Open Access Library Journal
2026, Volume 13, e14704
ISSN Online: 2333-9721
ISSN Print: 2333-9705

A Foundational Protocol for Reproducible Visualization in Multivariate Quantum Data

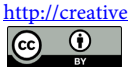
Cassio R. Cristani, Daniele Tessera

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) A Foundational Protocol for Reproducible Visualization in Multivariate Quantum Data. *Open Access Library Journal*, **13**: e14704.
https://doi.org/10.4236/oalib.1114704

Received: December 3, 2025
Accepted: February 8, 2026
Published: February 11, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

Abstract

The visualization of high-dimensional data is a cornerstone of modern scientific inquiry, particularly in quantum physics, where complex non-linear interactions define system behavior. While linear dimensionality reduction methods provide mathematical guarantees of reproducibility, they fail to capture the intricate manifolds underlying such data. Non-linear techniques like Uniform Manifold Approximation and Projection (UMAP) are therefore essential, but their stochastic optimization introduces a fundamental challenge: the lack of reproducibility across independent runs. In this work, we introduce a foundational protocol to establish UMAP as a reproducible tool for scientific visualization. We define explicit, quantitative criteria for embedding convergence, requiring that repeated executions of UMAP under fixed parameters consistently produce a single connected embedding with zero variance in the number of connected components. This criterion transforms UMAP from an exploratory heuristic into a deterministic mapping procedure. Applying the protocol to high-dimensional multivariate quantum data, we demonstrate that feature standardization promotes rapid and consistent convergence at substantially smaller neighborhood sizes, whereas raw data require careful parameter tuning to achieve reproducibility. Our framework provides a rigorous methodological foundation for distinguishing robust visual structures from stochastic artifacts, elevating non-linear visualization to a reproducible component of the scientific process.

Subject Areas

Big Data Search and Mining

Keywords

Reproducible Visualization, Dimensionality Reduction, UMAP, Nonlinear Manifold Learning, Quantum Many-Body Data, High-Dimensional Data

Analysis, Embedding Convergence, Scientific Visualization, Data Standardization

1. Introduction

The curse of dimensionality presents a fundamental barrier to understanding complex systems [1]. In fields such as genomics and quantum many-body physics [2] [3], the state of a system is described by hundreds or thousands of interdependent variables, creating a data landscape that is intrinsically high-dimensional and difficult to comprehend in its raw form. Visualization—the projection of data into a 2D or 3D space accessible to human comprehension—is therefore not merely an aid but a necessity for scientific discovery. Dimensionality reduction (DR) provides the mathematical foundation for this process, serving as the essential bridge between abstract, high-dimensional data and actionable insight [4].

While linear DR methods like Principal Component Analysis (PCA) [5] are computationally robust and provide a unique, reproducible projection, they face a critical limitation: they assume the data's intrinsic structure lies in a linear subspace. This assumption fails for the intricate, non-linear interactions that define many modern scientific domains. The coordinated expression of gene networks [6] [7], the entangled wavefunctions of quantum particles [8], or the emergent phenomena in complex materials [9] all manifest as complex, low-dimensional manifolds embedded within a high-dimensional space. To faithfully visualize these structures, non-linear dimensionality reduction (NLDR) is not just beneficial—it is crucial. Techniques like UMAP (Uniform Manifold Approximation and Projection) are specifically designed to unravel and preserve these non-linear relationships, making the invisible fabric of complex interactions visually apparent [10]-[13]. Other widely used NLDR methods, such as t-SNE, pursue similar goals of manifold visualization but are not the focus of the present work.

However, this powerful capability introduces a profound methodological challenge. Unlike their linear counterparts, NLDR methods are inherently stochastic and sensitive to hyperparameters. A UMAP embedding is the result of a non-convex optimization process, meaning that different runs on the same data can converge to different local minima, producing visually distinct—and sometimes contradictory—projections. This lack of determinism creates a crisis of reproducibility: if a visualization cannot be consistently reproduced, can it be trusted as a foundation for scientific reasoning? When one research group identifies a "cluster" or a "trajectory" in their UMAP plot, while another sees a different organization, there is no objective framework to determine which, if either, reflects a true property of the data versus an artifact of the algorithm's randomness [14] [15].

This paper addresses this critical gap by establishing a rigorous, protocol-driven framework for reproducible NLDR. We posit that for NLDR to be a reliable tool for scientific visualization, it must transcend its role as an exploratory heuristic

and adopt the reproducibility standards expected of any scientific measurement. We introduce a mathematically defined UMAP protocol that establishes formal convergence criteria, transforming subjective visual assessment into objective, quantitative validation. Through a systematic study of high-dimensional quantum data, we demonstrate how preprocessing and parameter selection dictate the pathway to reproducibility, providing clear guidelines for ensuring that visual discoveries are not stochastic artifacts but robust, verifiable features of the underlying data. Our work aims to fortify the crucial bridge between data and discovery, ensuring that the power of non-linear visualization is matched by the rigor required for conclusive scientific interpretation.

2. Data Organization and Structure

This work leverages high-dimensional quantum data to establish a rigorous visualization protocol, with a direct application to documenting the novel nonergodic metal (NEM) phase [16]. As visual confirmation is crucial for interpreting complex phase transitions, as done in [17], according to the pipeline presented in **Figure 1**, which produces an unsupervised machine learning phase classifier. Our objective with the current work is to produce reliable, human-interpretable projections of the underlying quantum state space.

The core data for each interaction strength λ forms a matrix $X^{(\lambda)} \in \mathbb{R}^{n\times d}$, where $n \approx 23000$ rows correspond to unique quantum eigenstates sampled from a trajectory across the generalized Aubry-André model, and $d = 750$ columns are the ordered components of each eigenstate's entanglement spectrum [18]. Our fundamental challenge is the dimensionality reduction $X^{(\lambda)} \to Y^{(\lambda)}$, where $Y^{(\lambda)} \in \mathbb{R}^{n\times 2}$ is a 2D embedding suitable for visual analysis.

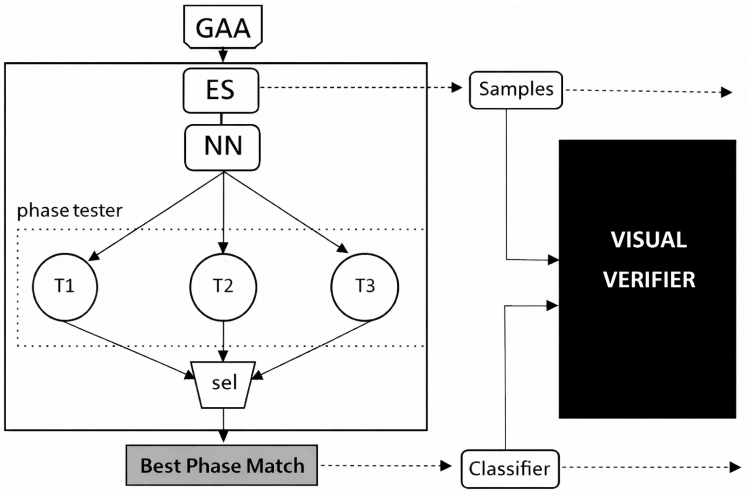


Figure 1. Research pipeline for quantum phase classification, with post-visualization.

The feature space exhibits a pronounced multiscale character, spanning an immense dynamic range. This is illustrated in **Figure 2**, which presents the median trajectory for each feature using all sample space for the whole λ datasets, plot-

ted on a logarithmic scale. The highest feature values are on the order of 10^0, while the lowest values approach 10^{-17}. To contextualize this scale, the difference in variance between the dominant and subtle features is comparable to the difference in mass between a school bus and the planet Pluto.

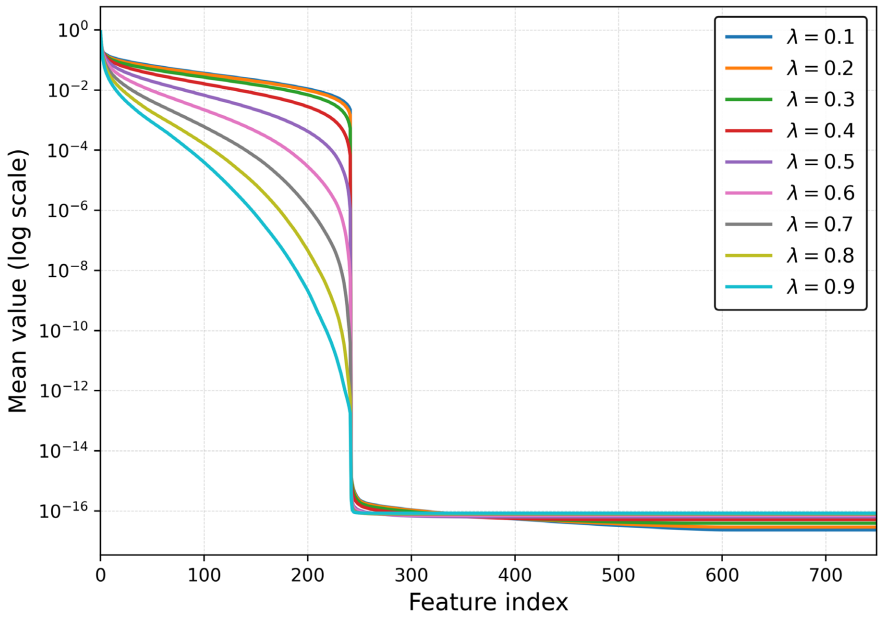


Figure 2. Feature-wise mean across all λ, presented on a logarithmic scale.

3. A Reproducibility Protocol for UMAP

To establish UMAP as a reliable component of the scientific process for analyzing quantum phenomena, it is imperative to guarantee that its visual outputs are not artifacts of stochastic optimization but faithful, reproducible representations of the underlying data structure. We introduce a formal protocol centered on a rigorous definition of embedding convergence, which provides the necessary foundation for trustworthy visual analysis.

3.1. UMAP Operation

Uniform Manifold Approximation and Projection (UMAP) [13] constructs a low-dimensional embedding through a two-stage process. First, it builds a fuzzy topological representation of the high-dimensional data by identifying neighborhoods through a k-nearest neighbor (KNN) graph. The parameter N (n_neighbors) controls the scale of this construction: small N values capture fine-grained local structure, while larger N values emphasize broader global relationships.

The second stage employs stochastic gradient descent to optimize a low-dimensional layout that preserves this topological structure. The optimization applies attractive forces between neighboring points and repulsive forces between non-neighbors, minimizing a cross-entropy loss function. This stochastic process, combined with random initialization, means that independent runs can produce visually distinct embeddings from the same data and parameters.

3.2. The Significance of Convergent Embeddings

The core challenge in applying stochastic dimensionality reduction techniques like UMAP to scientific discovery is distinguishing meaningful geometric structure from computational artifacts. A UMAP embedding that varies significantly across runs with identical parameters offers little scientific value, as its visual representation cannot be reliably interpreted or communicated.

The ideal scenario is convergence to a unique and stable geometry. However, this is complicated by large sample sizes and extreme dynamic ranges in multivariate feature spaces, both common characteristics of scientific datasets. When multiple independent executions of UMAP under a fixed configuration nonetheless produce identical connected structures, it provides compelling evidence that these structures are intrinsic properties of the preprocessed data's topology rather than stochastic artifacts. Achieving this convergent state—where the number of connected substructures is consistently one across all runs with zero variance—transforms UMAP from an exploratory heuristic into a deterministic mapping procedure.

This convergence is the cornerstone of scientific utility. Once a unique embedding is guaranteed for a given dataset and preprocessing method, it can serve as a definitive visual map for consistently interpreting and explaining that dataset. Different scientists can independently arrive at the same visualization, enabling unified interpretation, direct comparison of results, and collaborative hypothesis testing on a stable geometric representation. Furthermore, by achieving this convergent state for both raw and standardized data, one can then visually inspect whether the two preprocessing pathways reveal the same underlying global structure, providing a powerful cross-validation of the observed phenomena.

3.3. Mathematical Framework for Convergence

To formalize this notion, we define a discrete random variable $C_{\theta,f}$ representing the number of connected substructures in an embedding. For a fixed UMAP parameter set $\theta=\left(N,\text{min\_dist},\cdots\right)$ and preprocessing function f, we execute m independent runs to obtain embeddings $\left\{Y^{(1)},\cdots,Y^{(m)}\right\}$. For each embedding, $C^{(s)}=C\left(Y^{(s)}\right)$ is the number of connected components in its KNN graph for a given N.

Our primary theoretical claim is that for a suitably preprocessed dataset, the stochasticity of UMAP can be controlled via the neighborhood parameter N. Specifically, we posit that $C_{N,f}$ converges in probability to 1 as N increases:

$$\lim_{N\to\infty} P\left(\left|C_{N,f}-1\right|\geq\epsilon\right)=0 \text{ for any } \epsilon>0. \tag{1}$$

Justification: As N increases, the number of neighborhoods with a population N decreases, and the k-NN graph underlying UMAP's initial topological representation becomes less fractioned (more connected). In the limiting case where N approaches the sample size, the graph becomes fully connected, possessing exactly one connected component. The optimization process, while sto-

chastic, aims to find a low-dimensional layout that preserves this connectivity structure. For sufficiently large N, the attractive forces imposed by the global connectivity dominate the repulsive forces, constraining the optimization to produce a single, coherent structure. Standardization facilitates this convergence by ensuring the graph's connectivity reflects balanced global correlations rather than being fragmented by disparate feature scales.

Based on this convergence behavior, we define a rigid, empirical criterion for a configuration to be considered minimally reproducible in practice. A configuration (θ, f) is necessary reproducible if, across a finite number of runs m, the following condition is met:

$$\min_s C^{(s)} = \max_s C^{(s)} = 1 \text{ and } \mathrm{Var}\left[C^{(s)}\right] = 0 \tag{2}$$

This condition ensures that every execution produces a single, connected structure with zero empirical variance, guaranteeing a unique connected geometry for scientific analysis.

3.4. Protocol Implementation

The implementation of our reproducibility protocol follows a rigorous pipeline designed to systematically evaluate stability across the entire phase space of the quantum model:

1) **Systematic Data Generation:** The protocol is applied across the full spectrum of quantum phases. We generate datasets for multiple interaction strengths, $\lambda \in \{\lambda_1, \lambda_2, \cdots, \lambda_k\}$, ensuring the analysis covers diverse quantum regimes.

2) **Preprocessing:** For each λ, we generate both the raw dataset $X_{\text{raw}}^{(\lambda)}$ and the standardized dataset $X_{\text{std}}^{(\lambda)}$.

3) **Multi-Scale Screening:** For each dataset $X^{(\lambda)}$ and for each $N \in \{10, 15, 20, 30, 40, 50\}$, we execute $m = 10$ independent UMAP runs. The choice of $m = 10$ represents a balance between computational feasibility and statistical sensitivity: given the binary nature of the convergence criterion (zero versus non-zero variance in the number of connected components), ten trials are sufficient to reliably detect empirical instability while keeping the total computational cost manageable for large datasets.

4) **Consistency Assessment:** For each configuration (λ, N, f), we compute the empirical distribution of C across the m runs and apply the convergence criteria from Equation (2).

5) **Convergence Verification:** For each λ and preprocessing method f, we identify the minimal $N_{\min}^{(\lambda, f)}$ for which the configuration achieves scientific reproducibility.

6) **Robustness Validation:** We aggregate results across all λ values. A preprocessing method f is considered *universally robust* if it achieves scientific reproducibility for all λ at a consistent and feasible N.

This protocol provides a mathematically grounded framework to identify UMAP configurations that converge to a unique, stable embedding. The subsequent ap-

plication of this protocol determines which configurations, if any, meet the criteria for scientific reproducibility, thereby objectively evaluating UMAP's suitability for rigorous visual analysis of high-dimensional quantum data.

4. Methodology and Results for Substructure Convergence Analysis

To mirror the default behaviour of exploratory data analysis, each UMAP execution was initialized using the algorithm's random initialization (*i.e.*, without a fixed seed). As a result, the stochastic gradient descent trajectory differs across runs, making $C_{\theta,f}$ an empirical random variable that quantifies variability in the embedding's substructures as a function of N. We estimate the distribution of $C_{\theta,f}$ from $m=10$ independent trials.

This section reports the results of applying our reproducibility protocol to nine multivariate quantum datasets spanning interaction strengths $\lambda=0.1$ to 0.9, with comprehensive statistics for the number of connected substructures summarized in Table 1. Across all interaction regimes, the number of disconnected substructures—measured as connected components in the two-dimensional KNN graph—decreases monotonically with increasing N, consistent with theoretical expectations that larger neighborhoods enhance global connectivity by expanding UMAP's local simplicial complexes.

Table 1. Empirical distribution of the random variable C (number of UMAP connected components) comparing raw and standardized data for neighborhood sizes $N=10,15,20,30,40,50$. **Bold italic values** indicate fully connected embeddings (Mean = 1, Var. = 0).

Dataset	N	Raw Data				Standard Scaled			
		Min	Max	Mean	Var.	Min	Max	Mean	Var.
$\lambda=0.1$	10	1	3	1.9	0.49	1	3	1.6	0.44
	15	1	1	1.0	0.00	1	1	1.0	0.00
	20	1	1	1.0	0.00	1	1	1.0	0.00
	30	1	1	1.0	0.00	1	1	1.0	0.00
	40	1	1	1.0	0.00	1	1	1.0	0.00
	50	1	1	1.0	0.00	1	1	1.0	0.00
$\lambda=0.2$	10	3	5	3.9	0.49	1	3	2.0	0.59
	15	2	2	2.0	0.00	1	1	1.0	0.00
	20	1	2	1.4	0.24	1	1	1.0	0.00
	30	1	1	1.0	0.00	1	1	1.0	0.00
	40	1	1	1.0	0.00	1	1	1.0	0.00
	50	1	1	1.0	0.00	1	1	1.0	0.00
$\lambda=0.3$	10	12	16	14.5	1.85	1	2	1.4	0.24
	15	4	7	5.7	0.81	1	1	1.0	0.00
	20	3	5	4.2	0.76	1	1	1.0	0.00

Continued

	30	2	2	2.0	0.00	1	1	1.0	0.00
	40	2	2	2.0	0.00	1	1	1.0	0.00
	50	1	1	1.0	0.00	1	1	1.0	0.00
$\lambda = 0.4$	10	20	32	25.9	8.29	3	5	3.5	0.45
	15	6	9	7.8	1.56	1	1	1.0	0.00
	20	2	4	2.4	0.44	1	2	1.1	0.09
	30	1	1	1.0	0.00	1	1	1.0	0.00
	40	1	1	1.0	0.00	1	1	1.0	0.00
	50	1	1	1.0	0.00	1	1	1.0	0.00
$\lambda = 0.5$	10	31	40	36.1	5.48	3	7	4.5	1.44
	15	15	20	17.8	1.96	1	2	1.1	0.09
	20	10	17	13.0	2.99	1	1	1.0	0.00
	30	3	7	5.0	1.21	1	1	1.0	0.00
	40	2	4	3.0	0.40	1	1	1.0	0.00
	50	3	4	3.4	0.24	1	1	1.0	0.00
$\lambda = 0.6$	10	37	49	42.8	13.54	4	8	6.0	1.21
	15	10	23	18.2	15.76	2	3	2.2	0.16
	20	5	10	7.3	1.82	2	2	2.0	0.00
	30	2	4	3.3	0.61	1	2	1.9	0.09
	40	2	4	3.1	0.69	1	2	1.1	0.09
	50	2	3	2.3	0.21	1	1	1.0	0.00
$\lambda = 0.7$	10	69	83	75.5	16.24	13	20	15.5	3.84
	15	34	44	40.2	9.55	1	3	2.3	0.41
	20	20	23	22.0	1.00	1	4	2.1	0.69
	30	10	14	12.0	1.80	1	2	1.3	0.21
	40	3	11	6.9	5.29	1	1	1.0	0.00
	50	3	4	3.1	0.09	1	1	1.0	0.00
$\lambda = 0.8$	10	71	84	77.9	22.85	15	20	17.4	3.24
	15	32	44	36.5	11.83	2	6	4.0	1.21
	20	15	23	18.2	5.38	1	2	1.2	0.16
	30	9	13	11.1	1.69	1	1	1.0	0.00
	40	1	5	3.3	1.00	1	1	1.0	0.00
	50	1	3	1.5	0.45	1	1	1.0	0.00
$\lambda = 0.9$	10	82	93	87.8	11.56	14	22	19.6	4.45
	15	27	40	34.2	11.16	2	7	4.6	1.85
	20	16	27	22.6	10.24	1	2	1.2	0.16
	30	6	11	7.8	3.17	1	1	1.0	0.00
	40	4	7	5.6	0.64	1	1	1.0	0.00
	50	2	4	2.6	0.44	1	1	1.0	0.00

For raw data, small neighborhood sizes ($N = 10,15$) produce highly fragmented embeddings, often with tens of disconnected components at intermediate interaction strengths ($\lambda \approx 0.5$ - 0.8). These regimes correspond to unstable manifolds that are insufficiently sampled, where stochastic initialization yields qualitatively distinct topologies across runs, reflected in high variance. As N increases, connectivity constraints are relaxed and components merge, reducing stochastic sensitivity and stabilizing the embedding.

These results indicate that substructure convergence in UMAP can be interpreted as a discrete topological transition governed by the connectivity of the neighborhood graph: fragmented simplicial complexes at small N progressively merge until a single connected manifold emerges. The zero-variance condition marks the onset of global topological stability under random initialization and parameter noise.

In summary, convergence toward a single connected embedding is a general property of UMAP on complex multivariate datasets. This process is accelerated and stabilized by standardization, which enforces uniform metric scaling, and is guaranteed at sufficiently large neighborhood sizes. The resulting criterion provides a principled method for identifying the minimal N required for topologically stable and reproducible embeddings, offering a rigorous foundation for UMAP-based analysis in mathematical and quantum physics contexts.

4.1. Standardized Data Produces a Single Connected Structure

For standardized data, UMAP embeddings consistently converge to a single connected structure across all nine quantum datasets when the neighborhood size N is sufficiently large. In most datasets ($\lambda = 0.1, 0.2, 0.3, 0.4, 0.7, 0.8, 0.9$), this convergence occurs already at $N = 15$ - 30 , while for all datasets $N = 50$ guarantees a single structure. The variance column for standardized data in **Table 1** shows a clear trend: as N increases, the number of disconnected substructures systematically reduces, ultimately reaching one. This behavior confirms that standardization stabilizes local and global relationships in high-dimensional quantum data, enabling reproducible structural interpretations.

4.2. Fragmentation in Raw Data Embeddings

In contrast, embeddings generated from raw data exhibit pronounced fragmentation and are highly sensitive to initialization. At small neighborhood sizes ($N = 10$), the mean number of disconnected substructures ranges from 1.9 ($\lambda = 0.1$) to 87.8 ($\lambda = 0.9$), with substantial variance, indicating highly irregular and fractured embeddings. Increasing N progressively reduces fragmentation, but a single connected structure is achieved only in 4 of 9 datasets ($\lambda = 0.1, 0.2, 0.3, 0.4$). For intermediate regimes ($\lambda = 0.5, 0.6$) and higher interaction strengths ($\lambda = 0.7, 0.8, 0.9$), multiple disconnected substructures persist even at $N = 50$, demonstrating that raw data cannot reliably produce a fully connected embedding in all cases.

4.3. Preliminary Insights for Reliable Non-Linear Dimensionality Reduction

The pursuit of reliable visualization for complex multivariate quantum data presents a significant challenge. This investigation, conducted across nine large-scale datasets (approximately 23,000 samples with 750 features), explored the stability of UMAP embeddings as a potential path forward. The experimental protocol, involving 10 independent executions for each neighborhood parameter (N), provides initial evidence, though further validation is needed.

A central finding from these preliminary tests, summarized in Table 2, concerns the optimization landscape. The raw data's high-dimensional, fragmented nature appears to create a complex topography with many local minima, slowing convergence, and often leading to multiple substructures. In contrast, standardized data facilitates a more direct path to a single, deterministic embedding by presenting a smoother global structure for the algorithm to capture.

Table 2. Convergence results across datasets.

λ	N_{std}	**Std.**	N_{raw}	**Raw**
0.1	15	√	15	√
0.2	20	√	–	×
0.3	30	√	–	×
0.4	30	√	–	×
0.5	30	√	–	×
0.6	40	√	50	√
0.7	40	√	50	√
0.8	40	√	50	√
0.9	50	√	50	√
Total	–	**9/9**	–	**4/9**

These observations begin to outline a methodological approach for generating more reliable explanations of quantum data through visualization:

Data Preprocessing as a Stabilizer: The consistent performance of standardized data suggests that preprocessing which promotes global coherence can be a powerful tool for stabilizing the embedding process and mitigating local fragmentation.

Navigating the Optimization Landscape: The results indicate that the choice of neighborhood size is pivotal. Larger neighborhood sizes appear to help the algorithm navigate past local minima in complex data, a consideration that becomes especially relevant when preprocessing is not applied.

Convergence as a Diagnostic Metric: Monitoring variance in substructure counts across multiple runs is a practical diagnostic. A consistent, single structure across executions can bolster confidence that the resulting visualization reflects a stable global geometry rather than a transient local configuration.

Interpreting with an Awareness of Stability: Embeddings that are sensitive to initial conditions may offer valuable but partial insights. Acknowledging this instability can lead to a more nuanced interpretation, where multiple runs are seen as exploring a landscape of possible structures inherent to the data.

This preliminary work suggests that a conscious approach to data preparation and parameter selection, coupled with simple validation checks, can pave the way for more trustworthy visualizations. For complex multivariate quantum data, such practices may ultimately lead to more reliable explanations and a deeper understanding of the underlying quantum phenomena.

4.4. Limitations and Future Work

While our protocol establishes a necessary condition for reproducible visualization—the convergence to a stable, connected embedding—we acknowledge that reproducibility alone does not guarantee physical interpretability. A legitimate critique is that by increasing the neighborhood parameter N or applying standardization, we might be artificially connecting genuinely distinct physical regions, effectively smoothing over physically significant discontinuities.

However, we posit that achieving a single, reproducible structure is a fundamental prerequisite for any robust physical interpretation. The high fragmentation observed in raw data embeddings at small N values, characterized by high variance across runs, represents a state of analytical chaos where no stable interpretation is even possible. Our protocol systematically navigates this stochasticity to arrive at a deterministic endpoint. The resulting connected structure provides a stable canvas upon which physically meaningful patterns (such as gradients, clusters, or topological voids) can be reliably identified and studied.

In this work, we focus on establishing this methodological foundation of reproducibility. A comprehensive validation of the physical meaning embedded within these stable structures—a visual cross-referencing of UMAP patterns with independent physical observables—remains a critical task for future work. Such an analysis is challenging in this context due to the lack of a clear ground truth and the high dimensionality of the original space, which complicates quantitative measures such as trustworthiness and continuity. Future research will involve a detailed visual comparison across the quantum phase diagram and the development of specialized metrics to quantify how well the embedding preserves the physical relationships between a representative subset of states. Future work will incorporate pseudo ground-truth physical observables as color overlays on the converged embeddings, enabling a direct visual cross-validation and providing a more suitable and intuitive tool for physical inspection.

5. Conclusions

This work establishes a rigorous protocol for rendering UMAP a reproducible tool in scientific visualization. By framing reproducibility as the convergence toward a single, connected embedding topology, we introduced quantitative criteria that

transform a stochastic algorithm into a verifiable mapping process. Applied to complex multivariate quantum datasets, our protocol revealed a consistent and interpretable pattern: as the neighborhood parameter N increases, the embedding undergoes a discrete topological transition from fragmented to unified manifolds, converging to a deterministic structure once global connectivity percolates through the data graph.

Standardization plays a decisive role in this transition. By rescaling the features, it equalizes the contribution of disparate energy scales and promotes faster convergence at smaller N, effectively stabilizing the manifold reconstruction. In contrast, raw data require larger neighborhoods to achieve the same level of reproducibility, underscoring the sensitivity of distance metrics to feature-scale imbalance. Across all interaction regimes, however, the limiting behavior is universal: UMAP asymptotically converges to a single, stable embedding once sufficient neighborhood connectivity is reached.

The implications extend beyond the present application. The proposed reproducibility protocol provides a general framework for distinguishing robust visual patterns from stochastic artifacts across domains using non-linear embeddings. It offers a methodological foundation for integrating visualization within the standards of scientific reproducibility, where a figure represents not a random snapshot of computation, but a stable property of the underlying data manifold. Future work will expand this framework to include geometric stability metrics and explore how convergence correlates with physically meaningful invariants in quantum phenomena.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bellman, R. (1966) Dynamic Programming. *Science*, **153**, 34-37. https://doi.org/10.1126/science.153.3731.34

[2] Libbrecht, M.W. and Noble, W.S. (2015) The Nature of Machine Learning in High-Dimensional Data. *Nature Reviews Genetics*, **16**, 728-740.

[3] Carrasquilla, J. and Melko, R.G. (2017) Machine Learning Phases of Matter. *Nature Physics*, **13**, 431-434. https://doi.org/10.1038/nphys4035

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

[5] Jolliffe, I.T. and Cadima, J. (2016) Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A*: *Mathematical*, *Physical and Engineering Sciences*, **374**, Article 20150202. https://doi.org/10.1098/rsta.2015.0202

[6] Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., *et al.* (2019) Visualizing Structure and Transitions in High-Dimensional Biological Data. *Nature Biotechnology*, **37**, 1482-1492. https://doi.org/10.1038/s41587-019-0336-3

[7] Hu, Q., Lu, X., Xue, Z. and Wang, R. (2025) Gene Regulatory Network Inference

during Cell Fate Decisions by Perturbation Strategies. *npj Systems Biology and Applications*, **11**, Article No. 23. https://doi.org/10.1038/s41540-025-00504-2

[8] Brown, M.R., *et al.* (2025) Machine Learning the Entanglement Spectrum of Disordered Quantum Spin Liquids. *Physical Review X*, **15**, Article 021045.

[9] Kim, B., Jin, J., Wang, Z., He, L., Christensen, T., Mele, E.J., *et al.* (2023) Three-Dimensional Nonlinear Optical Materials from Twisted Two-Dimensional Van Der Waals Interfaces. *Nature Photonics*, **18**, 91-98. https://doi.org/10.1038/s41566-023-01318-6

[10] Tenenbaum, J.B., Silva, V.D. and Langford, J.C. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, **290**, 2319-2323. https://doi.org/10.1126/science.290.5500.2319

[11] Roweis, S.T. and Saul, L.K. (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, **290**, 2323-2326. https://doi.org/10.1126/science.290.5500.2323

[12] Van der Maaten, L. and Hinton, G. (2008) Visualizing Data Using T-Sne. *Journal of Machine Learning Research*, **9**, 2579-2605.

[13] McInnes, L., Healy, J., Saul, N. and Melville, J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Journal of Open Source Software*, **3**, Article 861. https://doi.org/10.21105/joss.00861

[14] 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

[15] Donaldcito, A., Smith, J., *et al.* (2022) Reproducibility of Machine Learning Algorithms in Single-Cell Data Analysis. *Nature Methods*, **19**, 1047-1055.

[16] 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 245701. https://doi.org/10.1103/physrevlett.121.245701

[17] 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

[18] 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 010504. https://doi.org/10.1103/physrevlett.101.010504