A PRINCIPAL SUBMANIFOLD-BASED APPROACH FOR CLUSTERING AND MULTISCALE RNA CORRECTION

BY MENGHAO WU[1,a], ZHIGANG YAO[2,b]

[1]*Department of Statistics and Data Science, National University of Singapore, 117546 Singapore,* [a]*menghao.wu@u.nus.edu*

[2]*Department of Statistics and Data Science, National University of Singapore, 117546 Singapore,* [b]*zhigang.yao@nus.edu.sg*

RNA structure determination is essential for understanding its biological functions. However, the reconstruction process often faces challenges, such as atomic clashes, which can lead to inaccurate models. To address these challenges, we introduce the principal submanifold (PSM) approach for analyzing RNA data on a torus. This method provides an accurate, low-dimensional feature representation, overcoming the limitations of previous torus-based methods. By combining PSM with DBSCAN, we propose a novel clustering technique, the principal submanifold-based DBSCAN (PSM-DBSCAN). Our approach achieves superior clustering accuracy and increased robustness to noise. Additionally, we apply this new method for multiscale corrections, effectively resolving RNA backbone clashes at both microscopic and mesoscopic scales. Extensive simulations and comparative studies highlight the enhanced precision and scalability of our method, demonstrating significant improvements over existing approaches. The proposed methodology offers a robust foundation for correcting complex RNA structures and has broad implications for applications in structural biology and bioinformatics.

1. Introduction.

1.1. *Background.* In the realm of life sciences, ribonucleic acid (RNA) is a pivotal molecule that undertakes various biological functions. Beyond its role as a genetic information conveyor and participant in protein synthesis, RNA is instrumental in regulating gene expression, catalyzing biochemical reactions, and other critical biological processes (Kieft et al., 2002; Moore, 1998; Jinek et al., 2014; Chen, 2020). Research has revealed that RNA's functionality is largely dictated by its 3-dimensional structure rather than merely its nucleotide sequence. Consequently, a profound understanding of RNA's spatial configuration is required to comprehend its functional mechanisms and decode its multifaceted roles in cellular processes (Doudna and Cech, 2002; Lilley, 2000).

In recent years, the development of high-resolution structural determination techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (Cryo-EM) has significantly enhanced our understanding of the interplay between RNA structure and function (Shen, Cai and Tinoco Jr, 1995; Jain, Richardson and Richardson, 2015). However, reconstructing RNA's backbone conformation remains challenging, often leading to RNA clashes, where atoms are too close based on van der Waals radii, resulting in physically implausible structures (Batsanov, 2001; Sponer et al., 2018). These clashes highlight potential inaccuracies in structural prediction and require advanced computational correction methods. Effective clash correction is essential for accurate RNA modeling, as unresolved clashes can distort structural interpretations and hinder insights into RNA's roles in catalysis, binding, and regulation. Improving RNA structural reliability through clash correction enables more precise models, crucial for studying RNA's functions in biological systems and diseases

---

(Serganov and Nudler, 2013). As RNA research advances, computational refinement tools are increasingly vital, not only in academia but also in medical and biotechnological applications, aiding the development of RNA-based therapeutics and diagnostics (Zhu et al., 2022; Lu et al., 2024).

For RNA clashes, most attention is focused on suites, which are segments of an RNA molecule spanning from one sugar ring to the next (Murray et al., 2003). A comprehensive understanding of suites is achieved by adopting a dual-scale perspective that encompasses both microscopic and mesoscopic scales. At the microscopic scale, analysis concentrates on the atomic-level details of a suite's RNA backbone, characterized by its dihedral angles that capture the conformational intricacies of the alternating phosphate and ribose sugar units with attached nitrogenous bases. In contrast, the mesoscopic scale expands the analysis to a sequence of suites, emphasizing the geometric positioning of sugar ring centers along the RNA strand to capture the overall shape and form of the molecule. At this level, researchers examine interactions among multiple nucleotides, which give rise to structural features such as helices, loops, and bulges (Shivashankar, 2002). Size-and-shape analysis at the mesoscopic scale thus provides a framework for understanding the geometric configuration of RNA molecules beyond the atomic details (Dryden and Mardia, 2016).

1.2. *Existing methods.* Currently, there are two main methods to correct RNA clashes. One method is based on molecular dynamics. According to approximations based on biophysical and chemical laws, simulated atoms can move toward the position of lowest energy, as exemplified by ERRASER (Chou et al., 2013), which exhaustively samples nucleotide conformations and refines them using an energy function and electron-density correlation. ERRASER effectively corrects the majority of geometric errors while maintaining or improving agreement with diffraction data. However, it has limitations: the process is time-consuming and, despite its precision, this method often fails to resolve all clashes.

Another method is data-driven by considering the geometric structure of RNA. The main challenge is performing a high-dimensional clustering of RNA suites on the torus. Traditional methods include linear and non-linear methods. Linear methods, such as PCA, efficiently capture variance but fail to fit non-linear structures such as a torus, leading to information loss. non-linear methods, including Isomap (Tenenbaum, Silva and Langford, 2000), LLE (Roweis and Saul, 2000; Donoho and Grimes, 2003), t-SNE (Van der Maaten and Hinton, 2008), and UMAP (McInnes, Healy and Melville, 2018), better capture complex geometries by preserving local manifold structures. However, they lack strong theoretical guarantees, are computationally expensive at scale, and are sensitive to parameter choices, often failing to retain global structures.

To handle the RNA data on the high-dimensional torus, Wiechers et al. (2023) proposed an RNA correction method MINT-AGE-CLEAN, a data-driven method utilizing MINT-AGE (Mardia et al., 2022) to categorize RNA data into different classes, before correcting RNA clashes across both microscopic and mesoscopic scales. It overcomes the limitations of traditional dimensionality reduction methods by mapping RNA data on the torus. The central idea is Torus PCA (tPCA) (Eltzner, Huckemann and Mardia, 2018), which involves a transformation — known as Torus-to-Stratified-Sphere (TOSS) mapping — that converts the data from a torus to a sphere. Although this approach allows the adaptation of principal nested spheres analysis (PNS) (Jung, Dryden and Marron, 2012), a limitation still exists: tPCA fails to accurately capture the geometric structure of RNA data in its native space (e.g., the torus). Therefore, there is still room for improvement by exploring the hidden structure of its intrinsic space.

1.3. *Main Contribution.* For high-dimensional clustering, the general approach is to encode features into a low-dimensional space prior to clustering. The quality of low-dimensional representation is crucial, as it directly impacts the preservation of the data's intrinsic structure and, consequently, the overall clustering performance. Instead of clustering methods such as MINT-AGE, which rely on tPCA before clustering, we adopt the Principal Submanifold (PSM) approach (Yao et al., 2023; Yao, Eltzner and Pham, 2026) to handle high-dimensional features. Our work makes the following key contributions:

- PSM can accurately fit low-dimensional principal submanifolds with more general structures. Unlike existing methods like tPCA, which rely on spherical transformations and cause unignorable information loss, PSM preserves both local and global geometric features directly on the torus.

- We introduce PSM-DBSCAN, a clustering method that combines PSM with the density-based DBSCAN algorithm (Ester et al., 1996). PSM provides a precise low-dimensional representation, enabling DBSCAN to detect clusters more accurately, even in complex and noisy data. Compared to tPCA-DBSCAN, MINT-AGE and traditional methods such as hierarchical and spectral clustering, PSM-DBSCAN consistently achieves higher accuracy and stability.

- We evaluate PSM-DBSCAN for RNA correction, demonstrating its practical utility in real-world scenarios. The enhanced clustering performance of PSM-DBSCAN leads to more accurate RNA correction by ensuring more reliable categorization of RNA data. Comparative studies show that our method outperforms MINT-AGE-CLEAN, achieving superior results in resolving atomic clashes and modeling complex RNA structures.

Without delving into details, we illustrate the improvement in two steps: dimensionality reduction and clustering performance. Figure 1(a) illustrates samples on a 2-dimensional torus in a 3-dimensional space, situated around a 1-dimensional curve. We can see that PSM accurately fits samples onto a 1-dimensional curve (Figure 1(b)), while tPCA fits samples onto a circle (Figure 1(c)), thereby losing critical information of data. This highlights PSM's advantage in aligning the samples more precisely with the underlying low-dimensional structure, preserving the intrinsic geometric information in the data and retaining both local and global geometric features.

For clustering, we consider a high-dimensional scenario where samples are situated around three 1-dimensional curves on a 7-dimensional torus, with added Gaussian noise (Figure 2(a)). Figure 2(b) demonstrates that PSM can effectively distinguish samples from different curves, forming three well-defined and compact clusters. In contrast, in Figure 2(c), MINT-AGE fails to differentiate between samples from distinct curves, resulting in a larger number of clusters and mistakenly grouping samples from two separate curves into the same cluster.

1.4. *Outline.* In Section 2, we introduce the framework of the principal submanifold. We first describe the basic setup, then present PSM for dimensionality reduction and PSM-DBSCAN for clustering, followed by our method for RNA multiscale correction. In Section 3, we provide numerical simulations that demonstrate the superior performance of PSM and the associated clustering techniques. Section 4 details the RNA correction in real data. Finally, Section 5 summarizes the main findings and contributions of our study and discusses several promising future works.
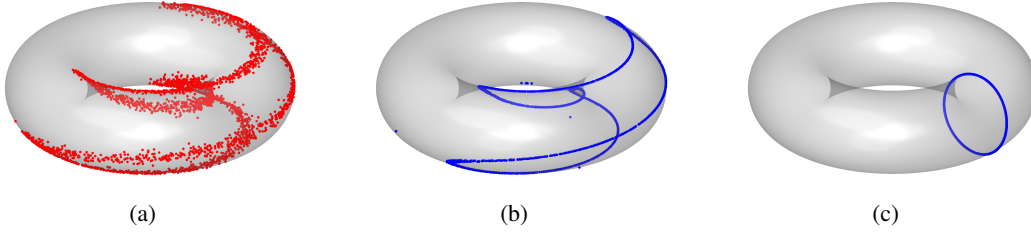
FIG 1. *An intuitive illustration of dimensionality reduction on a 2-dimensional torus using PSM and tPCA. Points (red) indicate the original noisy sample on the torus (a), while points (blue) represent the results from PSM (b) and tPCA (c).*
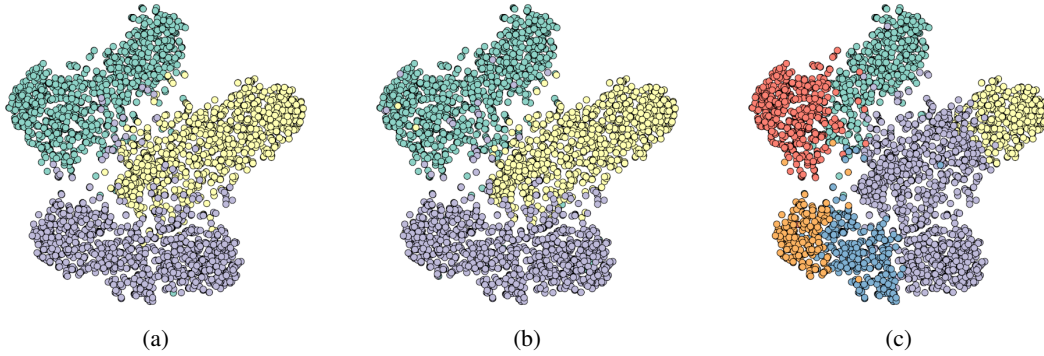

FIG 2. *An intuitive illustration of clustering performance on the high-dimensional data using t-SNE. (a) shows the visualization of three different categories of points, (b) shows the clustering performance of PSM-DBSCAN and (c) shows MINT-AGE.*

2. Methodology.

2.1. *Preliminary.* At the microscopic scale, the geometric relationship between atoms can be characterized by dihedral angles. As shown in Figure 3, four atoms within an RNA suite (specifically, $O - P - O - C$) can form a dihedral angle, denoted by α, in a 3-dimensional space. A dihedral angle is inherently periodic, ranging from 0 to 2π, which aligns naturally with the circular structure of $\mathbb{S}^1$. When considering multiple dihedral angles—each corresponding to a degree of rotational freedom along the RNA backbone—their collective multidimensional nature suggests a higher-dimensional torus. As the Cartesian product of circles, the torus elegantly accounts for angular periodicity, thereby enabling seamless transitions between rotations and enhancing computational simulations and structural predictions. At the microscopic scale, we focus on seven dihedral angles—$(\delta_i, \epsilon_i, \zeta_i, \alpha_{i+1}, \beta_{i+1}, \gamma_{i+1}, \epsilon_{i+1})$—for RNA suite i, as illustrated in Figure 4.

For convenience, seven dihedral angles will be generally denoted as $x \in \mathbb{R}^D$ (i.e. $D = 7$ in the RNA case), where $x = (x^{(1)}, \ldots, x^{(D)})^T$ with each $x^{(j)} \in [0, 2\pi]$ representing the j-th angle. Consequently, x can be considered as embedded on a D-dimensional torus $\mathbb{T}^D$. Accordingly, we can define the samples $\mathcal{X} = \{x_i\}_{i=1}^N$ on the torus $\mathbb{T}^D$. $\mathbb{T}^D$ is defined as the product of D circles:

$$\mathbb{T}^D := [0, 2\pi]^D / \sim,$$

where $\sim$ denotes that 0 and 2π are identified. We can define the geodesic distance between two samples $x_1 = (x_1^{(1)}, \ldots, x_1^{(D)})^T$ and $x_2 = (x_2^{(1)}, \ldots, x_2^{(D)})^T$ on the torus $\mathbb{T}^D$, as the canonical product distance

$$\mathrm{d}_{\mathbb{T}^D}(x_1, x_2) = \sqrt{\sum_{j=1}^{D} \mathrm{d}_{\mathbb{T}}(x_1^{(j)}, x_2^{(j)})^2},$$

where the canonical distance is given by:

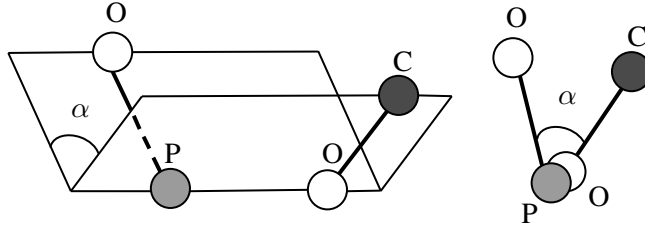$$\mathrm{d}_{\mathbb{T}}(x_1^{(j)}, x_2^{(j)}) = \min(|x_1^{(j)} - x_2^{(j)}|, 2\pi - |x_1^{(j)} - x_2^{(j)}|).$$

FIG 3. *Illustration of the dihedral angles in the RNA backbone between the atom. Reproduced from Mardia (2013).*

Moreover, at the mesoscopic scale, RNA involves additional landmark positions of sugar rings in 3-dimensional space and considering their relative configurations. Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$, where $s_i \in \mathbb{R}^3$ represents the position of the i-th sugar ring center. The size-and-shape space Λ_3^N is defined as:

$$\Lambda_3^N = \{[S] \mid S = (s_1, \ldots, s_N)^T \in \mathbb{R}^{N \times 3}\},$$

where $[S]$ denotes the equivalence class of S under proper Euclidean transformations (ET), including rotations $R \in \Xi(3)$ and translations $v \in \mathbb{R}^3$:

(1) $\quad \mathrm{ET}.S = (Rs_1 + v, Rs_2 + v, \ldots, Rs_N + v)^T, \ [S] = \{\mathrm{ET}.S \mid S \in \Xi(3) \times \mathbb{R}^3\}.$

The Procrustes distance d_Λ between two configurations $[S]$ and $[W]$ in size-and-shape space is given by:

(2) $$\mathrm{d}_\Lambda([S], [W]) = \min_{\mathrm{ET} \in \Xi(3) \times \mathbb{R}^3} \|S - \mathrm{ET}.W\|_2,$$

with the Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^{N \times 3}$. This provides comprehensive insights into mesoscopic geometry, revealing structural variations, similarities, and potential functional motifs in RNA. At the mesoscopic scale, for RNA suite i, we focus on the k suites before and k suites after this central suites, as show in Figure 4.

Since we have defined the geodesic distance in microscopic and mesoscopic scale spaces respectively, we can define the Fréchet mean μ at both two scale space $\mathbb{T}^D$ and Λ_3^N:

$$\mu_{\mathbb{T}} \in \arg\min_{x \in \mathbb{T}^D} \sum_{i=1}^{N} \mathrm{d}_{\mathbb{T}^D}(x, x_i)^2, \quad \mu_\Lambda \in \arg\min_{[S] \in \Lambda_3^N} \sum_{i=1}^{N} \mathrm{d}_{\Lambda_3^N}([S], [S_i])^2.$$
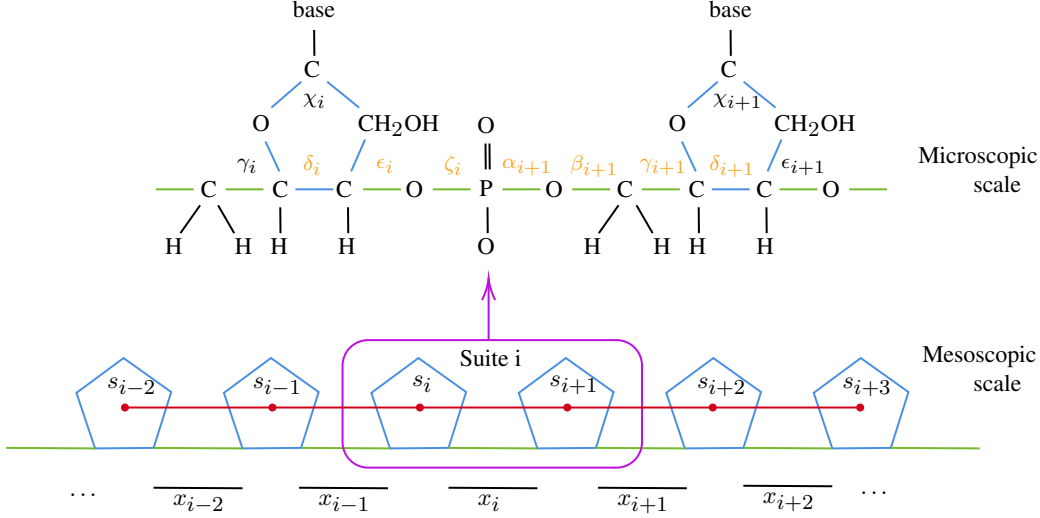
FIG 4. *Visualization of an RNA backbone at microscopic and mesoscopic scale respectively, illustrating seven dihedral angles of i-th suite at microscopic scale and six centers $(k = 2)$ of the sugar rings (from s_{i-2} to s_{i+3}) defining shape of i-th suite at mesoscopic scale.*

2.2. *Statistical model for PSM.* To analyze $\mathcal{X}$ on the torus $\mathbb{T}^D$, tPCA was introduced, which practically projects torus data onto a sphere and applies PNS for low-dimensional representation. Unlike linear methods, it accounts for periodicity but has limitations: the TOSS mapping on the data may lose intrinsic geometric information, and PNS may not fully capture the geometric structure of the torus, leading to additional information loss. See Appendix B for mathematical details of tPCA.

To overcome the challenges posed by torus setting and the weakness of tPCA, we need a new method which can accurately fit low-dimensional representation on the high-dimensional torus. Here, we adopt an advanced method, which comes from a series of works represented by principal submanifold (Yao et al., 2023; Yao, Eltzner and Pham, 2026), a generalization of the principal flow (Panaretos, Pham and Yao, 2014). This paper adopt the theories and methodologies and aims to fit a smooth low-dimensional principal submanifold within data situated on the manifold $\mathcal{M}$, which can capture the main geometric structures of data and minimize the information loss.

To accommodate our RNA setting, let $\mathcal{M} = \mathbb{T}^D$. We assume d^*-dimensional submanifolds class $\Psi = \{\psi : \psi \subset \mathbb{T}^D, \dim(\psi) = d^*, \psi \in C^2(\mathbb{T}^D)\}$, where $d^* < D$. Then, the population principal submanifold ψ^* is conceptually idealized to summarize the data by optimizing the expected squared geodesic distance with respect to the distribution τ. Formally, ψ^* is defined as:

$$\psi^* = \arg\min_{\psi' \in \Psi} \int \mathrm{d}_{\mathbb{T}^D}(x, \psi')^2 \tau(dx).$$

Given samples $\mathcal{X}$ on the torus $\mathbb{T}^D$, we define the empirical principal submanifold $\widehat{\psi}_N$ as follows:

$$\widehat{\psi}_N = \arg\min_{\psi' \in \Psi} \frac{1}{N} \sum_{i=1}^{N} \mathrm{d}_{\mathbb{T}^D}(x_i, \psi')^2.$$

2.3. *Algorithm for PSM.* Here, we elaborate PSM method in practice to calculate the empirical principal submanifold $\widehat{\psi}_N$ based on the samples $\mathcal{X}$, visualized in Figure 5. The algorithm of PSM method is outlined in Algorithm 1.
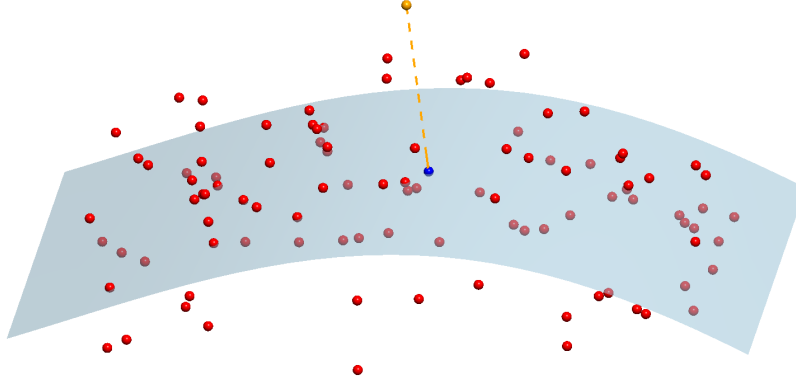
FIG 5. *Visualization of PSM. Original samples $\mathcal{X}$ (red) are on the $\mathcal{M}$. The curve surface (light blue) represents the empirical principal submanifold $\widehat{\psi}_N$. For a selected sample x_i (orange), PSM projects to $\widehat{\psi}_N$ through the estimated projection direction (orange dashed line) to fit the corresponding $\widetilde{x}_i$ (blue) on the $\widehat{\psi}_N$.*

1. *Augmentation*: To account for the periodicity of data on the torus, at the initialization stage, we augment $x_i \in \mathcal{X}$ to a $2D$-dimensional space as:

$$(3) \qquad x_i^{(j)} \mapsto y_i^{(j)} = (\cos(x_i^{(j)}), \sin(x_i^{(j)})), \ j = 1, \ldots, D.$$

2. *Iteration*: In the new $2D$-dimensional space, the algorithm iteratively updates x_i^t to x_i^{t+1} such that x_i^{t+1} is closer to ψ^*. At stage t, the neighbours $\mathcal{N}_{r_s}(x_i^t) = \mathbb{B}(x_i^t, r_s) \cap \mathcal{X}$ within the neighbourhood radius r_s and the neighborhood sample matrix $\mathcal{N}$ of x_i^t as:

$$(4) \qquad \mathcal{N} = (x_1, \ldots, x_{n_\mathcal{N}})^T, \ x_j \in \mathcal{N}_{r_s}(x_i^t), \ j = 1, \ldots, n_\mathcal{N},$$

where $n_\mathcal{N}$ is the number of elements in $\mathcal{N}_{r_s}(x_i^t)$.

If there exists a sufficient number of neighbors $\mathcal{N}_{r_s}(x_i^t)$, we can calculate the local mean $u_{x_i^t}$:

$$u_{x_i^t} = \frac{1}{n_\mathcal{N}} \sum_{x_j \in \mathcal{N}_{r_s}(x_i^t)} x_j,$$

and define the local mean matrix U:

$$(5) \qquad U = \mathbf{1} u_{x_i^t}^T,$$

where $\mathbf{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^{n_\mathcal{N}}$.

Then, given the dimension d of data to fit, where $1 \leq d \leq d^*$, we calculate singular vector matrix Ω_d, defined as:

$$(6) \qquad \Omega_d = (\omega_1, \omega_2, \ldots, \omega_{2D-d})^T,$$

where ω_i is the singular vector corresponding to the smallest i singular values of the local zero-centered neighborhood matrix $\mathcal{N} - U$.

Based on Ω_d, we can determine the projection direction towards to the empirical principal submanifold $\widehat{\psi}_N$:

$$(7) \qquad v_d = \Omega_d^T \Omega_d (x_i^t - u_{x_i^t}).$$

Algorithm updates x_i^{t+1} by $x_i^t - \lambda v_d$ using a learning rate λ and continue until movement falls below a predefined convergence threshold ρ: $\|x_i^{t+1} - x_i^t\|_2 \leq \rho$, ensuring stability of the algorithm.

Algorithm 1: PSM Algorithm

Result: Compute the empirical principal submanifold $\widehat{\psi}_N$ based on the sample set $\mathcal{X}$.
Input: Sample set $\mathcal{X}$, maximum number of iterations $maxiter$, learning rate λ, initial neighbourhood radius r_s, submanifold dimension d, convergence threshold ρ.
Initialization: Augment each $x_i \in \mathcal{X}$ by (3);
for $i \leftarrow 1$ **to** N **do**

    Init $x_i^0 = x_i$;
    for $t \leftarrow 1$ **to** $maxiter$ **do**

        Calculate $\mathcal{N} - U$ with the neighbourhood $\mathcal{N}_{r_s}(x_i^{t-1})$ of x_i^{t-1} by (4) and (5);
        Calculate the singular vector matrix Ω_d by (6) and the projection direction v_d by (7);
        Update x_i^t by $x_i^{t-1} - \lambda v_d$;
        if $\|x_i^t - x_i^{t-1}\|_2 \le \rho$ **then**
          | break;
        end

    end
    $\widetilde{x}_i = x_i^t$;
    Normalize $\widetilde{x}_i$ by (8) and map $\widetilde{x}_i$ back to $\mathbb{T}^D$ by (9).

end
Output: The fitted samples $\widetilde{\mathcal{X}} = \{\widetilde{x}_i\}_{i=1}^N$ as the d-dimentional $\widehat{\psi}_N$.

---

3. *Output*: Following the iterative process, the fitted $\widetilde{y}_i$ will be normalized to ensure the fitted samples remain on the intended torus $\mathbb{T}^D$. For each consecutive pair $\widetilde{y}_i^{(j)} = (\widetilde{y}_i^{(j),1}, \widetilde{y}_i^{(j),2})$, we normalize as:

$$(8) \qquad (\widetilde{x}_i^{(j),1}, \widetilde{x}_i^{(j),2}) = \frac{\widetilde{y}_i^{(j)}}{\|\widetilde{y}_i^{(j)}\|_2}, \quad \text{for } j = 1, \dots, D.$$

Then, we can map fitted $\widetilde{x}_i$ from the $2D$-dimensional space back to the torus $\mathbb{T}^D$ with mapping as:

$$(9) \qquad \arctan\left(\frac{\widetilde{x}_i^{(j),1}}{\widetilde{x}_i^{(j),2}}\right) \mapsto \widetilde{x}_i^{(j)}, \quad \text{for } j = 1, \dots, D.$$

The algorithm outputs the fitted samples $\widetilde{\mathcal{X}}^d = \{\widetilde{x}_i^d\}_{i=1}^N$, with $\widetilde{x}_i^d \in \widehat{\psi}_N \subset \mathbb{R}^D$. For convenience, without any confusion, we will omit the superscript d of $\widetilde{\mathcal{X}}^d$ from now on including the Algorithm 1, although it still denotes that the fitted sample $\widetilde{\mathcal{X}}$ is on the d-dimensional submanifold. If necessary, we will use $\widetilde{x}_i^d$ and $\widetilde{x}_i$ interchangeably.

2.4. *Clustering based on PSM.* Based on the fitted $\widetilde{\mathcal{X}}$, DBSCAN is applied for clustering. We note that any clustering method can be suitable for $\widetilde{\mathcal{X}}$. Here, we select DBSCAN, which is especially useful for identifying arbitrary-shaped clusters. We refer to this combined approach as PSM-DBSCAN.

There are two parameters associated with DBSCAN: r_C, which defines the neighborhood radius; and η, which is the threshold for density.

PSM-DBSCAN starts with an initial assignment for each sample from the samples $\widetilde{\mathcal{X}}$ to any of three sets: $\mathscr{C}$, $\mathscr{B}$, or $\mathscr{O}$ using the following rule:

$$\widetilde{x}_i \in \begin{cases} \mathscr{C}(r_C, \eta), & \text{if } \#(\mathcal{N}_{r_C}(\widetilde{x}_i)) \ge \eta, \\ \mathscr{B}(r_C, \eta), & \text{if } \#(\mathcal{N}_{r_C}(\widetilde{x}_i)) < \eta \ ; \ \exists \, \widetilde{x}_j \in \mathcal{N}_{r_C}(\widetilde{x}_i), \text{ s.t. } \#(\mathcal{N}_{r_C}(\widetilde{x}_j)) \ge \eta, \\ \mathscr{O}(r_C, \eta), & \text{otherwise.} \end{cases}$$

Based on the sets assigned to $\widetilde{\mathcal{X}}$, DBSCAN forms clusters $C_1, \ldots, C_r$ starting from each sample $\widetilde{x}_i \in \mathcal{C}$. Then, expanding cluster set $\mathcal{C}(\widetilde{x}_i)$ is defined as follows:

$$\mathcal{C}(\widetilde{x}_i) = \{\widetilde{x}_j \mid \widetilde{x}_j \in \mathcal{D}(\widetilde{x}_i), j = 1, \ldots, N\},$$

where we define the density-reachable set $\mathcal{D}(\widetilde{x}_i)$ of $\widetilde{x}_i$ as:

$$\mathcal{D}(\widetilde{x}_i) = \{\widetilde{x}_j \in D \mid \exists p_1, \ldots, p_n,$$
$$\text{s.t. } p_1 = \widetilde{x}_i, \ p_n = \widetilde{x}_j, \ p_{m+1} \in \mathcal{N}_{r_C}(p_m), \ p_{m+1} \in \mathcal{C}, \ for \ m = 1, \ldots, n\}.$$

Once the cluster sets are constructed through above density-based expansion process, each $\widetilde{x}_i$ is then assigned a label l_i, and cluster set $\mathcal{C}_h$ for cluster C_h can be defined as:

$$\mathcal{C}_h = \{\widetilde{x}_i \mid l_i = C_h\}, \ h = 1, \ldots, r.$$

We can identify the outliers set $\mathcal{C}_o$, which defined as:

$$\mathcal{C}_o = \left\{ \widetilde{x}_i \in \widetilde{\mathcal{X}} \,\middle|\, \widetilde{x}_i \notin \mathcal{C} \text{ and } \nexists \widetilde{x}_j \in \mathcal{N}_{r_C}(\widetilde{x}_i), \text{ s.t. } \widetilde{x}_j \in \mathcal{C} \right\}.$$

2.5. *Multiscale correction.* Wiechers et al. (2023) introduce the multiscale correction method that integrates RNA correction at both microscopic and mesoscopic scales. Based on the cluster set $\mathcal{C}_1, \ldots, \mathcal{C}_r$ generated by PSM-DBSCAN, which efficiently identifies the different RNA suites, we propose a method named as PSM-DBSCAN-MC (Figure 6).
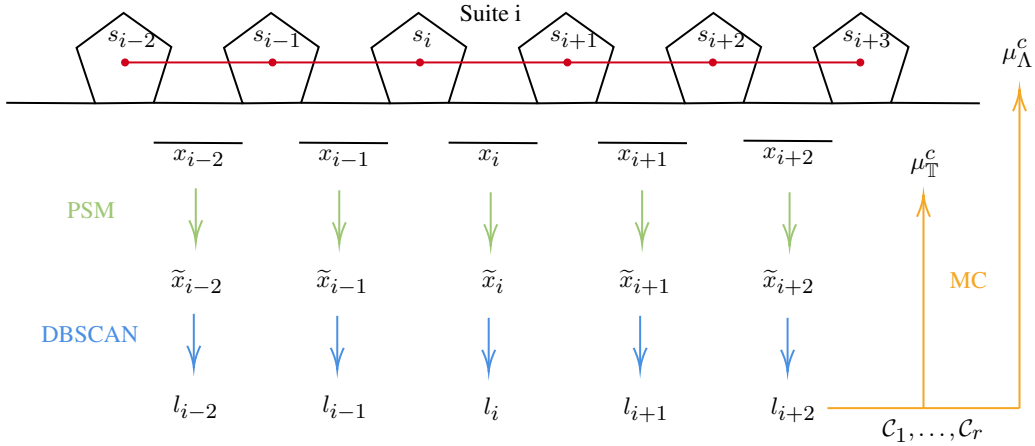

FIG 6. *Framework of PSM-DBSCASN-MC for RNA correction at microscopic and mesoscopic scale.*

Multiscale RNA correction starts with a clash suites sample set $\mathcal{X}^c$. For each clash suite x_i^c, its corresponding mesoscopic shape $[S^c]$ is defined as (1) by the landmark configuration matrix $S^c = (s_1, \ldots, s_{2k+2})^T$. Based on a clash-free suites sample set $\mathcal{X}$, we address each clash suite x_i^c as follows. First, we define a neighborhood $\mathcal{N}^c$ of x_i^c consisting of the κ-nearest clash-free suites in $\mathcal{X}$. This neighborhood is determined by the mesoscopic shape space distance (2):

$$\mathcal{N}^c := \{x_i \in \mathcal{X} : \#\{x_j \in \mathcal{X} : d_\Lambda([S_j], [S^c]) \leq d_\Lambda([S_i], [S^c])\} \leq \kappa\},$$

where $\#$ is the number of set elements and $[S_i]$ denotes corresponding mesoscopic shape $[S]$ of training suite x_i. Then, we identify the dominant class $\mathcal{C}_{h^c}$ of x_i^c within its neighborhood $\mathcal{N}^c$:

$$h^c \in \arg\max_{h=1,\ldots,r} \#(\mathcal{C}_h \cap \mathcal{N}^c).$$

After establishing $\mathcal{C}_{h^c}$, Fréchet mean $\mu_{\mathbb{T}}^c$ is used to correct the microscopic shape, representing the geometric center of the clash-free suites with the same cluster $\mathcal{C}_{h^c}$ within neighborhood $\mathcal{N}^c$:

$$\mu_{\mathbb{T}}^c \in \arg\min_{x \in \mathbb{T}^7} \sum_{x_i \in \mathcal{C}_{h^c} \cap \mathcal{N}^c} \mathrm{d}_{\mathbb{T}^7}(x, x_i)^2.$$

Fréchet mean $\mu_{\Lambda}^c = (\mu_1, \ldots, \mu_{2k+2})^T$ is used to correct the mesoscopic shape:

$$\mu_{\Lambda}^c \in \arg\min_{[S] \in \Lambda_3^{2k+2}} \sum_{i \in \{i | x_i \in \mathcal{C}_{h^c} \cap \mathcal{N}^c\}} \mathrm{d}_{\Lambda}([S], [S_i])^2.$$

The adjusted mesoscopic shape is then calculated by modifying the distances between the core sugar rings to the new length, determined through the Fréchet mean $\mu_{\mathbb{T}}^c$, while satisfying the constrained condition set $\{[S] = [s_1, \ldots, s_{2k+2}] \in \Lambda_3^{2k+2} : \|s_1 - s_{2k+2}\|_2 = \theta_1, \|s_{k+2} - s_{k+1}\|_2 = \theta_2\}$. The first condition adjusts the distance between the starting and ending landmarks to match the corresponding distance in the original mesoscopic shape, thereby ensuring compatibility with the larger RNA strand. The second condition aligns the distance between the central landmarks with the length of the revised suite. This adjustment involves an orthogonal projection of the size-and-shape W^* of the Fréchet mean μ_{Λ}^c for $i \in \{1, k+1, k+2, 2k+2\}$:

$$w_1^* = p\mu_1 + (1-p)\mu_{2k+2}, \quad w_{2k+2}^* = p\mu_{2k+2} + (1-p)\mu_1,$$

$$w_{k+1}^* = q\mu_{k+1} + (1-q)\mu_{k+2}, \quad w_{k+2}^* = q\mu_{k+2} + (1-q)\mu_{k+1},$$

where

$$p = \frac{1}{2}\left(1 + \frac{\|s_{2k+2} - s_1\|_2}{\|\mu_{2k+2} - \mu_1\|_2}\right), \quad q = \frac{1}{2}\left(1 + \frac{\|s_{k+2} - s_{k+1}\|_2}{\|\mu_{k+2} - \mu_{k+1}\|_2}\right).$$

and $w_i^* = \mu_i$ for otherwise.

PSM-DBSCASN-MC outputs a corrected microscopic shape $\mu_{\mathbb{T}}^c$ and its corresponding corrected mesoscopic shape $[W^*]$ to reflect the new, clash-free configuration of the RNA suites.

3. Simulations. We conduct a series of simulation experiments divided into two parts. In the first part, we systematically compare the dimensionality reduction performance of PSM and tPCA for data situated on the torus $\mathbb{T}^D$. In the second part, we compare the clustering performance of PSM-DBSCAN with other clustering methods.

3.1. *Experiment 1 (dimensionality reduction).* A series of subexperiments $(a - l)$ is conducted to compare PSM with tPCA. In each subexperiment, we generate 3000 samples on the torus $\mathbb{T}^D$, ranging from $D = 2, \ldots, 7$. Each subexperiment is repeated 10 times to assess the stability and robustness. For each D-dimensional torus, two subexperiments are conducted to generate samples $\mathcal{X}$ around a submanifold ψ with intrinsic dimension 1 and $D - 1$. Two types of submanifolds are considered: linear and non-linear. For the linear case, the samples $\mathcal{X}$

are generated as follows: all components $x_i^{(j)}$ (for $j = 1, \ldots, D$) are set equal to a value drawn from a Uniform$(0, 2\pi)$ sample. For the non-linear case, a relationship among the different components is established via a sinusoidal function: $x_i^{(m)} = \sin(2x_i^{(j)}) + \pi$. Gaussian noise $\epsilon \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$ is added to x_i, with the covariance matrix $\Sigma = \sigma^2 I$ ($\sigma = 0.2$).

Traditional metrics often fail to accurately assess the effectiveness of dimensionality reduction methods on the manifold. To address this challenges, we use two novel metrics: approximation error AE and proportion of information retained P_d.

DEFINITION 3.1. *The approximation error (*AE*) for the fitted samples $\widetilde{\mathcal{X}}$ is defined as:*

$$\text{AE} = \sum_{i=1}^{N} \inf \mathrm{d}_{\mathbb{T}^D}(\widetilde{x}_i, \psi).$$

Table 1 reports the approximation error, reflecting how well the low-dimensional data aligns with the underlying submanifold ψ. Overall, the approximation errors of the PSM method are consistently lower than those of tPCA for low-dimensional representation, especially when $d = 1$, which demonstrates its superior performance across most subexperimental scenarios. Specifically, the approximation errors of PSM range from 0.01 to 0.29, whereas tPCA errors range from about 0.11 to 0.43, clearly indicating that PSM significantly reduces the approximation error. This suggests that the low-dimensional representation fitted by PSM is closer to the intrinsic true structure of the samples than tPCA. Additionally, the standard deviations (shown in parentheses) for PSM are generally smaller than those for tPCA, indicating that PSM provides more stable and reliable approximations.

DEFINITION 3.2. *Given the d-dimensional fitted samples $\widetilde{\mathcal{X}}^d = \{\widetilde{x}_i^d\}_{i=1}^{N}$ and their corresponding $(d-1)$-dimensional fitted samples $\{\widetilde{x}_i^{d-1}\}_{i=1}^{N}$, the residuals R_d is defined as:*

$$\text{R}_d = \frac{1}{N} \sum_{i=1}^{N} \mathrm{d}_{\psi^d}(\widetilde{x}_i^d, \widetilde{x}_i^{d-1})^2,$$

where d_{ψ^d} represents the geodesic distance on the d-dimensional submanifold ψ^d.

DEFINITION 3.3. *The proportion of information retained P_d in dimension d is defined as:*

$$\text{P}_d = \frac{\text{R}_d}{\sum_{j=2}^{D} \text{R}_j}.$$

TABLE 1

Comparison of approximation error between PSM and tPCA. Standard deviations computed based on 10 repetitions of simulations are shown in parentheses, and all standard deviations are scaled by 10^{-4}.

		PSM					
D	Exp1	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
2	a	**.01**(11)					
	b	**.05**(58)					
3	c	**.02**(43)	**.15**(24)				
	d	**.07**(50)	**.09**(111)				
4	e	**.03**(24)	.17(43)	**.24**(30)			
	f	**.11**(75)	**.11**(48)	.19(56)			
5	g	**.03**(9)	**.18**(37)	**.24**(48)	**.29**(17)		
	h	**.15**(158)	**.12**(101)	.19(26)	**.23**(21)		
6	i	**.01**(14)	**.07**(31)	**.11**(26)	**.14**(17)	**.17**(9)	
	j	**.11**(29)	.13(31)	.11(39)	.12(7)	**.15**(12)	
7	k	**.02**(17)	**.07**(36)	**.12**(11)	**.15**(28)	**.17**(15)	**.19**(11)
	l	**.12**(32)	.13(37)	.12(34)	.13(36)	**.14**(15)	**.16**(11)
		tPCA					
D	Exp1	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
2	a	.29(40)					
	b	.43(978)					
3	c	.26(21)	.15(28)				
	d	.22(295)	.14(36)				
4	e	.30(48)	.17(40)	.25(24)			
	f	.23(429)	.15(51)	**.17**(46)			
5	g	.29(26)	.22(57)	.25(32)	.32(19)		
	h	.16(363)	.15(36)	**.17**(39)	.25(26)		
6	i	.28(48)	.22(40)	.13(19)	.16(14)	.19(11)	
	j	.17(42)	**.12**(14)	.11(10)	.12(11)	.16(18)	
7	k	.28(69)	.28(53)	.15(66)	.17(27)	.19(16)	.21(18)
	l	.17(20)	**.11**(12)	**.11**(7)	.13(5)	.16(10)	.19(14)

Table 2 reports the proportion of information preserved in the low-dimensional representation. Overall, it is clear that PSM consistently achieves higher proportions of retained information in the 1-dimensional representation (ranging from 0.81 to 0.99) compared to tPCA (ranging from 0.43 to 0.94). Specifically, PSM outperforms tPCA by approximately 5% to 50% in terms of information preservation, demonstrating that PSM has a clear advantage in effectively preserving the primary information in low-dimensional representations across various manifold dimensions D.

In summary, PSM produces a more accurate low-dimensional representation of the original data on the torus $\mathbb{T}^D$, which is closer to the intrinsic low-dimensional geometric structure of the data while preserving more feature information. Therefore, PSM offers significant advantages over tPCA, making it particularly valuable for downstream tasks.

TABLE 2
Comparison of proportion of information retained between PSM and tPCA. Standard deviations computed based on 10 repetitions of simulations are shown in parentheses, and all standard deviations are scaled by 10^{-4}.

PSM							
D	Exp1	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$
2	a	**.81**(218)	.19(218)					
	b	**.90**(52)	.10(52)					
3	c	**.98**(4)	.01(4)	.01(8)				
	d	**.96**(14)	.02(14)	.02(29)				
4	e	**.99**(1)	.01(1)	.00(2)	.00(1)			
	f	**.96**(7)	.02(6)	.01(18)	.01(21)			
5	g	**.99**(1)	.01(1)	.00(1)	.00(1)	.00(3)		
	h	**.95**(3)	.02(2)	.01(4)	.01(9)	.01(9)		
6	i	**.99**(0)	.01(0)	.00(0)	.00(1)	.00(0)	.00(2)	
	j	**.95**(2)	.01(1)	.01(3)	.01(4)	.01(14)	.01(15)	
7	k	**.99**(0)	.01(0)	.00(0)	.00(0)	.00(1)	.00(5)	.00(22)
	l	**.95**(2)	.01(0)	.01(1)	.01(3)	.01(6)	.00(23)	.01(22)

tPCA							
D	Exp1	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$
2	a	.44(253)	.56(253)					
	b	.43(356)	.57(356)					
3	c	.94(3)	.05(6)	.01(7)				
	d	.82(9)	.15(67)	.03(63)				
4	e	.92(3)	.06(31)	.01(32)	.01(62)			
	f	.82(32)	.13(21)	.03(135)	.02(89)			
5	g	.89(2)	.06(11)	.03(11)	.01(16)	.01(16)		
	h	.80(16)	.11(15)	.05(26)	.02(104)	.02(66)		
6	i	.90(0)	.06(1)	.02(2)	.01(7)	.01(21)	.00(25)	
	j	.83(2)	.10(2)	.05(1)	.01(7)	.01(21)	.00(32)	
7	k	.90(0)	.06(0)	.03(2)	.01(0)	.00(3)	.00(3)	.00(0)
	l	.83(2)	.10(2)	.05(1)	.01(2)	.01(4)	.01(24)	.00(32)

3.2. *Experiment 2 (clustering).* A series of subexperiments $(a-l)$ is conducted to evaluate the performance of PSM-DBSCAN. In each subexperiment, we generate 1000 samples $\mathcal{X}$ for each distinct submanifold ψ on the torus $\mathbb{T}^D$, ranging from $D=2,\ldots,7$. The true label of each sample x_i is determined by the submanifold to which it belongs. Specifically, subexperiment $(a-e)$, (g), (i) and (k) contain 3 submanifolds, subexperiment (f), (h), (j) and (l) contain 5 submanifolds. Each subexperiment is repeated 10 times to assess the stability and robustness. Two types of submanifolds are considered: linear and non-linear. For the linear case, samples $\mathcal{X}$ are generated on the lines parameterized by the equation: $\mathbf{x}_\psi = \mathbf{p}_\psi + z(\mathbf{q}_\psi - \mathbf{p}_\psi)$ for $z \in [0,1]$. We randomly generate the starting point $\mathbf{p}_\psi$ and the ending point $\mathbf{q}_\psi$. For the non-linear case, a relationship among the different components $x_i^{(m)}$ is established via a sinusoidal function: $x_i^{(m)} = \alpha_\psi \sin(\beta_\psi x_i^{(j)}) + \gamma_\psi$, where $\alpha_\psi \in \{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}, \beta_\psi \in \{\frac{1}{2}, 1, 2\}, \gamma_\psi \in \{-\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi\}$. Gaussian noise $\epsilon \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$ is added to x_i, with the covariance matrix $\Sigma = \sigma^2 I$ ($\sigma \in \{0.3, 0.4, 0.5\}$).

Figure 7 illustrates the clustering performance of various methods. Here, we present the subexperiments $(i-l)$ with $D=6$ and 7. The results for the remaining subexperiments are provided in Appendix C. In this simulation, we select several clustering methods to systematically evaluate the performance of PSM-DBSCAN. First, we choose DBSCAN to assess the impact of PSM for dimensionality reduction on the clustering. Next, we adopt tPCA-DBSCAN to examine the impact of using tPCA as an alternative dimensionality reduction method. We also employ MINT-AGE to contrast the clustering performance achieved via tPCA with other clustering approaches. Finally, we include two classic clustering methods: hierarchical clustering (HC) and spectral clustering (SC) to compare with these well-established methods.

Figure 7(b) presents the clustering performance of DBSCAN. This indicates that DBSCAN fails to accurately identify the underlying submanifold structure in these high-dimensional settings, even though it is designed for density-based and shape-sensitive clustering. Specifically, in subexperiment (i), although DBSCAN correctly clusters one of the submanifolds (in yellow), it fails to separate the remaining two submanifolds (in blue and green), as these samples are too close to each other for DBSCAN to distinguish effectively on the high-dimensional torus. Figure 7(c) shows that tPCA-DBSCAN also cannot cluster accurately based on the underlying geometric structure and groups samples from different submanifolds into the same cluster. In contrast, as highlighted in the red box, Figure 7(d) demonstrates that PSM-DBSCAN can accurately cluster samples according to the submanifold on which they reside in each subexperiment. MINT-AGE, as shown in Figure 7(e), produces fine clustering but yields an excessively large number of clusters, significantly exceeding the actual number of underlying submanifolds. Although MINT-AGE attempts refine clustering, it still does not guarantee accurate identification of the underlying submanifold structure. In the subexperiment (i), (k) and (l), MINT-AGE incorrectly groups data from distinct submanifolds into a single cluster. Finally, in Figure 7(f) and (g), the two classical clustering methods have similar performances. It can be observed that the clustering primarily relies on the Euclidean distances between the samples, grouping those relatively close to each other into the same cluster. Within each cluster, the samples typically exhibit a locally concentrated distribution, and the results indicate that the data are partitioned into several regions where the samples are spatially proximate. However, the underlying geometric structures of the samples are not taken into account, rendering these methods ineffective in distinguishing samples from different submanifolds.

In order to further quantitatively compare the performance of different clustering methods, we use two classical performance metrics for clustering: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), which are defined as follows:

DEFINITION 3.4.   *Given the samples $\mathcal{X}$ divided into two different sets of clusters $\mathcal{U}$ and $\mathcal{V}$, where $\mathcal{U}$ is the set of true labels and $\mathcal{V}$ is the set of labels produced by a clustering algorithm. Define the contingency table $\mathbf{L}$ where n_{ij} is the number of samples that are in both cluster $\mathcal{U}_i$ in $\mathcal{U}$ and cluster $\mathcal{V}_j$ in $\mathcal{V}$. Let a_i and b_j be the sums of the rows and columns of $\mathbf{L}$, respectively:*

$$a_i = \sum_j n_{ij}, \quad b_j = \sum_i n_{ij}.$$

The Rand Index (RI) is then calculated as:

$$\mathrm{RI} = \frac{\sum_{ij} \binom{n_{ij}}{2} + \sum_{ij} \binom{a_i - n_{ij}}{2} + \sum_{ij} \binom{b_j - n_{ij}}{2}}{\binom{n}{2}}.$$

The Expected Rand Index (ERI) under random labeling is given by:

$$\mathrm{ERI} = \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}^2}.$$

ARI is then defined as the normalized difference between RI and ERI:

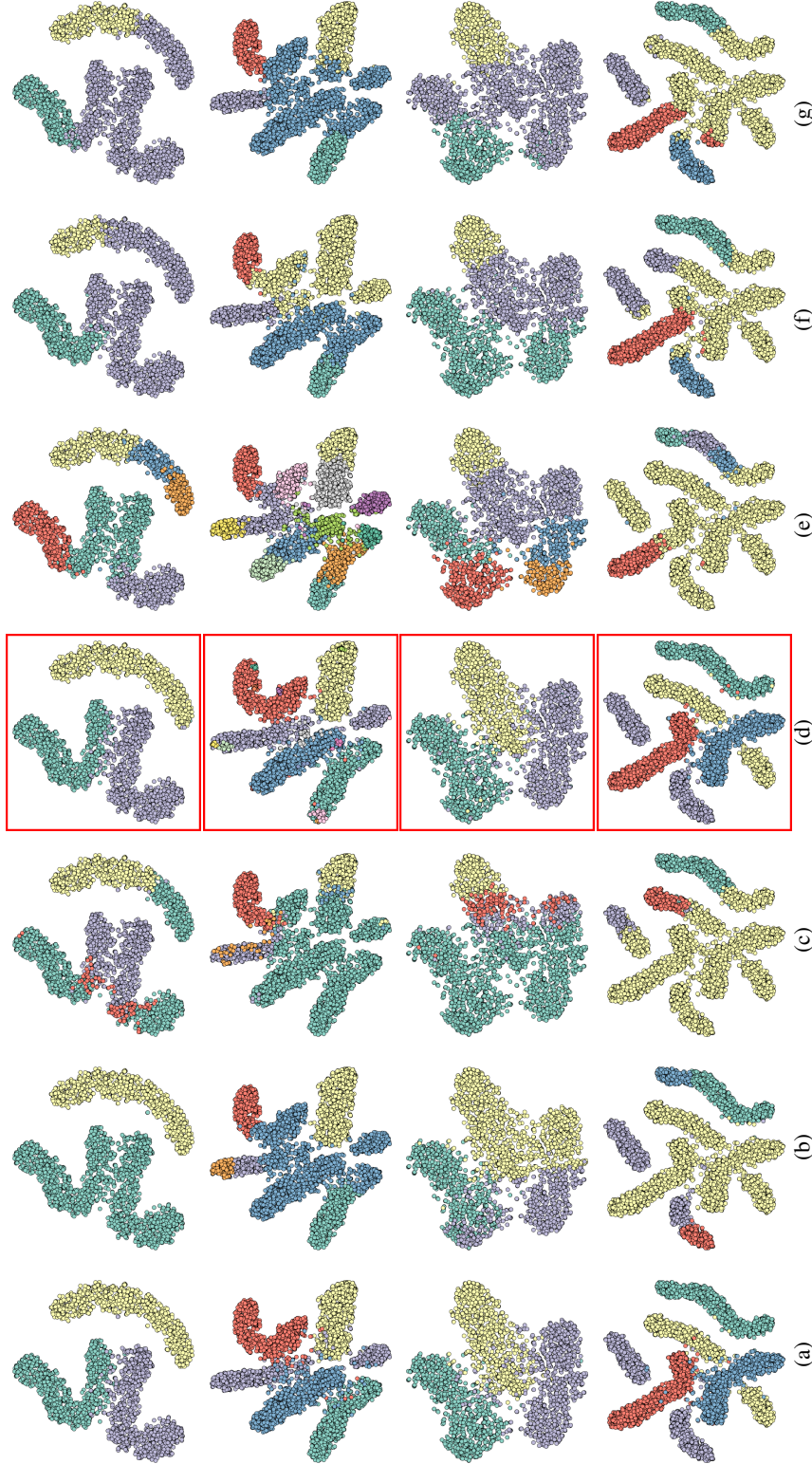$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{ERI}}{\max(\mathrm{RI}) - \mathrm{ERI}}.$$

FIG 7. *An intuitive illustration of clustering performance for subexperiment* $(i - l)$. *Samples are shown in the angle space and visualized using t-SNE.* (a) *shows true labels of simulated samples generated by the different submanifold,* (b) *shows the clusters obtained through DBSCAN method,* (c) *shows the clusters of tPCA-DBSCAN,* (d) *shows the clusters of PSM-DBSCAN (highlighted in red box),* (e) *shows the clusters of MINT-AGE,* (e) *shows the clusters of hierarchical clustering and* (f) *shows the clusters of spectral clustering. Different clusters are represented by different colors. The subexperiments are correspond to* $D = 6$ *(rows 1 and 2) and* $D = 7$ *(rows 3 and 4).*

DEFINITION 3.5. *Define $H(\mathcal{U})$ and $H(\mathcal{V})$ as the entropies of $\mathcal{U}$ and $\mathcal{V}$, respectively, and $I(\mathcal{U}; \mathcal{V})$ as the mutual information between $\mathcal{U}$ and $\mathcal{V}$:*

$$H(\mathcal{U}) = -\sum_i \frac{\#(\mathcal{U}_i)}{N} \log \frac{\#(\mathcal{U}_i)}{N}, \quad H(\mathcal{V}) = -\sum_j \frac{\#(\mathcal{V}_j)}{N} \log \frac{\#(\mathcal{V}_j)}{N},$$

$$I(\mathcal{U}; \mathcal{V}) = \sum_{i,j} \frac{\#(\mathcal{U}_i \cap \mathcal{V}_j)}{N} \log \frac{N \#(\mathcal{U}_i \cap \mathcal{V}_j)}{\#(\mathcal{U}_i) \#(\mathcal{V}_j)}.$$

NMI is then defined as the ratio of mutual information and the mean of the entropies of $\mathcal{U}$ and $\mathcal{V}$:

$$\text{NMI} = \frac{2I(\mathcal{U}; \mathcal{V})}{H(\mathcal{U}) + H(\mathcal{V})}.$$

Table 3 provides a quantitative comparison of clustering performance between PSM-DBSCAN and other clustering methods using ARI and NMI. Table 3 highlights that PSM-DBSCAN consistently achieves the highest ARI and NMI in all subexperiments across different dimensional scenarios, indicating that the clustering produced by PSM-DBSCAN is more closely aligned with the true labels and captures the underlying data structure more effectively. Overall, PSM-DBSCAN surpasses the second-best clustering methods by approximately from 20% to 50% in ARI and from 10% to 40% in NMI, demonstrating a significant and consistent advantage in clustering performance compared with other methods. In contrast, although standalone DBSCAN shows high ARI and NMI in the subexperiments with $D = 2$, as the dimension D increases, it achieves low ARI and NMI while PSM-DBSCAN still attains high ARI and NMI. Additionally, DBSCAN exhibits substantial variability in performance, as indicated by the large standard deviations (shown in parentheses), which suggests that this method lacks consistency and robustness. This underscores the limited ability of DBSCAN to handle high-dimensional scenarios and the necessity of dimensionality reduction via PSM. Compared to DBSCAN, tPCA-DBSCAN consistently shows lower performance in terms of ARI and NMI, suggesting that dimensionality reduction by tPCA hinders DBSCAN's ability to capture the intrinsic structure. MINT-AGE also falls short in ARI and NMI, reflecting its limited ability to cluster accurately in challenging scenarios, which indicates that the low-dimensional representation obtained through PSM can more effectively guide clustering than that obtained via tPCA. Classical clustering methods, such as hierarchical clustering and spectral clustering, also fail to perform effectively. Hierarchical clustering consistently yields poor ARI and NMI due to its inability to handle the non-linear and high-dimensional features of the data. Similarly, spectral clustering struggles to identify the underlying submanifold structures and to cluster accurately. Compared with these widely used clustering methods, PSM-DBSCAN demonstrates the advantages of clustering on manifolds.

In summary, the simulations suggest that PSM-DBSCAN consistently exhibits improved performance relative to competing methods, demonstrating superior accuracy and robustness in clustering tasks on the torus. By effectively addressing the challenges of high dimensionality and noise through its integration with PSM, PSM-DBSCAN emerges as a powerful and reliable tool for clustering data on manifolds. The observed differences between PSM-DBSCAN and other methods underscore the critical importance of leveraging PSM to enhance clustering accuracy, stability, and adaptability.

TABLE 3
Comparison of clustering performance between DBSCAN, tPCA-DBSCAN, PSM-DBSCAN, MINT-AGE, hierarchical clustering and spectral clustering. Standard deviations computed based on 10 repetitions of simulations are shown in parentheses and scaled by 10^{-2}.

		DBSCAN		tPCA-DBSCAN		PSM-DBSCAN	
D	Exp2	ARI($\uparrow$)	NMI($\uparrow$)	ARI($\uparrow$)	NMI($\uparrow$)	ARI($\uparrow$)	NMI($\uparrow$)
2	a	.88 (10.98)	.87 (5.87)	.56 (0.20)	.73 (0.42)	**.91** (4.56)	**.89** (3.53)
	b	.79 (2.50)	.79 (2.65)	.56 (0.37)	.70 (1.56)	**.95** (4.95)	**.95** (3.86)
3	c	.54 (0.02)	.68 (0.03)	.80 (0.02)	.79 (0.07)	**.93** (0.01)	**.90** (0.01)
	d	.68 (17.09)	.77 (10.21)	.54 (2.85)	.66 (7.00)	**.97** (0.54)	**.95** (0.74)
4	e	.56 (0.25)	.70 (0.41)	.52 (3.93)	.60 (3.96)	**.95** (0.63)	**.92** (0.77)
	f	.71 (4.49)	.77 (3.19)	.34 (10.50)	.49 (10.36)	**.84** (10.08)	**.90** (4.79)
5	g	.63 (16.34)	.74 (9.51)	.45 (4.36)	.55 (2.33)	**.96** (0.69)	**.93** (0.94)
	h	.50 (6.66)	.62 (3.90)	.22 (9.71)	.33 (8.81)	**.82** (7.51)	**.82** (3.62)
6	i	.39 (10.87)	.61 (7.83)	.05 (2.95)	.21 (6.21)	**.85** (4.86)	**.82** (3.43)
	j	.50 (7.74)	.65 (12.98)	.37 (8.51)	.44 (4.42)	**.74** (1.22)	**.84** (0.64)
7	k	.43 (14.61)	.52 (11.30)	.24 (8.39)	.34 (11.60)	**.81** (2.05)	**.78** (0.56)
	l	.22 (9.23)	.31 (12.30)	.09 (0.58)	.26 (2.39)	**.94** (3.01)	**.93** (1.75)

		MINT-AGE		HC		SC	
D	Exp2	ARI($\uparrow$)	NMI($\uparrow$)	ARI($\uparrow$)	NMI($\uparrow$)	ARI($\uparrow$)	NMI($\uparrow$)
2	a	.39 (2.74)	.57 (1.61)	.38 (7.59)	.39 (8.46)	.24 (0.50)	.31 (0.39)
	b	.46 (8.04)	.64 (6.95)	.08 (5.67)	.12 (8.52)	.15 (3.54)	.19 (2.74)
3	c	.37 (4.17)	.55 (3.12)	.67 (4.56)	.71 (5.33)	.77 (9.16)	.77 (7.37)
	d	.33 (6.14)	.56 (2.39)	.19 (8.41)	.25 (6.42)	.38 (10.19)	.40 (8.23)
4	e	.48 (3.50)	.61 (1.54)	.56 (6.77)	.67 (7.56)	.60 (11.82)	.67 (10.55)
	f	.43 (6.51)	.54 (3.90)	.43 (12.90)	.45 (9.86)	.18 (2.23)	.26 (2.35)
5	g	.52 (2.68)	.62 (3.16)	.42 (7.24)	.54 (9.16)	.61 (2.36)	.66 (2.74)
	h	.37 (3.66)	.58 (2.32)	.24 (3.18)	.44 (1.47)	.29 (9.34)	.51 (7.45)
6	i	.39 (4.38)	.57 (4.57)	.30 (5.85)	.46 (7.74)	.28 (0.91)	.46 (1.38)
	j	.35 (5.21)	.58 (4.40)	.39 (9.39)	.58 (7.76)	.32 (0.74)	.57 (0.94)
7	k	.29 (1.37)	.48 (1.08)	.35 (5.85)	.44 (7.74)	.41 (0.91)	.27 (1.38)
	l	.52 (4.92)	.70 (2.50)	.32 (9.39)	.53 (7.76)	.27 (0.74)	.52 (0.94)

4. Application for RNA data. In this section, we apply PSM-DBSCAN for RNA correction and compare its performance with that of the MINT-AGE-CLEAN method. We analyze classical RNA datasets from the Protein Data Bank (Bank, 1971), a widely used bioinformatics database that primarily stores 3-dimensional structural data of biomolecules. These RNA datasets consist of 8,665 suites obtained from high-precision experimental X-ray measurements. Our analysis aims to demonstrate the efficacy of PSM-DBSCAN-MC in enhancing the accuracy and reliability of RNA clustering and clash correction, potentially offering substantial improvements over the MINT-AGE-CLEAN method.

4.1. *Clustering for RNA suites.* To evaluate the effectiveness of PSM-DBSCAN in clustering RNA suites, we conduct a comparative analysis with MINT-AGE. Figure 8 illustrates the clustering performance achieved by these methods, respectively. Both PSM-DBSCAN and MINT-AGE successfully identify the main RNA suites clusters, such as Class 1. However, the clustering results of PSM-DBSCAN are much clearer, with distinct boundaries between clusters, whereas MINT-AGE's results exhibit indistinct boundaries and non-contiguous clusters. Additionally, PSM-DBSCAN automatically detects outliers in a data-driven manner during the clustering process without requiring a predefined outlier proportion as in MINT-AGE, leading to more effective outlier identification. Moreover, PSM-DBSCAN uncovers new RNA suites clusters compared to MINT-AGE, achieving a more refined clustering. For instance, while MINT-AGE broadly clusters RNA suites into Class 3, PSM-DBSCAN subdivides this cluster into two distinct classes: Class 2 and Class 9. This refined clustering highlights the enhanced

capability of PSM-DBSCAN to detect subtle differences among RNA suites by effectively utilizing latent geometric characteristics, thereby achieving superior clustering performance and providing a new perspective on RNA suites clustering.
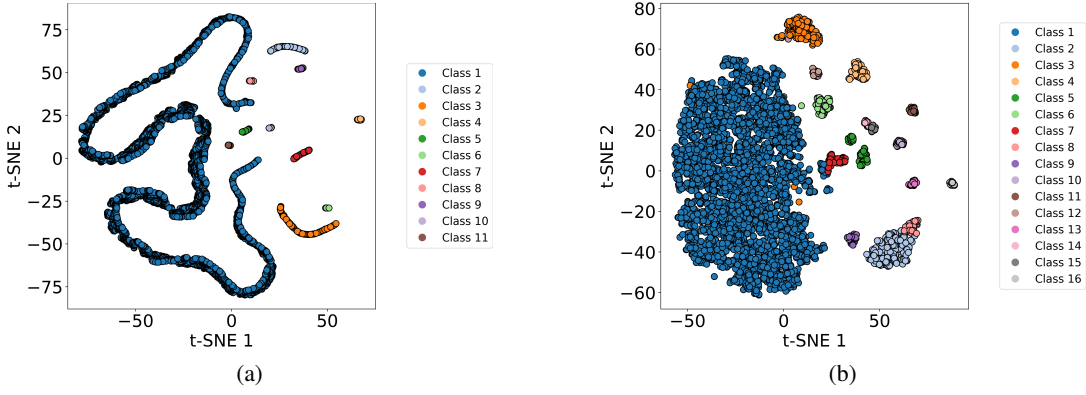

FIG 8. *The visualization of the clusters by t-SNE. (a) shows the clusters of PSM-DBSCAN, and (b) shows the clusters of MINT-AGE. Different clusters are represented by different colors.*

Since the true labels of the RNA suites are unknown in practice, we introduce two metrics for clustering to better compare the performance of the two clustering methods on RNA suites: the silhouette score (SI) and the Davies-Bouldin index (DB), both of which are widely used to assess clustering performance. The SI index measures cluster separation and cohesion, with higher values indicating better and more compact clusters. In contrast, the DB index assesses cluster similarity, with lower values indicating more distinct and well-separated clusters. Higher SI and lower DB values indicate that a clustering method performs well in terms of internal consistency and inter-cluster discrimination, thereby more accurately reflecting the intrinsic structure of the data.

DEFINITION 4.1. *Given a set of samples $\mathcal{X}$. For each sample x_i assigned to cluster C_m:*

- *Let $\mu^a(x_i)$ be the mean distance between x_i and all other samples in the same cluster C_j:*

$$\mu^a(x_i) = \frac{1}{\#(\mathcal{C}_m) - 1} \sum_{\substack{x_j \in \mathcal{C}_j \\ j \neq i}} \mathrm{d}(x_i, x_j),$$

- *Let $\mu^b(x_i)$ be the minimum mean distance between x_i and all samples in any other cluster C_l, for $l \neq m$:*

$$\mu^b(x_i) = \min_{l \neq m} \{\mathrm{d}(x_i, \mathcal{C}_l)\} = \min_{l \neq m} \left\{ \frac{1}{\#(\mathcal{C}_l)} \sum_{x_j \in \mathcal{C}_l} \mathrm{d}(x_i, x_j) \right\}.$$

For each sample x_i, $\mathrm{si}(x_i)$ is defined as:

$$\mathrm{si}(x_i) = \frac{\mu^b(x_i) - \mu^a(x_i)}{\max\{\mu^a(x_i), \mu^b(x_i)\}}.$$

SI *for a clustering method is then defined as the mean of $\mathrm{si}(x_i)$ of all samples:*

$$\mathrm{SI} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{si}(x_i).$$

DEFINITION 4.2. *Given a set of samples $\mathcal{X}$, let r denote the number of clusters. Let C_j be a cluster with centroid u_j. For each cluster C_j:*

- *Let Θ_j be the mean distance of all samples in $\mathcal{C}_j$ to the centroid u_j:*

$$\Theta_j = \frac{1}{\#(\mathcal{C}_j)} \sum_{x_i \in \mathcal{C}_j} \|x_i - u_j\|_2,$$

DB *is defined as the mean similarity between each cluster C_j and its most similar cluster C_m:*

$$\mathrm{DB} = \frac{1}{r} \sum_{j=1}^{r} \max_{l \neq m} \left\{ \frac{\Theta_j + \Theta_m}{\|u_j - u_m\|_2} \right\}.$$

Table 4 provides a quantitative comparison of the performance of PSM-DBSCAN and MINT-AGE using two key clustering evaluation metrics: SI and DB. Table 4 highlights the superior performance of PSM-DBSCAN in both metrics. PSM-DBSCAN achieves a higher SI, outperforming MINT-AGE by approximately 10%, indicating that it forms clusters that are both well-separated and cohesive. In addition, PSM-DBSCAN attains a lower DB, reflecting more compact and distinct clusters than MINT-AGE. These results demonstrate that PSM-DBSCAN effectively groups RNA suites into intrinsic and meaningful clusters, which is critical for RNA correction.

In summary, PSM-DBSCAN provides a more effective tool for clustering RNA data compared to MINT-AGE. It consistently outperforms MINT-AGE in the both SI and DB metrics, demonstrating a superior ability to delineate and compactly represent RNA structures. These differences underscore the advantages of PSM-DBSCAN in uncovering representative and geometrically accurate clustering, thereby laying a strong foundation for subsequent RNA structure correction and analysis. By preserving the geometric properties of the data in the low-dimensional representation, effectively dealing with outliers, and forming well-defined clusters, PSM-DBSCAN establishes itself as a robust and reliable tool for RNA analysis tasks. This capability not only facilitates a deeper understanding of the relationships among RNA suites but also paves the way for accurate, data-driven corrections of structural clashes in RNA suites.

TABLE 4

Comparison of clustering results between PSM-DBSCAN and MINT-AGE for RNA data, using SI and DB as metrics to measure clustering performance.

	PSM-DBSCAN	MINT-AGE
SI($\uparrow$)	**0.65**	0.59
DB($\downarrow$)	**0.51**	0.82

4.2. *Multiscale correction for RNA clashes.* This subsection introduces a multiscale method for RNA correction that bridges the microscopic and mesoscopic scales. The PSM-DBSCAN-MC method leverages the geometric similarities between these scales to resolve clashes in RNA suites. By incorporating the PSM-DBSCAN clustering framework, PSM-DBSCAN-MC provides a robust and comprehensive approach to RNA correction.

Figure 9 illustrates the structures of 73 original RNA clash suites along with their corresponding corrected structures, generated using PSM-DBSCAN-MC and CLEAN-MINT-AGE, respectively. This detailed comparison highlights the differences in how these two methods resolve clashes at the atomic level. Compared to the RNA clash suites corrected by CLEAN-MINT-AGE, PSM-DBSCAN-MC produces clearer 3-dimensional structures

with a significant reduction in noise, offering new insights into RNA correction. Figure 10 presents four 2-dimensional projection slides, obtained via convex hull analysis, that illustrate the positional distribution of atoms in the 3-dimensional structures (see Figure 9) for different methods. It is evident that the distribution of atoms in RNA clash suites is highly chaotic and spans a wide range (in red). After applying the two correction methods, the atom distributions become more concentrated into distinct regions. The correction result of CLEAN-MINT-AGE shows that the atoms remain dispersed over a relatively wide region (in green), whereas the correction by the PSM-DBSCAN-MC method yields several well-defined, concentrated regions of atoms (in blue). A full movie of projection slides at different Z can be found at https://github.com/zhigang-yao/RNA-Clash-Correction/blob/main/RNA_correction_slides.gif. Meanwhile, Figure 11 provides a comparative visualization of changes in dihedral angle distributions resulting from different RNA correction methods. In the corrected 2-dimensional dihedral angle distributions, PSM-DBSCAN-MC produces a more concentrated pattern with fewer deviations (in green). In contrast, although CLEAN-MINT-AGE corrects RNA clashes, the resulting RNA suites exhibit broader and more sporadic dihedral angle distributions (in purple), indicating that some structural noise remains unresolved. In summary, PSM-DBSCAN-MC offers a novel approach to RNA clash correction, resulting in clearer 3-dimensional structures.
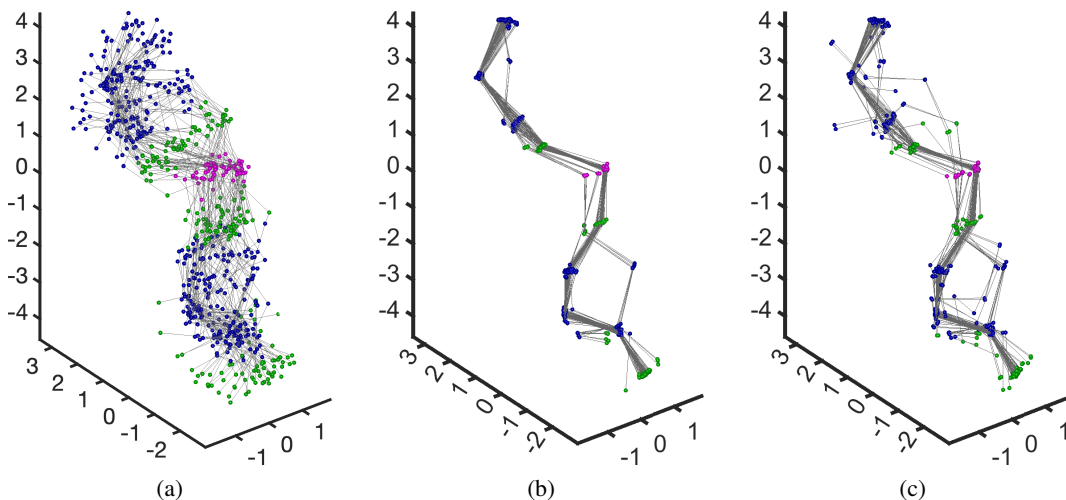

FIG 9. *Illustration of RNA clash suites and corresponding correction. Different colors of points represent different atoms of RNA backbone suites, with carbon (blue), oxygen (green) and phosphorus (pink). (a) is the 73 clash suites, (b) is the clash correction by PSM-DBSCAN-MC and (c) is the clash correction by CLEAN-MINT-AGE.*

5. Discussion. In this paper, we explore the problem of fitting effective and accurate low-dimensional representations of data on the high-dimensional torus and, based on these representations, address the challenges encountered in high-dimensional clustering on the torus. We propose PSM-DBSCAN, a robust framework that integrates PSM-based dimensionality reduction with DBSCAN. PSM accurately fits low-dimensional data on the torus while preserving its geometric features, thereby overcoming the shortcomings of existing methods such as tPCA. By producing accurate low-dimensional representations, PSM-DBSCAN significantly improves clustering performance on the high-dimensional torus.
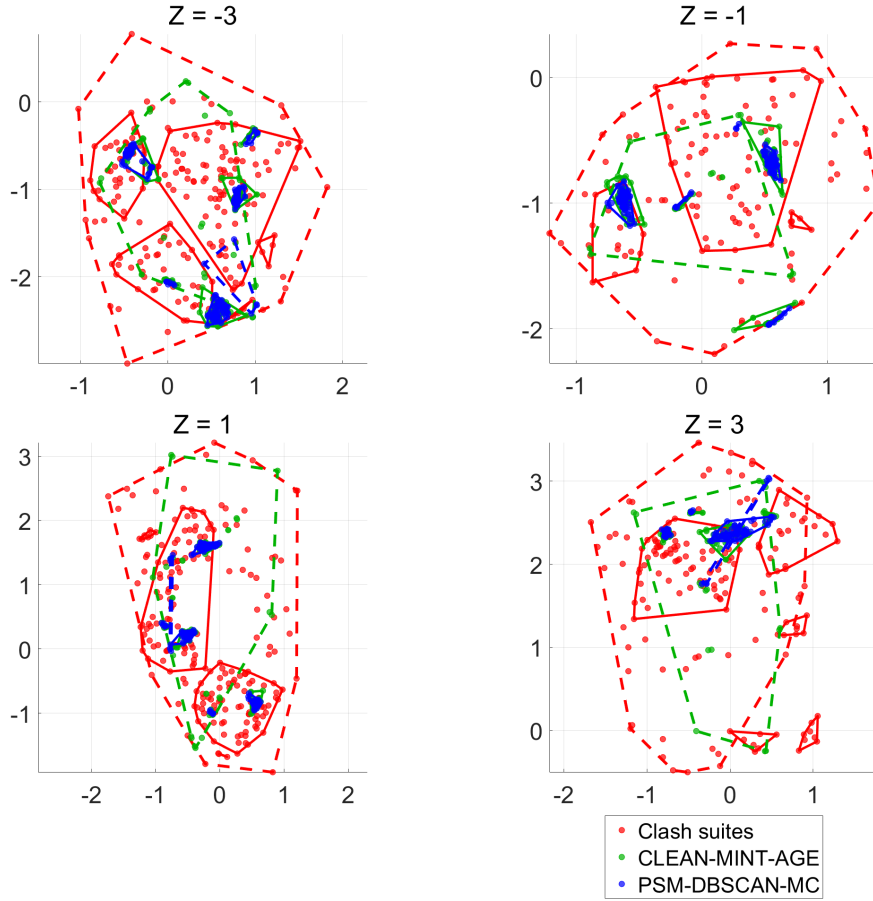
FIG 10. *Illustration of the 2-dimensional projection slides of RNA's 3-dimensional atomic structure. The structure is equally segmented into four intervals along the Z-axis, and atoms within a $Z \pm 1$ range are projected onto the plane. Solid lines delineate the convex hulls of atoms in densely populated regions, while dashed lines outline the convex hulls of dispersed atoms.*

In simulation experiments, we demonstrate the effectiveness of the proposed PSM framework, highlighting its superior capability in capturing intrinsic structures and achieving accurate clustering results on toroidal manifold data. By applying this novel clustering method to RNA suites data, we develop PSM-DBSCAN-MC for multiscale RNA correction. This work bridges the gaps in current RNA structural analysis methodologies by addressing the challenges of dimensionality reduction and clustering in high-dimensional RNA data on the torus. The proposed PSM-DBSCAN-MC not only enhances RNA correction but also provides a versatile framework for broader applications in RNA structure analysis. As our understanding of RNA continues to evolve, the principal submanifold-based analysis framework will be instrumental in advancing both theoretical insights and practical applications in RNA structure.

The current study demonstrates the effectiveness of the proposed PSM framework for toroidal data. Future research could focus on extending the applicability of the PSM framework to data on other types of manifold. In addition, improving the computational efficiency of the PSM framework by implementing parallel computing and leveraging GPU acceleration represents another promising direction for future work.
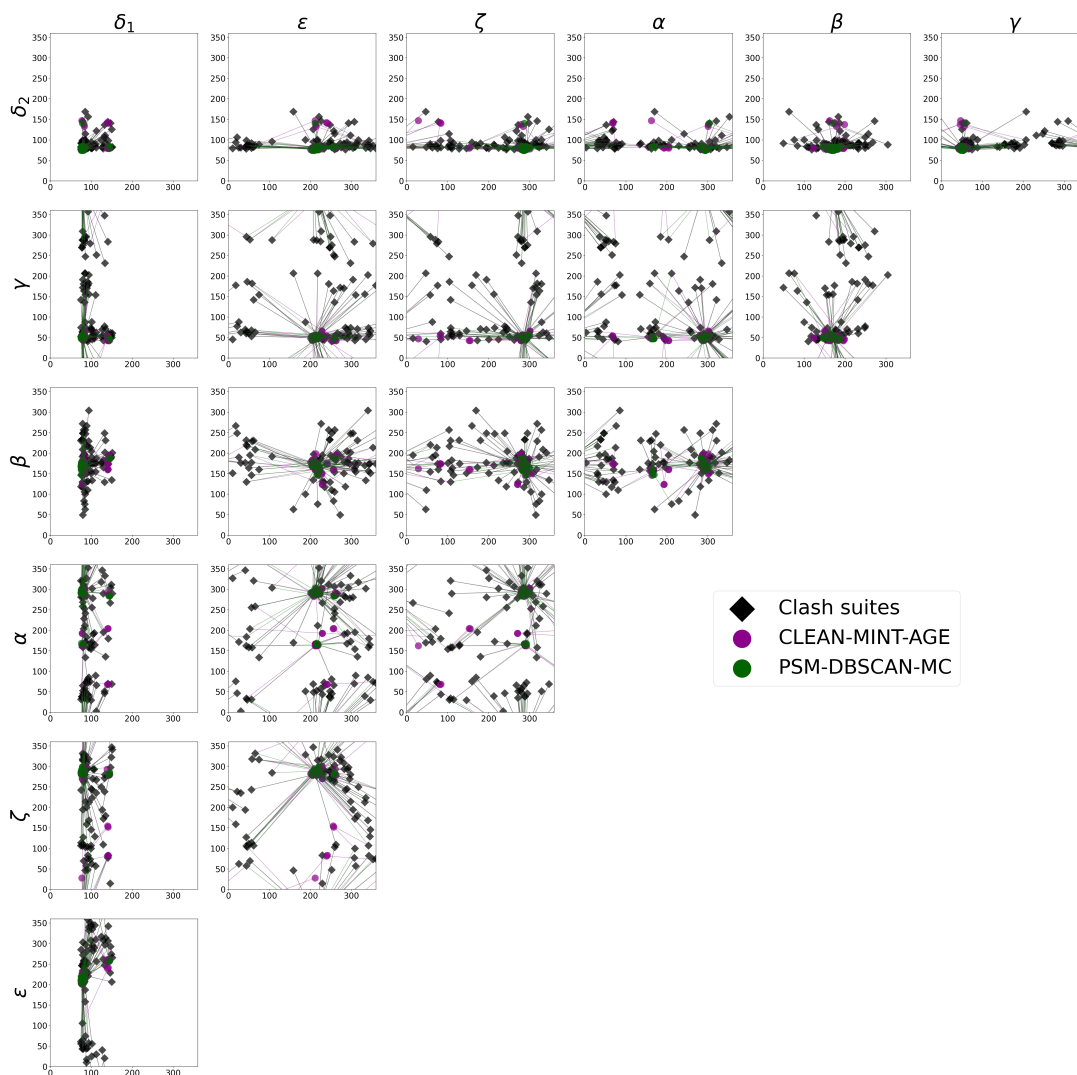
FIG 11. *A comparative analysis through scatterplots showcasing all 2-dimensional dihedral angle pairs of RNA structures. Original RNA clashes are marked in the points (black), indicating the starting point prior to RNA correction. Corrections for RNA clashes using CLEAN-MINT-AGE are depicted in the points (purple), while corrections using PSM-DBSCAN-MC are depicted in the points (green).*

REFERENCES

BANK, P. D. (1971). Protein data bank. *Nature New Biology* **233** 10–1038.

BATSANOV, S. S. (2001). Van der Waals radii of elements. *Inorganic Materials* **37** 871–885.

CHEN, L.-L. (2020). The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nature Reviews Molecular Cell Biology* **21** 475–490.

CHOU, F.-C., SRIPAKDEEVONG, P., DIBROV, S. M., HERMANN, T. and DAS, R. (2013). Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods* **10** 74–76.

DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* **100** 5591–5596.

DOUDNA, J. A. and CECH, T. R. (2002). The chemical repertoire of natural ribozymes. *Nature* **418** 222–228.

DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis: with Applications in R*. John Wiley & Sons.

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *The Annals of Applied Statistics* **12** 1332 – 1359.

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* **96** 226–231.

FRELLSEN, J., MOLTKE, I., THIIM, M., MARDIA, K. V., FERKINGHOFF-BORG, J. and HAMELRYCK, T. (2009). A probabilistic model of RNA conformational space. *PLoS Computational Biology* **5** e1000406.

JAIN, S., RICHARDSON, D. C. and RICHARDSON, J. S. (2015). Computational methods for RNA structure validation and improvement. In *Methods in Enzymology*, **558** 181–212. Elsevier.

JINEK, M., JIANG, F., TAYLOR, D. W., STERNBERG, S. H., KAYA, E., MA, E., ANDERS, C., HAUER, M., ZHOU, K., LIN, S. et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343** 1247997.

JUNG, S., DRYDEN, I. L. and MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568.

KIEFT, J. S., ZHOU, K., GRECH, A., JUBIN, R. and DOUDNA, J. A. (2002). Crystal structure of an RNA tertiary domain essential to HCV IRES-mediated translation initiation. *Nature Structural Biology* **9** 370–374.

LILLEY, D. M. (2000). Structures of helical junctions in nucleic acids. *Quarterly Reviews of Biophysics* **33** 109–159.

LU, S., TANG, Y., YIN, S. and SUN, L. (2024). RNA structure: implications in viral infections and neurodegenerative diseases. *Advanced Biotechnology* **2** 3.

MARDIA, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society Series C: Applied Statistics* **62** 487–514.

MARDIA, K. V., WIECHERS, H., ELTZNER, B. and HUCKEMANN, S. F. (2022). Principal component analysis and clustering on manifolds. *Journal of Multivariate Analysis* **188** 104862.

MCINNES, L., HEALY, J. and MELVILLE, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

MOORE, P. B. (1998). The three-dimensional structure of the ribosome and its components. *Annual Review of Biophysics and Biomolecular Structure* **27** 35–58.

MURRAY, L. J., ARENDALL III, W. B., RICHARDSON, D. C. and RICHARDSON, J. S. (2003). RNA backbone is rotameric. *Proceedings of the National Academy of Sciences* **100** 13904–13909.

PANARETOS, V. M., PHAM, T. and YAO, Z. (2014). Principal flows. *Journal of the American Statistical Association* **109** 424–436.

ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

SERGANOV, A. and NUDLER, E. (2013). A decade of riboswitches. *Cell* **152** 17–24.

SHEN, L. X., CAI, Z. and TINOCO JR, I. (1995). RNA structure at high resolution. *The FASEB journal* **9** 1023–1033.

SHIVASHANKAR, G. (2002). Mesoscopic biology. *Pramana* **58** 439–442.

SPONER, J., BUSSI, G., KREPL, M., BANÁŠ, P., BOTTARO, S., CUNHA, R. A., GIL-LEY, A., PINAMONTI, G., POBLETE, S., JURECKA, P. et al. (2018). RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chemical Reviews* **118** 4177–4338.

TENENBAUM, J. B., SILVA, V. D. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* **9** 2579-2605.

WIECHERS, H., ELTZNER, B., MARDIA, K. V. and HUCKEMANN, S. F. (2023). Learning torus PCA-based classification for multiscale RNA correction with application to SARS-CoV-2. *Journal of the Royal Statistical Society Series C: Applied Statistics* **72** 271–293.

YAO, Z., ELTZNER, B. and PHAM, T. (2026). Principal Sub-manifolds. *Statistica Sinica* **36** 1–41.

YAO, Z., XIA, Y., TRAN, D. V. and ZHANG, Z. (2023). Hunting principal sub-manifolds: new theories and methods. *Technical Report*.

ZHU, Y., ZHU, L., WANG, X. and JIN, H. (2022). RNA-based therapeutics: an overview and prospectus. *Cell Death & Disease* **13** 644.

Appendices

APPENDIX A: RNA DIHEDRAL ANGLES

In the RNA backbone, repeating sugar-phosphate units form the structural framework, with the nitrogenous bases extending from the sugar. The backbone consists of alternating phosphate groups and pentose sugars, which are connected by phosphodiester bonds. As shown in Figure 12, the angle α is formed by the phosphorus atom (P), the oxygen atom in the phosphate group ($O5'$), and the carbon atom in the sugar ring ($C5'$). This angle influences the orientation of the phosphate group relative to the sugar, thereby affecting the flexibility and overall geometry of the RNA chain. Angles β and γ involve the carbon atoms of the sugar ring ($C5', C4', C3'$) and are essential for determining the sugar pucker configuration, which dictates whether the sugar adopts an 'endo' or 'exo' conformation and directly influences RNA secondary structure formation. Angles δ, ϵ, and ζ define the connections involving the oxygen atoms attached to the phosphate groups ($O3', O5'$); these angles play a critical role in the backbone conformation between nucleotides and are vital for the overall topology and folding of the RNA molecule. Finally, angle χ involves the nitrogenous base attached to the sugar ($C1'$) and is essential for base stacking interactions, which are important for the stability of RNA secondary and tertiary structures.
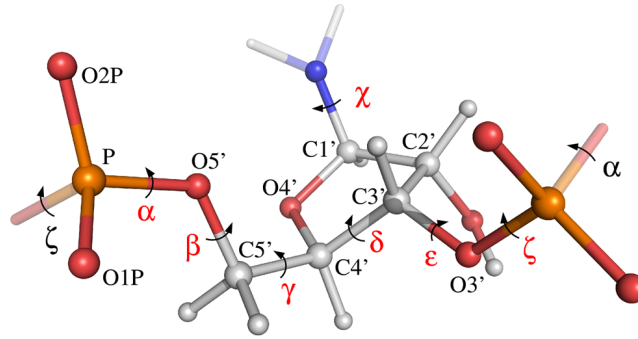

FIG 12. *Structural components at a microscopic scale: (P, O, C) (atoms) and bonds in a segment of the RNA backbone, with $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi)$ (dihedral angles) marked to indicate the rotational freedom around each bond. Reproduced from Frellsen et al. (2009).*

APPENDIX B: tPCA

Recall the $\mathbb{T}^D$ defined in Section 2.1, tPCA uses TOSS mapping, which is a pivotal step where samples $\mathcal{X}$ on the torus $\mathbb{T}^D$ are mapped to the sphere $\mathbb{S}^D := \{z \in \mathbb{R}^{D+1} : \|z\|_2 = 1\}$. The process is visualized in Figure 13. For $\mathbb{T}^D$, TOSS mapping Φ to the sphere $\mathbb{S}^D$ involves a complex series of transformations:

$$\Phi(x) : (x^{(1)}, x^{(2)}, \ldots, x^{(D)})^T \mapsto \left(y^{(1)}, y^{(2)}, \ldots, y^{(D+1)} \right)^T,$$

where y are given by:

$$
y^{(j)} = \begin{cases}
\cos\left(\frac{x^{(1)}}{2}\right), & j = 1, \\[2ex]
\left(\prod_{k=1}^{j-1} \sin\frac{x^{(k)}}{2}\right)\cos\frac{x^{(j)}}{2}, & j = 2,\dots,D-1, \\[2ex]
\left(\prod_{k=1}^{D-1} \sin\frac{x^{(k)}}{2}\right)\cos x^{(D)}, & j = D, \\[2ex]
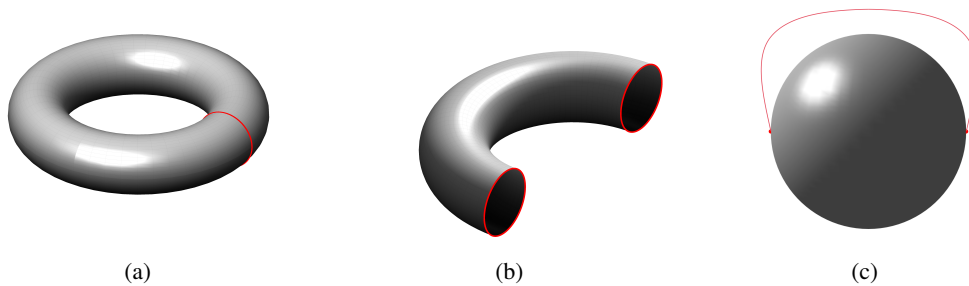\left(\prod_{k=1}^{D-1} \sin\frac{x^{(k)}}{2}\right)\sin x^{(D)}, & j = D+1.
\end{cases}
$$

(a)   (b)   (c)

FIG 13. *The process of TOSS mapping on 2-dimensional torus. (a) shows the original 2-dimensional torus in the 3-dimensional space. (b) shows that TOSS mapping cuts the torus through the red circles. (c) shows that the transformation collapses the ends of the cylinder, turning it into a sphere with identified poles. The transformation involves adjusting the torus angles so they correspond to polar angles on the sphere.*

After mapping the torus data onto the sphere, tPCA uses PNS for dimensionality reduction. Let $\mathbb{S}^k$ be a k-dimensional sphere embedded in $\mathbb{R}^{k+1}$. The PNS (small) are defined through a sequence of nested subspheres $\mathbb{S}_{V,\phi}$ as follows:

$$\mathbb{S}_{V,\phi} = \mathcal{A}_{V,\phi} \cap \mathbb{S}^k,$$

where $\mathcal{A}_{V,\phi}$ is an affine subspace of $\mathbb{R}^{k+1}$ defined by:

$$\mathcal{A}_{V,\phi} = \{z \in \mathbb{R}^{k+1} : V^T z = \phi, V^T V = I\}.$$

Here, ϕ is a vector in $\mathbb{R}^{k-\ell}$ representing the offsets from the origin in the directions orthogonal to the subsphere.

The core of PNS involves decomposing the sphere into a series of nested subspheres, each of decreasing dimension, starting from the full-dimensional sphere $\mathbb{S}^d$ and moving inward to $\mathbb{S}^{d-1}$, $\mathbb{S}^{d-2}$, and so on, until reaching $\mathbb{S}^0$, which is a singular point. This sequence of spheres forms a nested structure where each sphere is contained within the next higher dimension sphere. PNS involves identifying the optimal subsphere for each dimension by minimizing the sum of squared residuals between the geodesic distances of the samples and the corresponding sphere. Specifically, considering a given combination of torus coordinates, the objective distance function to be minimized is formulated as:

$$d(\mathbb{S}_{V,\phi}, x) = \min\Big\{ \arccos\left(y^T \pi_{\mathbb{S}_{V,\phi}}(y)\right),$$

$$\min_{d=0,\dots,k-1}\left(\arccos\left(\sqrt{\sum_{j=1}^{d+1} y_j^2}\right) + \arccos\left(\bar{y}^{d^T}\pi_{S_{V,\phi}}(\bar{y}^d)\right)\right)\Big\},$$

where $y = \Phi(x)$ and $\bar{y}^d$ is a suitable projection of y.

APPENDIX C: SUPPLEMENTARY SIMULATIONS EXPERIMENTS 2

The subexperiments of experiment 2 for dimensions $D = 2$, $D = 3$, $D = 4$, and $D = 5$ are illustrated in Figure 14 and Figure 15. These figures provide detailed clustering performance, enabling a comprehensive comparison between PSM-DBSCAN and other clustering methods in various dimensional settings.
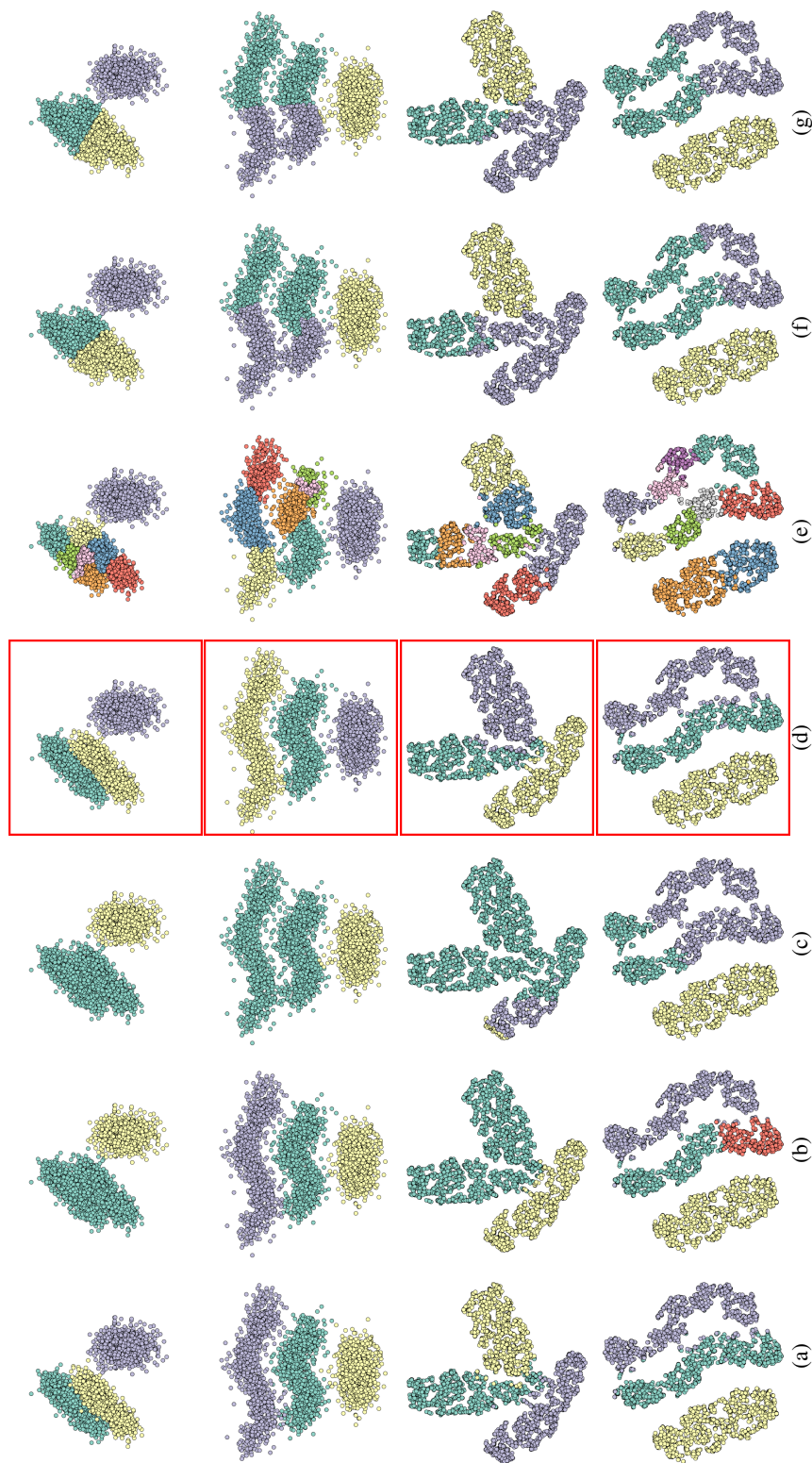
FIG 14. *An intuitive illustration of clustering performance for subexperiment* $(a - d)$. *Rows* 1 *and* 2 *correspond to subexperiments for* $D = 2$, *rows* 3 *and* 4 *correspond to* $D = 3$.
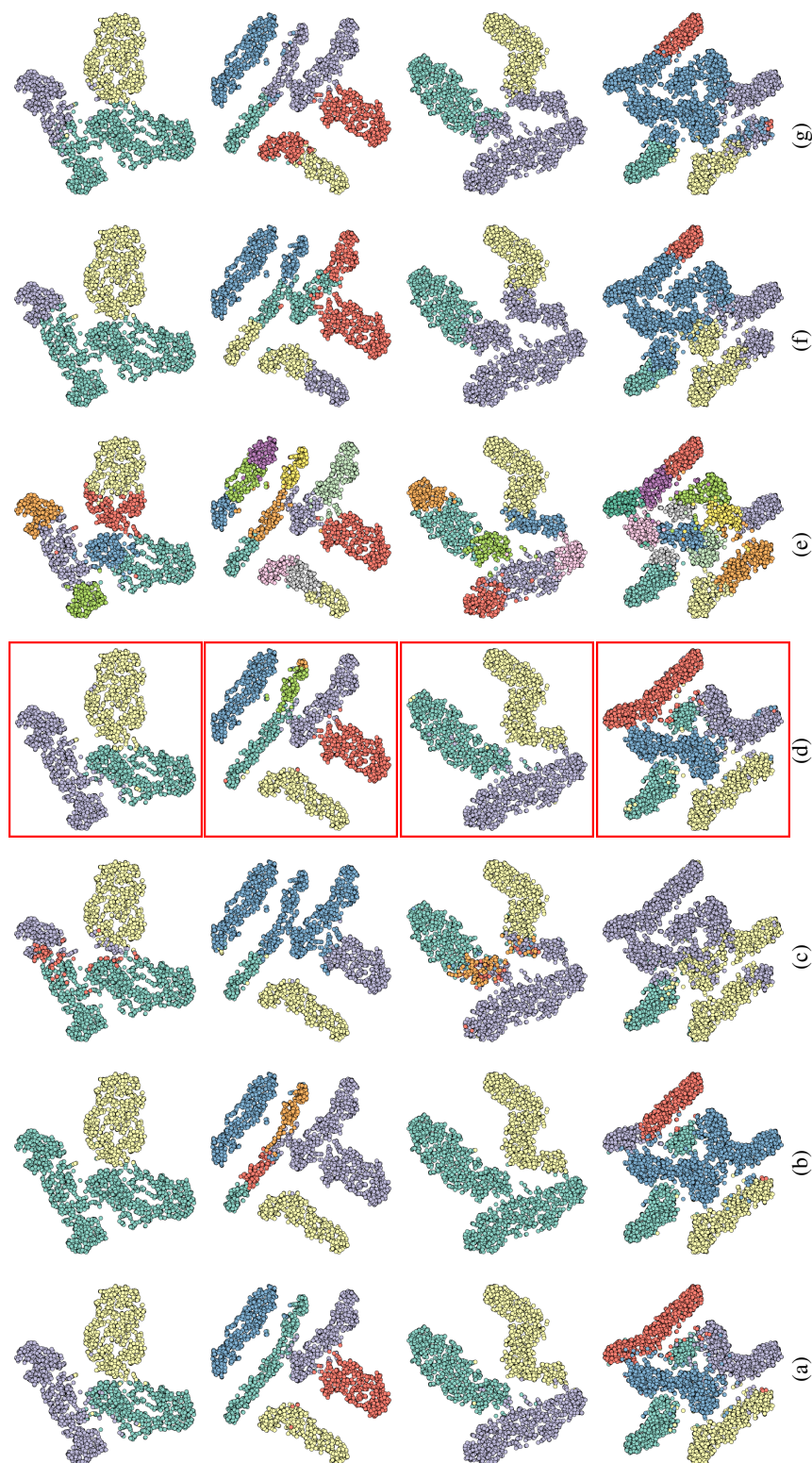
FIG 15. *An intuitive illustration of clustering performance for subexperiment* $(e - h)$. *Rows 1 and 2 correspond to subexperiments for* $D = 4$, *rows 3 and 4 correspond to* $D = 5$.