similar overall structure.

At $N = 5$, we observe some minor differences caused by the flexibility of small neighborhoods. This allows many subsets to be closely embedded, increasing susceptibility to local minima. Nonetheless, our evaluation indicates that the overall structure remains consistent across runs, which can be visually confirmed.

Comparing raw and scaled data for the same N, we find that UMAP captures a similar pattern in the data, revealing two branches that are not easily perceived by the eye using PCA. Interestingly, this pattern is rounded in the raw data while appearing rectangular in the scaled data. This indicates that, regardless of whether standard scaling is used, we can achieve a comparable overall structure.

Our scores show that full reproducibility is achieved using $N \geq 15$ for absolute values and $N \geq 10$ for standardized data. Furthermore, we find that structural stability is not sensitive to varying N when data is scaled, providing a solid foundation for validating a property that can accelerate the discovery of general structures. Using absolute values, we also achieve structural stability with $N \geq 10$, which can save computational resources in less demanding tasks.

One important conclusion from this set of experiments is that we can quickly achieve a stable embedding of UMAP by using standardization to reveal the overall structure, thereby reducing dependency on N analysis. This finding is valuable

for several lines of research where fixed conditions need to be scrutinized. The next section will explore one such application, focusing on comparisons in class organization, with the right column.

5. Visual Cross-Validation between PCA and UMAP for Class Organization

Visual inspection of embeddings can be conducted under two conditions: with hidden labels and with Active Labels. The first condition, where labels are hidden, represents a more realistic and challenging scenario. It closely resembles how real-world datasets are typically encountered, without pre-established classifications. This setting is particularly relevant when the ground truth is inaccessible or when conducting blind evaluations that do not rely on class estimators. However, it has limited interpretability, especially in complex datasets.

The second condition serves as a visual aid by introducing active labels, which enhance the interpretability of sample distributions across different embeddings. Even when pseudo-labels are employed—providing no direct categorical information—they act as a consistent reference, allowing for a structured comparison between PCA and UMAP.

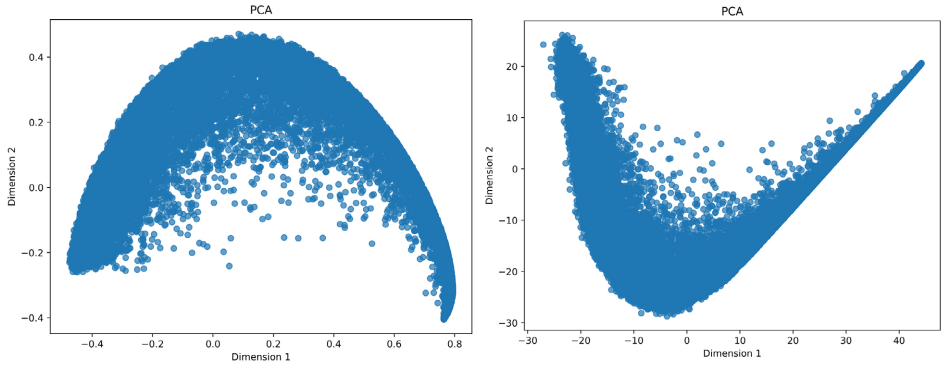
To offer a more realistic and comprehensive approach to visual analysis, we invite readers to try out a test provided with all relevant materials, datasets, Active Pseudo-Labels, and an interactive visualization tool included in the supplementary material. This setup enables readers to fully explore the capabilities of the 2D/3D embeddings and quickly verify the reported results in a real-world scenario. This ensures a deeper understanding of data structures, variations in density, and potential outliers.

5.1. Hidden Label Condition

In this analysis, we visually inspect 2D scatter plots of the absolute values and scaled values. transformed dataset for PCA and UMAP using hidden label condition, which means only the points distribution is available without any metadata consult. This evaluation focuses on checking the visual silhouette of the data and gives the first idea of the structure in low-dimensional space. We also want to check if the same components evaluated in the previous section might be perceived directly on 2D embeddings under a hidden label condition.

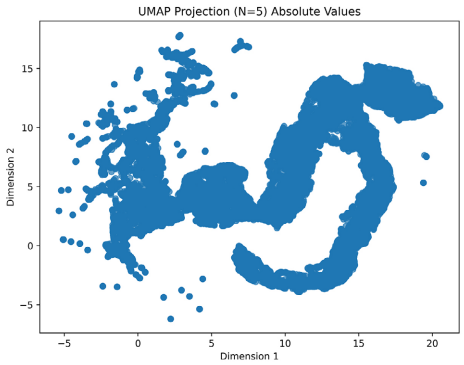
Figure 5 presents embeddings of absolute values in the left column and scaled values in the right column. The first row illustrates PCA results, while subsequent rows show UMAP results with neighborhood parameters $N = 5$, 10, and 15.

Under the hidden label condition, comparing UMAP and PCA directly is challenging due to their differing coordinate systems. Even comparing PCA results across different preprocessing methods is not feasible. However, UMAP's convergence allows for meaningful comparisons across its parameter configurations. Through visual assessment of Figure 5, we derive consistent individual scores for structural consistency (S), outlier presence (O), and continuity (C).

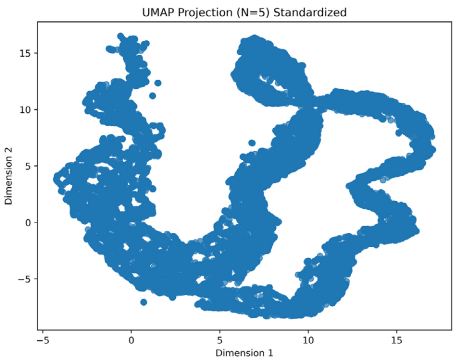


(a) PCA Absolute Values

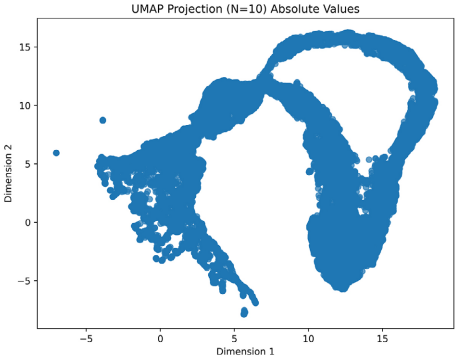
(b) PCA Standard Scaled Values



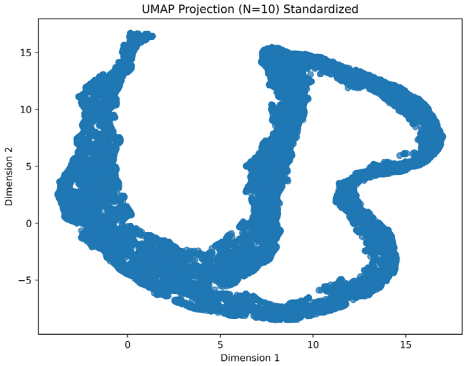
(c) UMAP N=5 Absolute Values



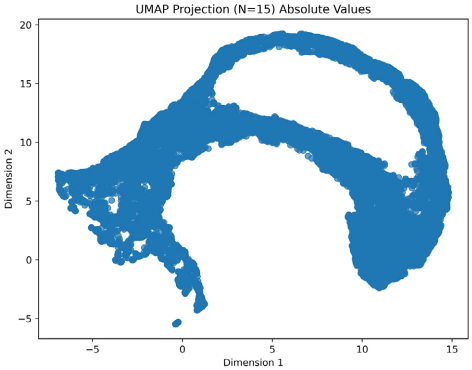
(d) UMAP N=5 Standard Scaled Values



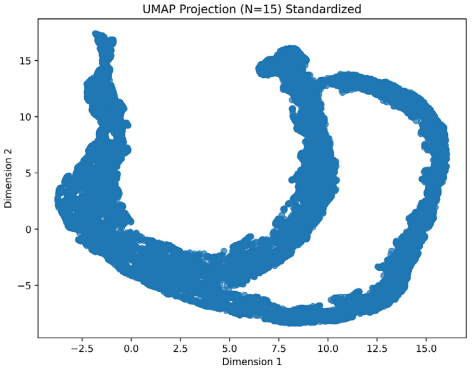
(e) UMAP N=10 Absolute Values



(f) UMAP N=10 Standard Scaled Values



(g) UMAP N=15 Absolute Values



(h) UMAP N=15 Standard Values

Figure 5. Hidden label UMAP embeddings using absolute values (left) and standard scaled values (right).

In complex visualization scenarios involving high-dimensional data and densely packed points, relying solely on 2D representations can lead to unreliable results for detailed inspections. Introducing a third axis enhances visual analysis by adding depth, allowing for the discovery of patterns that might be obscured in 2D projections due to overlapping regions. However, even 3D visualizations can be problematic if they use fixed angles or static viewpoints, as these can also result in overlapping regions and hinder the identification of subtle relationships within the data.

A significant advantage of visual inspection is the ability to explore embeddings dynamically through rotation, zooming, and interactive navigation. These features enable a more precise analysis of complex structures, highlighting subtle relationships that might remain hidden in 2D representation or even in a static 3D visualization.

Unfortunately, the full potential of these interactive tools cannot be effectively communicated through written descriptions or static images. An interactive experience can be done using instructions within supplementary material.

5.2. Active Pseudo-Label Condition

Dimension-reduction assessment and cluster assessment address two complementary questions. To evaluate a clustering, we freeze the low-dimensional coordinates produced by a single embedding method and ask how well competing label assignments reveal the resulting structure. Conversely, to evaluate an embedding itself, we freeze the true labels and ask how faithfully different coordinate systems preserve the class structure, thereby quantifying the distortion introduced by each dimensionality-reduction technique.

In this work, our primary objective is to compare two different coordinate systems for visualizing complex multivariate quantum data. Because true labels are not available by the nature of the problem, we do not discuss how labels are generated. A complete methodology and discussion on unsupervised label estimation—compared with physically motivated pseudo-labels—can be found in [15]. For the purposes of this study, we adopt a single reference set of labels produced using K-Means, solely as a visual aid.

We also avoid complex quantum-physics interpretations at this stage of the methodology evaluation, since our goal is to characterize the distortion introduced by different coordinate systems. If we gather sufficient evidence that the coordinate systems preserve class ordering and avoid mixing samples, future work may substitute these generic labels with physically grounded, hypothesis-oriented labels to enable a deeper investigation of specific quantum behaviors.

The Active Pseudo Labels (APS) therminology in this work means that we are using the same labels in both coordinate systems in a 1 to 1 mapping. Each sample has the same label that was generated a priori for both visualizations. By fixing labels, we can more easily interpret how the sample groups are moving between two different coordinate systems. APL can also be replaced with new or improved estimators that may contain relevant data, offering a strong fit for incremental

knowledge frameworks. This substitution can be beneficial for observations and for refining hypotheses toward the current unknown ground truth.

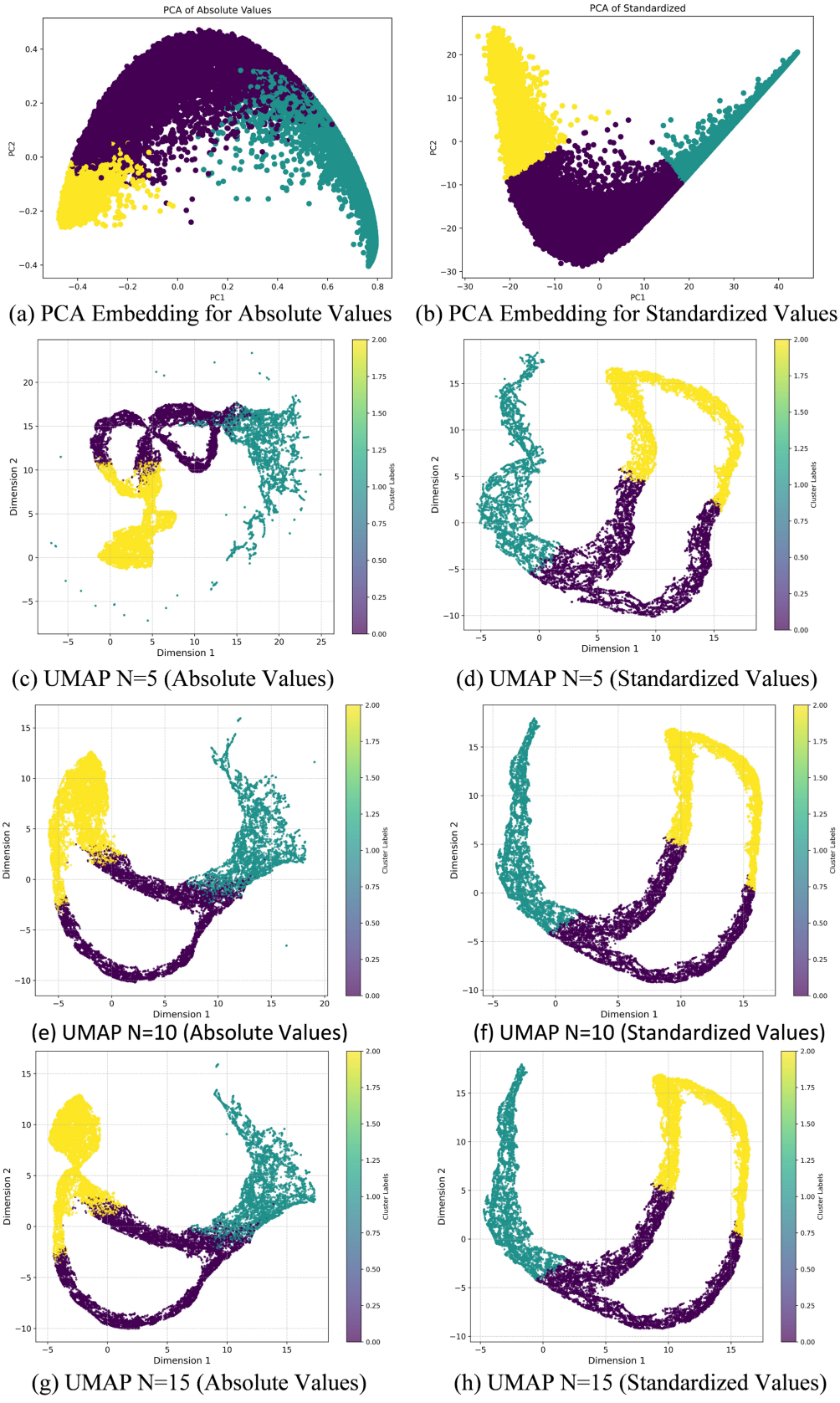
APL provides insights into the structural properties of embeddings. From a scientific standpoint, the validity of hypotheses often relies on the quality and reliability of the labels used. However, for visual inspection, APL can facilitate concrete inferences about reproducibility and group organization. To better understand the differences between PCA and UMAP, we examine their performance using 2D embeddings. By using this approach, we can gain insight into how effectively each method captures structural consistency, identifies outliers, and maintains continuity.

In this study, we will not perform cluster analysis, but it is essential to justify our comparison methodology. Generally, there are two types of mixture—also named class organization or labeling—found in any dataset: heterogeneous, where each group is well-separated, and homogeneous, where the data points are mixed, and no group pattern can be distinguished.

To determine the type of organization present in our data, we will compare UMAP with PCA and assess the similarities. Our evaluation will focus solely on class organization, which means we aim to understand whether the class structure remains stable, regardless of the coordinate system. We will not perform a strict one-to-one mapping; instead, we will assume that the intra-class organization might vary. In other words, our class evaluation relaxes the requirement for the commutativity of samples within the same group. We want to observe that classes that are distant in PCA are also distant in UMAP, and classes that are neighbors in PCA should also be neighbors in UMAP.

Figure 6 illustrates the cross-validation results for PCA and UMAP. 3D embeddings are produced using absolute for left column and scaled values for the right column. As discussed in the previous section, different preprocessing steps lead to different coordinate systems. In Section 4, it was shown that the projection in Figure 2(a) trims the tail by considering only the features up to feature 242, beyond which no contribution to the principal components is observed. In contrast, Figure 2(b) displays the projection using a standard scaler, where all features contribute almost equally to the principal components, utilizing the entirety of the data. Despite the differences between the two versions, an important conclusion can be drawn when using a fixed Active Pseudo-Label: there is consistency in class preservation. In both cases, the yellow and cyan classes remain in opposite positions, indicating that while the relative distances have changed, the classes themselves are not mixed. Which serves as a foundation for future experiments, where we will replace the given APLs for actual estimators containing quantum traits information.

Looking for UMAP embeddings, we observed that UMAP's stable versions (Figures 6(d)-(h)) corresponded well with PCA in terms of class organization, as they maintained the heterogeneity of the classes. We can see that the yellow and cyan classes respect the order informed by PCA since they were positioned at opposite ends.



(a) PCA Embedding for Absolute Values (b) PCA Embedding for Standardized Values

(c) UMAP N=5 (Absolute Values) (d) UMAP N=5 (Standardized Values)

(e) UMAP N=10 (Absolute Values) (f) UMAP N=10 (Standardized Values)

(g) UMAP N=15 (Absolute Values) (h) UMAP N=15 (Standardized Values)

Figure 6. PCA and UMAP embeddings using absolute values (left) and standard scaled values (right).

In Figure 6(c), the control scores reveal concerns with the embedding, as both scores $R = 0$ and $St = 0$ indicate issues. This embedding is particularly complex for visual inspection. By using the reproducibility score to ensure interpretability, we can visually assess the lack of clarity in this example.

Focusing on the lower right substructure formed by the yellow and purple classes, we can roughly identify where the purple segment should be integrated into the main structure. However, determining the placement of the yellow segment is much more complex. In this case, a new execution is required—one that achieves at least a stability score of one—to ensure a minimally viable embedding.

Our experiments demonstrate that we are achieving a low continuity score $C = 0$ in 92% of executions. indicating that only one out of approximately twelve executions yields a continuous embedding. This underscores the necessity of multiple executions to obtain reliable results and highlights the volatility in embedding stability.

This example effectively illustrates the practical utility of the control metrics introduced in this study. These metrics prove to be highly reliable in predicting scenarios where embedding reproducibility and stability fall below acceptable thresholds, thereby serving as critical indicators for assessing the quality of dimensionality reduction outcomes.

Importantly, our findings provide initial evidence that embeddings deemed reproducible through our metrics exhibit class correspondence with PCA, suggesting that UMAP does not introduce randomness that disrupts the inherent class order. Furthermore, we observe that stable embeddings, as identified by our metrics, are also capable of maintaining class organization consistent with PCA. This correspondence not only validates the effectiveness of our control metrics but also implies that leveraging these metrics can streamline the analysis process, making it more efficient and fluid.

6. Discussion

6.1. Main Findings and Interpretability

To better understand the interplay between variance, outliers, structural stability, reproducibility, and interpretability in dimensionality reduction, we propose a conceptual model presented in Figure 7. This model illustrates how the characteristics of principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) correspond each other as the variance is adjusted.

PCA provides a strong foundation for the high variance analysis of datasets but falls short in detailed visual inspection due to its tendency to produce high-density embeddings. It does not offer a direct way to visualize low-variance features without preprocessing, which may introduce biases in the underlying physics of the data.

In contrast, UMAP offers significant visual enhancements by projecting data into a new coordinate system that reduces embedding density while preserving and explicitly revealing both local and global relationships. This makes UMAP

particularly insightful for fields lacking ground truth, as it provides a path for incremental hypothesis validation and knowledge discovery. UMAP introduces a control variable N, which can serve as a surrogate for variance levels when the original data relationships are preserved. However, when data is preprocessed with standard scaling, UMAP loses sensitivity to the N parameter. This leads to quicker convergence for structure generalization but at the cost of cutting off small variance edges that cannot be perceived or reconstructed.

As variance decreases, several notable effects emerge:

• The number of outliers and substructures increases due to the formation of more neighborhoods given the flexibility of constraints;

• UMAP present a control variable N, which can surrogate variance levels when the original relationship is preserved on the data.

• Structural continuity diminishes due to the increasing number of neighborhood regions, which complicates the composition of the final 3D embedding;

• Reproducibility and stability degrade, making it more likely to reach unique embeddings and increasing the challenge for visual analysis;

• Interpretability through the correspondence between linear (LDR) and nonlinear (NLDR) dimensionality reduction methods becomes difficult once the mathematical guarantees of linear methods are lost

At this point, UMAP continues to capture low-variance relationships that are inherently present in the data. The implication is that when we lack the capacity to interpret these relationships, we may reject them due to insufficient understanding or waste time trying to stabilize embeddings against the odds. This shift highlights a critical trade-off: the extent of risk we are willing to accept in our analyses.

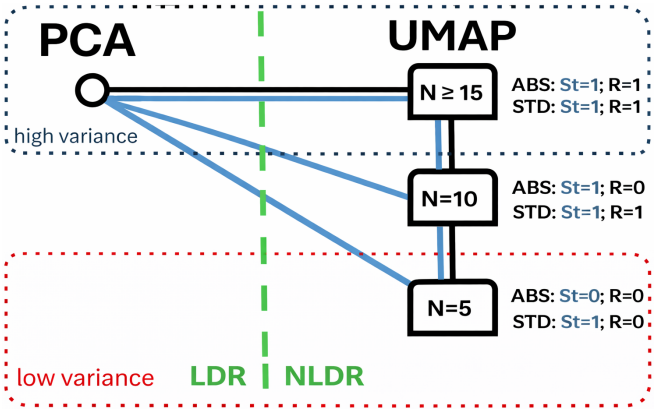


Figure 7. Association among tested configurations for PCA and UMAP.

6.2. Reproducibility vs. Discovery: A Necessary Trade-Off?

A crucial question that arises is: Are we prepared to accept reduced reproducibility to figure out how to interprete unique embedding, or should we explore alternative approaches that ensure stability as a requisite?

The challenge is particularly relevant for low-variance relationships, where

these interactions exist as factual elements of the data, yet they often elude human intuition, numerical methods, and conventional experimental methodologies. If scientific methodology inquiries are overly fixated on high-variance patterns, we risk discarding meaningful but subtle interactions.

The issue may lie not in the validity of low-variance relationships but in the limitations of our current interpretative frameworks. Are we inadvertently truncating meaningful connections due to biases in experimental design that prioritize direct cause-effect relationships? Instead of dismissing small-variance interactions as noise—because outliers and structure does not stabilize, which is a by-product of not enforcing reproducibility in method operation. Perhaps, we should consider retraining our models—and perhaps even our scientific intuition—to capture these subtleties' tail effects. This shift could allow us to uncover richer, more nuanced patterns in complex systems, broadening the horizons of data-driven discovery.

7. Conclusions

In this work, we present a visual inspection framework designed to facilitate the interdisciplinary analysis of complex high-dimensional quantum data. By integrating perspectives from quantum physics, mathematics, and data science, the framework provides a structured approach to exploratory analysis, leveraging 3D embedding maps as a shared ground for hypothesis formulation, testing, and accelerating scientific discovery.

We applied this framework to evaluate a quantum dataset using dimensionality reduction (DR) techniques, comparing the linear approach of PCA with the non-linear method of UMAP. Despite its deterministic nature, PCA is highly sensitive to preprocessing and requires data standardization to ensure fair weighting of independent features across different scales. When applied to absolute values, PCA inherently prioritizes high-variance features while treating small-scaled values as noise, leading to the elimination of approximately 67% of the dataset and limiting its ability to reveal fine-grained structures. This bias also results in the misrepresentation of minority features, effectively obscuring fat-tail events. Standardization mitigates this issue, increasing the explainable variance from 94.5% to 96.6% and preserving the representation of tail effects that were otherwise lost.

For UMAP, we conducted a sensitivity analysis of parameter N to assess its interaction with data preprocessing. When applied to standardized values, UMAP showed little sensitivity to N, consistently converging to a similar structure across all configurations. While minor variations were observed for small $N = 5$, the overall shape remained stable. We attribute this stability to the standardization process, which smooths low variances and introduces a bias toward high-variance features, preventing UMAP from capturing subtle similarities. This, in turn, allows for rapid convergence and reliable interpretability with reduced computational effort.

When applying UMAP to absolute values—preserving physical symmetries in

magnitude and scale—we found that it is sensitive to parameter selection and does not always converge to the same structure. This behavior presents a challenge for fair evaluation, as UMAP exhibits a dual nature: it can either enhance fine-detail resolution, producing less reproducible embeddings that offer new interpretative insights, or generate stable embeddings suitable for controlled scientific analysis, aligning with the structure observed in standardized data. A key advancement of our proposed metrics is their ability to define these limits and mitigate variance by tuning N, allowing UMAP to fulfill both roles without requiring data rescaling.

By aggregating our proposed cross-validation scores, we identified parameter regimes that reinforce trustworthiness in structural stability, minimize stochastic artifacts, and enhance the reliability of unsupervised visualization techniques for high-dimensional quantum data. UMAP demonstrated structural stability for $N \geq 10$ and reproducibility for $N \geq 15$, yielding embeddings comparable to those obtained with standardized data.

A cross-validation analysis using Active Pseudo-Labels (APL) between PCA and UMAP confirmed that UMAP preserves class organization relative to PCA, regardless of whether the data is in its original or scaled form. This suggests that UMAP does not introduce gross distortions to the intrinsic quantum structures in the tested scenario. Future work will refine this approach by replacing generic pseudo-labels with physics-informed labels to draw more precise conclusions. Nonetheless, our findings establish an important baseline by demonstrating that UMAP maintains consistency with PCA's class organization while offering greater flexibility.

The interpretability in low-variance is an open conflict because if each execution produces a different embedding, the solution becomes less scientifically reliable, as there is no guarantee that a given structure can be consistently reached again. One way to address this issue is by developing a mathematical approach that ensures robustness for low-variance, akin to PCA's high-variance stability, though this can be complex and computationally expensive.

An alternative strategy is a hierarchical decomposition of the problem, treating variance levels as incremental stages in a structured exploration. This approach begins with stable, high-variance structures—such as those produced by UMAP with scaled data—and progressively incorporates low-variance substructures. By formalizing this transition, researchers can balance discovery and comprehension while maintaining scientific integrity.

UMAP emerges as a versatile bridge between structured analysis and exploratory discovery, requiring minimal preprocessing while capturing both the stable class organization observed in PCA and the intricate low-variance symmetries that PCA overlooks. These subtle structures, though less reproducible, may be essential for uncovering rare fat-tail events in quantum systems.

Based on these findings, we conclude that PCA should always be applied with preprocessing to ensure fair representation of all features. Conversely, UMAP with absolute values demonstrates remarkable versatility, offering both stable high-

variance representations and the ability to explore low-variance structures, with control solely through tuning N. This flexibility makes UMAP a powerful tool for balancing reproducibility with exploratory analysis in quantum data visualization.

While this study is based on a single dataset from the GAA model, the proposed framework and the insights gained from UMAP's behavior and parameterization should be broadly applicable to other high-dimensional quantum datasets using different lambda potentials. Future work will extend this analysis by incorporating multiple datasets, providing a more comprehensive protocol for assessing the generalizability of the findings. A systematic exploration of the framework's performance across various types of data will further solidify its potential as a versatile tool for visualizing and analyzing complex systems in quantum mechanics and beyond.

Acknowledgements

I am deeply grateful to Professor Yi-Ting Hsu for generously providing the quantum data from her experiment, which forms the foundation of this research. I also wish to thank Professor Fausto Borgonovi for his mentorship and for the many insightful discussions on quantum mechanics, as well as Colin Beveridge for his valuable input on the quantum model, which helped clarify and strengthen the trustworthiness of the ideas presented here. Their contributions were indispensable and greatly enriched both the quality and the scope of this study.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Jolliffe, I. (2002) Principal Component Analysis. 2nd Edition, Springer-Verlag.

[2] McInnes, L., Healy, J. and Melville, J. (2018) UMAP: Uniform Manifold Ap Proximation and Projection for Dimension Reduction.

[3] Cunningham, J.P. and Ghahramani, Z. (2015) Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *The Journal of Machine Learning Research*, **16**, 2859-2900.

[4] Venna, J. and Kaski, S. (2007) Non-Linear Dimensionality Reduction as Informational Retrieval. *International Conference on Artificial Intelligence and Statistics*. *PMLR*, San Juan, 21-24 March 2007, 572-579.

[5] Chari, T. and Pachter, L. (2023) The Specious Art of Single-Cell Genomics. *PLOS Computational Biology*, **19**, e1011288. https://doi.org/10.1371/journal.pcbi.1011288

[6] Sun, Y. and Zhang, F. (2022) Optimization of Classification Results on Gene Expression Datasets Using Dimensionality Reduction. *CAIBDA* 2022; 2*nd International Conference on Artificial Intelligence, Big Data and Algorithms*. *VDE*, Nanjing, 17-19 June 2022, 1-11.

[7] Alhassan, A.M. and Wan Zainon, W.M.N. (2021) Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis. *IEEE Access*, **9**, 87310-87317. https://doi.org/10.1109/access.2021.3088613

[8] Li, X., Ganeshan, S., Pixley, J.H. and Das Sarma, S. (2015) Many-Body Localization and Quantum Nonergodicity in a Model with a Single-Particle Mobility Edge. *Physical Review Letters*, **115**, Article ID: 186601. https://doi.org/10.1103/physrevlett.115.186601

[9] Hsu, Y., Li, X., Deng, D. and Das Sarma, S. (2018) Machine Learning Many-Body Localization: Search for the Elusive Nonergodic Metal. *Physical Review Letters*, **121**, Article ID: 245701. https://doi.org/10.1103/physrevlett.121.245701

[10] Ganeshan, S., Pixley, J.H. and Das Sarma, S. (2015) Nearest Neighbor Tight Binding Models with an Exact Mobility Edge in One Dimension. *Physical Review Letters*, **114**, Article ID: 146601. https://doi.org/10.1103/physrevlett.114.146601

[11] Li, H. and Haldane, F.D.M. (2008) Entanglement Spectrum as a Generalization of Entanglement Entropy: Identification of Topological Order in Non-Abelian Fractional Quantum Hall Effect States. *Physical Review Letters*, **101**, Article ID: 010504. https://doi.org/10.1103/physrevlett.101.010504

[12] Nonato, L.G. and Aupetit, M. (2019) Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment. *IEEE Transactions on Visualization and Computer Graphics*, **25**, 2650-2673. https://doi.org/10.1109/tvcg.2018.2846735

[13] Bohm, J., Berens, J. and Kobak, J. (2022) Attraction-Repulsion Spectrum in Neighbor Embeddings. *Journal of Machine Learning Research*, **23**, 1-32.

[14] O'Keefe, J. and Dostrovsky, J. (1971) The Hippocampus as a Spatial Map. Preliminary Evidence from Unit Activity in the Freely-Moving Rat. *Brain Research*, **34**, 171-175. https://doi.org/10.1016/0006-8993(71)90358-1

[15] Beveridge, C., Hart, K., Cristani, C.R., Li, X., Barbierato, E. and Hsu, Y. (2025) Unsupervised Machine Learning for Detecting Mutual Independence among Eigenstate Regimes in Interacting Quasiperiodic Chains. *Physical Review B*, **111**, L140202. https://doi.org/10.1103/physrevb.111.l140202