# Unsupervised Co-Learning on $\mathcal{G}$-Manifolds Across Irreducible Representations

Yifeng Fan* Tingran Gao† Zhizhen Zhao‡

**Abstract**

We introduce a novel co-learning paradigm for manifolds naturally equipped with a group action, motivated by recent developments on learning a manifold from attached fibre bundle structures. We utilize a representation theoretic mechanism that canonically associates multiple independent vector bundles over a common base manifold, which provides multiple views for the geometry of the underlying manifold. The consistency across these fibre bundles provide a common base for performing unsupervised manifold co-learning through the redundancy created artificially across irreducible representations of the transformation group. We demonstrate the efficacy of the proposed algorithmic paradigm through drastically improved robust nearest neighbor search and community detection on rotation-invariant cryo-electron microscopy image analysis.

## 1 Introduction

Fighting with the *curse of dimensionality* by leveraging low-dimensional intrinsic structures has become an important guiding principle in modern data science. Apart from classical structural assumptions commonly employed in sparsity or low-rank models in high dimensional statistics [TWH15, CR09, CSPW09, RWY12, BBEKY13, Ver18, Wai19], recently it has become of interest to leverage more intricate properties of the underlying geometric model, motivated by algebraic or differential geometry techniques, for efficient learning and inference from massive complex datasets [CLL+05a, CLL+05b, NLKC06, OWNB17, BKSW18]. The assumption that high dimensional datasets lie approximately on a low-dimensional manifold, known as the *manifold hypothesis*, has been the cornerstone for the development of manifold learning [TSL00, RS00, DG03, BN02, BN03, BNS06, CL06, SW12, VG18] in the past few decades.

In many real applications, the low-dimensional manifold underlying the dataset of high ambient dimensionality enjoys additional structures that can be fully leveraged to gain deeper insights into the geometry of the data. One class of such examples arises in scientific fields such as cryo-electron microscopy (cryo-EM), where large numbers of random projections for a three-dimensional molecule generate massive collections of images that can be determined only up to in-plane rotations [SZSH11, ZS14]. Another source of examples is the application in computer vision and robotics, where a major challenge is to recognize and compare three-dimensional spatial configurations up to the action of Euclidean or conformal groups [Ken89, BBK08]. In these examples, the dataset of interest consists of images or shapes of potentially high spatial resolution, and admits a natural group action that plays the role of a nuisance or latent variable that needs to be "quotient out" before useful information can be distilled. In geometric terms, on top of a differentiable manifold $\mathcal{M}$ underlying the dataset of interest, as assumed in the manifold hypothesis, we also assume the manifold admits a smooth *right action* of a Lie group $\mathcal{G}$, in the sense that there is a smooth map $\phi : \mathcal{G} \times \mathcal{M} \to \mathcal{M}$ satisfying

$$\phi(e, m) = m \quad \text{and} \quad \phi(g_2, \phi(g_1, m)) = \phi(g_1 g_2, m)$$

for all $m \in \mathcal{M}$ and $g_1, g_2 \in \mathcal{G}$, where $e$ is the identity element of $\mathcal{G}$. A *left action* can be defined similarly. Such a group action reflects abundant information about the symmetry of the underlying manifold, with which one

*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61820, USA, e-mail: yifengf2@illinois.edu.

†Committee on Computational and Applied Mathematics, Department of Statistics, University of Chicago, Chicago, IL, 60637, USA, e-mail: tingrangao@galton.uchicago.edu.

‡Department of Electrical and Computer Engineering, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, 61820, USA, e-mail: zhizhenz@illinois.edu.

can study geometric and topological properties of the underlying manifold through the lens of the orbit, stablizer, or induced finite- or infinite-dimensional representations of $\mathcal{G}$. In modern differential and symplectic geometry literature, a smooth manifold $\mathcal{M}$ admitting the action of a Lie group $\mathcal{G}$ is often referred to as a $\mathcal{G}$-manifold (see e.g. [Mic08, §6], [Ric01, AA93] and references therein), and thus transformation-centered methodology has been proven fruitful [MFK94, Sch08, Mic08, Kob12] by several generations of geometers and topologists.

Recent development of manifold learning has started to digest and incorporate the additional information encoded in the $\mathcal{G}$-actions on the low-dimensional manifold underlying the high-dimensional data. In [LS18], the authors constructed a *steerable graph Laplacian* on the manifold of images — modeled as a rotationally invariant manifold (or $\mathrm{U}(1)$-manifold in geometric terms) — that serves the role of graph Laplacian in manifold learning but naturally encodes the rotational invariance by construction. In [LMQW18], the authors proposed a principal bundle model for image denoising, which achieved state-of-the-art performance by combining patch-based image analysis with rotationally invariant distances in microscopy [PZF96]. A major contribution of this paper is to provide deeper insights into the success of these group-transformation-based manifold learning techniques from the perspective of *multi-view learning* [SR08, Sun13, LYZ18] or *co-training* [BM98], and propose a family of new methods that fully utilize these additional information in a systematic way, by exploiting the inherent consistency across representation theoretic patterns. Motivated by the recent line of research bridging manifold learning with principal and associated fibre bundles [SW12, SW16, Gao16, FZ19b, FZ19a], we point out that to a $\mathcal{G}$-manifold admitting a principal bundle structure is naturally associated as many vector bundles as the number of distinct irreducible representations of the transformation group $\mathcal{G}$, and each of these vector bundles provide a separate "view" towards unveiling the geometry of the common *base manifold* on which all the fibre bundles reside.

More specifically, the main contributions of this paper are summarized as follows: (1) establish a new unsupervised co-learning paradigm on $\mathcal{G}$-manifold and propose an optimal alignment affinity measure for high-dimensional data points that lie on or close to a lower dimensional $\mathcal{G}$-manifold, using both the local cycle consistency of group transformations on the manifold (graph) and the algebraic consistency of the unitary irreducible representations of the transformations; (2) introduce the invariant moments affinity in order to bypass the computationally intensive pairwise optimal alignment search and efficiently learn the underlying local neighborhood structure; and (3) empirically demonstrate that our new framework is robust to noise and apply it to improve cryo-EM image classification.

## 2 Related Work

**Manifold Learning:** Initiated from early explorations [TSL00, RS00], more recently [BNS06, SR08, MBM16] provided reproducing kernel Hilbert space frameworks for scalar and vector valued kernel and interpreted the manifold assumption as a specific type of regularization; [BN02, BN03, CL06] used the estimated eigenfunctions of the Laplace–Beltrami operator to parametrize the underlying manifold; [HS11b, HS11a, SZSH11] investigated into the representation theoretic pattern of an integral operator acting on certain complex line bundles over the unit two-sphere naturally arising from cryo-EM image analysis; [SW12, SW16, Gao16] demonstrated the benefit of using differential operators defined on fibre bundles over the manifold, instead of the Laplace–Beltrami operator on the manifold itself, in manifold learning tasks. Recently, [FZ19b, FZ19a, GZ19, GFZ19] proposed to utilize the consistency across multiple irreducible representations of a compact Lie group to improve spectral decomposition based algorithms.

**Co-training and Multi-view Learning:** In their seminal work [BM98], Blum and Mitchell demonstrated both in theory and empirically that distinct "views" of a dataset can be combined in to improve the performance of learning tasks, through their complementary yet consistent prediction for unlabelled data. Similar ideas exploiting the consistency of the information contained in different sets of features has long existed in statistical literature such as canonical correlation analysis [Ket71]. Since then, multi-view learning has remained a powerful idea percolating through aspects of machine learning ranging from supervised and semi-supervised learning to active learning and transfer learning [FHM$^+$06, MMK06, SH10, CWB11, SN05, SR08, KI11, KRD11]. See surveys [Sun13, XTX13, ZXXS17, LYZ18] for more detailed accounts.

# 3 Geometric Motivation

Motivated by [LMQW18, LS18], we consider a $\mathcal{G}$-manifold $\mathcal{M}$ admitting a principal bundle structure. We collect the definition of fibre bundle and principal bundle here for convenience; see [BGV03] for more details. Briefly speaking, a fibre bundle is a manifold which is locally diffeomorphic to a product space, and a principal fibre bundle is a fibre bundle with a natural group action on its "fibres."

**Definition 3.1** (Fibre Bundle)**.** *Let $\mathcal{M}, \mathcal{B}, \mathcal{F}$ be three differentiable manifolds, and let $\pi : \mathcal{M} \to \mathcal{B}$ denote a smooth surjective map between $\mathcal{M}$ and $\mathcal{B}$. We say that $\mathcal{M} \xrightarrow{\pi} \mathcal{B}$ (or just $\mathcal{M}$ for short) is a **fibre bundle** with typical fibre $\mathcal{F}$ over $\mathcal{B}$ if $\mathcal{B}$ admits an open cover $\mathscr{U}$ such that $\pi^{-1}(U)$ is diffeomorphic to product space $U \times \mathcal{F}$ for any open set $U \in \mathcal{U}$. For any $x \in \mathcal{B}$, we denote $\mathcal{F}_x := \pi^{-1}(x)$ and call it the **fibre** over $x$.*

**Definition 3.2** (Principal Bundle)**.** *Let $\mathcal{M}$ be a fibre bundle, and $\mathcal{G}$ a Lie group. We call $\mathcal{M}$ a **principal $\mathcal{G}$-bundle** if (1) $\mathcal{M}$ is a fibre bundle, (2) $\mathcal{M}$ admits a right action of $\mathcal{G}$ that preserves the fibres of $\mathcal{M}$, in the sense that for any $m \in \mathcal{M}$ we have $\pi(m) = \pi(g \cdot m)$, and (3) For any two points $p, q \in \mathcal{M}$ on the same fibre of $\mathcal{M}$, there exists a group element $g \in \mathcal{G}$ satisfying $p \cdot g = q$.*

If $\mathcal{M}$ is a principal $\mathcal{G}$-bundle over $\mathcal{B}$, any representation $\rho$ of $\mathcal{G}$ on a vector space $V$ induces an **associated vector bundle** over $\mathcal{B}$ with typical fibre $V$, denoted as $\mathcal{M} \times_\rho V$, defined as a quotient space

$$\mathcal{M} \times_\rho V := \mathcal{M} \times V \big/ \sim$$

where the equivalence relation is defined by $(m \cdot g, v) \sim (m, \rho(g)\, v)$ for all $m \in \mathcal{M}$, $g \in \mathcal{G}$, and $v \in V$. This construction gives rise to as many different associated vector bundles as the number of distinct representations of the Lie group $\mathcal{G}$. This allows us to study the $\mathcal{G}$-manifold $\mathcal{M}$, as a principal $\mathcal{G}$-bundle, through tools developed for learning an unknown manifold from attached vector bundle structures, such as *vector diffusion maps* (VDM) [SW12, SW16]. We consider each of these associated vector bundles as a distinct "view" towards the unknown data manifold $\mathcal{M}$, as the representations inducing these vector bundles are different. In the rest of this paper, we will illustrate with several examples how to design learning and inference algorithms that exploit the inherent consistency in these associated vector bundles by representation theoretic machinery. Unlike the co-training setting where the consistency is induced from the labelled samples onto the unlabelled samples, in our unsupervised setting no labelled training data is provided and the consistency lies purely in the geometry of the manifold.

# 4 Methods

In many applications of interest, the Lie group is compact and thus always admits *unitary* irreducible representations. By the renowned Peter–Weyl theorem, any square integrable function $f \in L_2(\mathcal{G})$ can be decomposed as

$$f(g) = \sum_{k=0}^{\infty} d_k \mathrm{Tr}\left[F_k \rho_k(g)\right], \quad \text{and} \quad F_k = \int_{\mathcal{G}} f(g) \rho_k^*(g) d\mu_g, \tag{4.1}$$

where each $\rho_k : \mathcal{G} \to \mathbb{C}^{d_k \times d_k}$ is a *unitary irreducible representation (irrep)* of $\mathcal{G}$ with dimension $d_k \in \mathbb{N}$. This is the compact Lie group analogy of the standard Fourier series over the unit circle. The "generalized Fourier coefficient" $F_k$ in (4.1) is defined by the integral taken with respect to the Haar measure on $\mathcal{G}$.

## 4.1 Problem Setup

Given a dataset $\{x_1, \ldots, x_n\} \subset \mathbb{R}^l$, we assume that the data lie on or close to a low dimensional smooth manifold $\mathcal{M}$ of intrinsic dimension $d \ll l$, and that $\mathcal{M}$ is a $\mathcal{G}$-manifold admitting the structure of a principal $\mathcal{G}$-bundle with a compact Lie group $\mathcal{G}$. The data space $\mathcal{M}$ is closed under $\mathcal{G}$ if $g \cdot x \in \mathcal{M}$ for all $g \in \mathcal{G}$ and $x \in \mathcal{M}$, where '$\cdot$' denotes the group action. The $\mathcal{G}$-invariant distance between two data points is defined as

$$d_{ij} = \min_{g \in \mathcal{G}} \|x_i - g \cdot x_j\|, \quad \text{and} \quad g_{ij} = \arg\min_{g \in \mathcal{G}} \|x_i - g \cdot x_j\|. \tag{4.2}$$

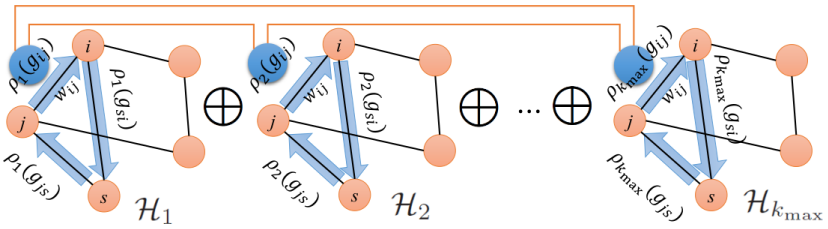


Figure 1: Illustration of unsupervised co-learning across irreducible representations. Within each graph of single representation the cycle consistency of the group transformation holds $\rho_k(g_{js})\rho_k(g_{si})\rho_k(g_{ij}) \approx I_{d_k \times d_k}$; moreover, the irreps should be consistent algebraically along the orange lines connecting the blue dots representing transformations on the edges. The unsupervised co-learning paradigm exploits all such consistencies.

**Algorithm 1:** Weight Matrix Filtering

**Input:** Initial graph $G = (V, E)$ with $n$ nodes and the corresponding edge weights $w_{ij}$ and transformation group $g_{ij}$, cutoff parameter $m_k$ for $k = 1, \ldots, k_{\max}$, and diffusion time $t$

**Output:** The filtered weight matrix $\widetilde{W}_{k,t}$

**for** $k = 1, \ldots, k_{\max}$ **do**
- Construct the block weight matrix $W_k$ of size $nd_k \times nd_k$ and the normalized symmetric matrix $\widetilde{A}_k$ according to (4.3) and (4.5)
- Compute the largest $m_k d_k$ eigenvalues $\lambda_1^{(k)} \geq \lambda_2^{(k)}, \geq, \ldots, \geq \lambda_{m_k d_k}^{(k)}$ of $\widetilde{A}_k$ and the corresponding eigenvectors $\{u_l^{(k)}\}_{l=1}^{m_k d_k}$
- **for** $i = 1, \ldots, n$ **do**
  - Construct the $\mathcal{G}$-equivariant mapping, $\psi_t^{(k)} : i \mapsto \left[\eta_{2t}(\lambda_1)^{1/2} u_1^{(k)}(i), \ldots, \eta_{2t}(\lambda_{m_k})^{1/2} u_{m_k d_k}^{(k)}(i)\right]$
  - Compute the singular value decomposition of $\psi_t^{(k)}(i) = U\Sigma V^*$
  - Compute the normalized mapping $\widetilde{\psi}_t^{(k)}(i) = UV^*$.
- **end**
- Vertically concatenate $\widetilde{\psi}_t^{(k)}(i)$ to form the matrix $\Psi_t^{(k)}$ of size $nd_k \times m_k d_k$
- Construct the filtered and normalized weight matrix $\widetilde{W}_{k,t} = \Psi_t^{(k)} \left(\Psi_t^{(k)}\right)^*$

**end**

where $\|\cdot\|$ is the Euclidean distance on the ambient space $\mathbb{R}^l$ and $g_{ij}$ is the associated optimal alignment. We assume that the optimal alignment is unique and construct an undirected graph $G = (V, E)$ based on distance (4.2) using the $\epsilon$-neighborhood criterion, i.e. $(i, j) \in E$ iff $d_{ij} < \epsilon$, or $\kappa$-nearest neighbor criterion, i.e. $(i, j) \in E$ iff $j$ is one of the $\kappa$ nearest neighbors of $i$. The edge weights $w_{ij}$ are defined using a kernel function on the $\mathcal{G}$-invariant distance $w_{ij} = K_\sigma(d_{ij})$. In many applications, noise in the observational data affects the estimations of $\mathcal{G}$-invariant distances $d_{ij}$ and optimal alignments $g_{ij}$. This results in errors in the edge connections in the $\epsilon$-neighborhood graph or $\kappa$-nearest neighbor graph, and connects points on $\mathcal{B}$ where the underlying geodesic distances are large. To recover the underlying clean geometric graph structure on $\mathcal{B}$, we propose an unsupervised co-learning framework using the irreps of $\mathcal{G}$. The main intuition is to systematically explore the cycle consistency of the transformations of the principal bundles across all irreps (see Fig. 1).

## 4.2 Weight Matrices Using Irreps

We extend VDM using multiple irreps of the compact Lie group $\mathcal{G}$. Given the graph structure $G = (V, E)$ with $n$ nodes and the group transformations $g_{ij}$, we define a set of weight matrices that takes into account both the scalar edge connection weight and the corresponding alignment group using unitary irreps,

$$W_k(i, j) = \begin{cases} w_{ij}\rho_k(g_{ij}) & (i, j) \in E, \\ 0 & \text{otherwise}, \end{cases} \tag{4.3}$$

where $w_{ij} = w_{ji}$ and $\rho_k(g_{ji}) = \rho_k^*(g_{ij})$ for all $(i,j) \in E$. Therefore the matrix $W_k$ is a block matrix with $n \times n$ blocks of size $d_k \times d_k$. The degree of node $i$ is: $\deg(i) := \sum_{j:(i,j)\in E} w_{ij}$ and the degree matrix $D$ is a block diagonal matrix and the $(i,i)$-block of the matrix $D(i,i) = \deg(i) I_{d_k \times d_k}$. We construct the normalized matrix $A_k = D^{-1} W_k$. The Hilbert space $\mathcal{H}$, as a unitary representation of the compact Lie group $\mathcal{G}$, admits an isotypic decomposition $\mathcal{H} = \bigoplus \mathcal{H}_k$, where a function $f$ is in $\mathcal{H}_k$ if and only if $f(xg) = g^k f(x)$. The matrix $A_k$ is an averaging operator for vector fields in $\mathcal{H}_k$, i.e.,

$$(A_k z_k)(i) = \frac{1}{\deg(i)} \sum_{j:(i,j)\in E} w_{ij} \rho_k(g_{ij}) z_k(j). \tag{4.4}$$

Then the averaging operator $A_k$ is similar to the Hermitian matrix

$$\widetilde{A}_k = D^{-1/2} W_k D^{-1/2}, \tag{4.5}$$

which has real eigenvalues and orthonormal eigenvectors $\{\lambda_l^{(k)}, u_l^{(k)}\}_{l=1}^n$ and all the eigenvalues are within $[-1,1]$. For simplicity, here we assume that data points are uniformly distributed on $\mathcal{B}$. If the data are non-uniformly distributed, we apply the normalization proposed in [CL06] to $W_k$. The matrices $L_k = I - \widetilde{A}_k$ are the normalized graph connection Laplacians.

VDM defines the affinity between $i$ and $j$ as $\|\widetilde{A}_1^{2t}(i,j)\|_{\mathrm{HS}}^2$, the squared Hilbert-Schmidt norm of the $d_1 \times d_1$ matrix $\widetilde{A}_1^{2t}(i,j)$, which takes into account all paths of length $2t$, where $t$ is a positive integer. It measures both the connectivity and the amount of agreement between their transformations at $k=1$. The affinity is larger when the path transformations are in agreement, and is smaller when they differ. We generalize this to compute the HS norm of the filtered and normalized weight matrix $\widetilde{W}_{k,t} = \eta_{2t}(\widetilde{A}_k)$, where $\eta_{2t}(\cdot)$ denotes a spectral filter for the graph adjacency matrices, for example $\eta_{2t}(\lambda) = \lambda^{2t}$. With top eigenvalues and eigenvectors $\{\lambda_l^{(k)}, u_l^{(k)}\}_{l=1}^{m_k d_k}$, we define a $\mathcal{G}$-equivariant embedding:

$$\psi_t^{(k)} : i \mapsto \left[\eta_{2t}(\lambda_1)^{1/2} u_1^{(k)}(i), \ldots, \eta_{2t}(\lambda_{m_k d_k})^{1/2} u_{m_k d_k}^{(k)}(i)\right]. \tag{4.6}$$

We also normalize $\psi_t^{(k)}(i)$ to ensure that the diagonal blocks of $\widetilde{W}_{k,t}$ are identity. The steps for filtering weight matrices are detailed in Alg. 1.

## 4.3 Optimal Alignment Affinity

VDM and frequency-$k$ VDM [FZ19b] only explore the transformation consistency at a single representation, which is achieved by computing the filtered and normalized weight matrix. As shown in Fig. 1, it is advantageous to couple the information under different irreps, similar to unsupervised multi-view learning. We apply the algebraic relation among different $\widetilde{W}_{k,t}$'s according to the generalized Fourier transform in (4.1) and define the optimal alignment affinity,

$$S_t^{\mathrm{OA}}(i,j) = \max_{g\in\mathcal{G}} \frac{1}{k_{\max}} \left| \sum_{k=1}^{k_{\max}} \mathrm{Tr}\left[\widetilde{W}_{k,t}(i,j)\rho_k(g)\right] \right|, \tag{4.7}$$

which can be evaluated using generalized FFTs [MR97] (see Alg. 2). For the nearest neighbor pairs $(i,j)$, the alignment group can be estimated as

$$\tilde{g}_{ij} = \arg\max_{g\in\mathcal{G}} \frac{1}{k_{\max}} \left| \sum_{k=1}^{k_{\max}} \mathrm{Tr}\left[\widetilde{W}_{k,t}(i,j)\rho_k(g)\right] \right|. \tag{4.8}$$

## 4.4 Invariant Moments Affinity

Searching for the optimal alignment as described above is challenging and time consuming. Therefore, we use invariant features to speed up the computation of the similarity measure. The *power spectrum*, the Fourier transform of the auto-correlation defined as $P_f(k) = F_k F_k^*$, is transformation invariant since under the right

**Algorithm 2:** Optimal Alignment Affinity

**Input:** Filtered and normalized weight matrices $\widetilde{W}_{k,t}$ for $k = 1, \dots, k_{\max}$
**Output:** The optimal alignment affinity $S_t^{\text{OA}}$

```
for i = 1, ..., n do
    for j = 1, ..., n do
        Compute the optimal alignment affinity S_t^OA for (i, j) based on (4.7)
    end
end
```

**Algorithm 3:** Power Spectrum Invariant Affinity

**Input:** Filtered and normalized weight matrices $\widetilde{W}_{k,t}$ for $k = 1, \dots, k_{\max}$
**Output:** The power spectrum invariant affinity $S_t^{\text{power spec}}$

```
for i = 1, ..., n do
    for j = 1, ..., n do
        for k = 1, ..., k_max do
            Compute the power spectrum feature P_{k,t}(i, j) as in (4.9)
        end
        Compute S_t^{power spec}(i, j) as the weighted sum of the trace of P_{k,t}(i, j) according to (4.9)
    end
end
```

action of $g \in \mathcal{G}$, the Fourier coefficients $F_k \to F_k \rho_k(g)$ and $P_{f\cdot g}(k) = F_k \rho_k(g) \rho_k(g)^* F_k^* = P_f(k)$. We compute the power spectrum $P_{k,t}$ of the filtered weight matrices $\widetilde{W}_{k,t}$ and define the corresponding affinity,

$$
\begin{aligned}
&S_t^{\text{power spec}}(i,j) = \frac{1}{k_{\max}} \left| \sum_{k=1}^{k_{\max}} \operatorname{Tr}\left[P_{k,t}(i,j)\right] \right|, \\
&\text{with } P_{k,t}(i,j) = \widetilde{W}_{k,t}(i,j) \widetilde{W}_{k,t}(i,j)^*.
\end{aligned} \tag{4.9}
$$

Previously, multi-frequency vector diffusion maps (MFVDM) proposed in [FZ19b] uses the power spectrum rotational invariant moments to combine the cycle consistencies of the in-plane rotations at different frequencies. Here, we extend it to general compact Lie group. The affinity $S_t^{\text{power spec}}$ combines the information at different irreps, however, it does not couple them and loses the relative *phase information*. Thus the affinity might be inaccurate under high level of noise.

Consider two unitary irreducible representations on vector spaces $\mathcal{H}_{k_1}$ and $\mathcal{H}_{k_2}$ of $\mathcal{G}$. For $\mathcal{G}$ is compact and $\mathcal{H}_{k_1}$ and $\mathcal{H}_{k_2}$ finite dimensional, there is a unique decomposition of $\rho_{k_1} \bigotimes \rho_{k_2}$ into a set of unitary irreducible representations $\rho_k$, $k \in \mathbb{N}$, where $\bigotimes$ is the Kronecker product of matrices, and we use $\bigoplus$ to denote direct sum. There exists $\mathcal{G}$-equivariant maps from $\mathcal{H}_{k_1} \bigotimes \mathcal{H}_{k_2} \to \bigoplus \mathcal{H}_k$, called generalized Clebsch–Gordan coefficients $C_{k_1,k_2}$ for compact Lie group $\mathcal{G}$ , which satisfies

$$
\rho_{k_1}(g) \bigotimes \rho_{k_2}(g) = C_{k_1,k_2} \left[ \bigoplus_{k \in k_1 \bigotimes k_2} \rho_k(g) \right] C_{k_1,k_2}^*. \tag{4.10}
$$

Using (4.10) and the fact that $C_{k_1,k_2}$ and $\rho_k$'s are unitary matrices, we have

$$
\left[\rho_{k_1}(g) \bigotimes \rho_{k_2}(g)\right] C_{k_1,k_2} \left[ \bigoplus_{k \in k_1 \bigotimes k_2} \rho_k^*(g) \right] C_{k_1,k_2}^* = I_{d_{k_1} d_{k_2} \times d_{k_1} d_{k_2}}.
$$

To systematically impose the algebraic consistency without solving the optimization problem in (4.7), we propose to use bispectrum invariants to define a new affinity. The triple correlation of a function $f$ on $\mathcal{G}$ can

**Algorithm 4:** Bispectrum Invariant Affinity

**Input:** Filtered and normalized weight matrices $\widetilde{W}_{k,t}$ for $k = 1, \dots, k_{\max}$

**Output:** The bispectral invariant affinity $S_t^{\text{bispec}}$

```
for i = 1, ..., n do
    for j = 1, ..., n do
        for k_1 = 1, ..., k_max do
            for k_2 = 1, ..., k_max do
                Compute the bispectral feature B_{k_1,k_2,t}(i,j) in (4.13)
            end
        end
        Compute S_t^bispec(i,j) as the weighted sum of the trace of B_{k_1,k_2,t}(i,j) according to (4.12)
    end
end
```

be defined as, $a_{3,f}(g_1, g_2) = \int_{\mathcal{G}} f^*(g) f(gg_1) f(gg_2) d\mu_g$. The bispectrum is the Fourier transform of the triple correlation $a_{3,f}$ and has the form

$$B_f(k_1, k_2) = \left[F_{k_1} \bigotimes F_{k_2}\right] C_{k_1,k_2} \left[\bigoplus_{k \in k_1 \bigotimes k_2} F_k^*\right] C_{k_1,k_2}^*, \tag{4.11}$$

and is $\mathcal{G}$-invariant [KM10, Feh10, Kon07]. The bispectrum has been used in several fields, including astrophysics [BJZ+16, WK00], statistics [Bri65, Kon08], cryo-electron microscopy data analysis [ZS14], and computer vision [Kon07, KM10]. The bispectrum $\mathcal{G}$-invariant affinity is defined as

$$S_t^{\text{bispec}}(i,j) = \frac{1}{k_{\max}^2} \left| \sum_{k_1=1}^{k_{\max}} \sum_{k_2=1}^{k_{\max}} \operatorname{Tr}\left[B_{k_1,k_2,t}(i,j)\right] \right|, \tag{4.12}$$

with

$$B_{k_1,k_2,t}(i,j) = \left[\widetilde{W}_{k_1,t}(i,j) \bigotimes \widetilde{W}_{k_2,t}(i,j)\right] C_{k_1,k_2} \left[\bigoplus_{k \in k_1 \bigotimes k_2} \widetilde{W}_{k,t}^*(i,j)\right] C_{k_1,k_2}^*. \tag{4.13}$$

If the transformations are consistent across different $k$'s, then the trace of $B_{k_1,k_2,t}$ in (4.13) should be large. The proposed new affinity measure takes into account the consistency of the transformation at each frequency and also enforces the algebraic consistency across different irreps.

## 4.5 Higher Order Invariant Moments

It is possible to design higher order invariant features to define pairwise affinity. For example, we can define the order-$d+1$ $\mathcal{G}$-invariant features as

$$M_{k_1,\dots,k_d} = \left[F_{k_1} \bigotimes \cdots \bigotimes F_{k_d}\right] C_{k_1,\dots,k_d} \left[\bigoplus_{k \in k_1 \bigotimes \cdots \bigotimes k_d} F_k^*\right] C_{k_1,\dots,k_d}^*, \tag{4.14}$$

where $C_{k_1,\dots,k_d}$ is the extension of the Clebsch–Gordan coefficients. However, using higher order spectra dramatically increases the computational complexity that grows exponentially with the order $d$. The bispectra are sufficient to enforce the consistency of the group transformations between nodes and across all irreps.

## 4.6 Computational Complexity

Filtering the normalized weight matrix involves computing the top $m_k d_k$ eigenvectors of the sparse Hermitian matrices $A_k$, for $k = 1, \dots, k_{\max}$, which can be efficiently evaluated using block Lanczos method [RST09], and

its cost is $O(nm_k d_k^2(m_k + l_k))$, where $l_k$ is the average number of non-zero elements in each row of $\widetilde{A}_k$. We compute the spectral decomposition for different $k$'s in parallel. Computing the power spectrum invariant affinity for all pairs takes $O(n^2 \sum_{k=1}^{k_{\max}} d_k^2)$ flops. The computational complexity of evaluating the bispectrum invariant affinity is $O(n^2(\sum_{k_1=0}^{k_{\max}} \sum_{k_2=0}^{k_{\max}} d_{k_1}^2 d_{k_2}^2))$. For the optimal alignment affinity, the computational complexity depends on the cost of optimal alignment search $C_a$ and the total cost is $O(n^2 C_a)$. For certain group structures, where FFTs are developed, the optimal alignment affinity can be efficiently and accurately approximated. However $C_a$ is still larger than the computation cost of invariants.

### 4.7 Examples with $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$

If the transformation parameter is a 2-D in-plane rotational angle $\alpha \in (0, 2\pi]$, i.e. $\mathcal{G} = \mathrm{SO}(2)$, the unitary irreps of the group are $\rho_k(\alpha) = e^{\imath k\alpha}$, where $\imath = \sqrt{-1}$. The dimensions of the irreps are $d_k = 1$, and $k_1 \bigotimes k_2 = k_1 + k_2$. The generalized Clebsch–Gordan coefficients is 1 for all $(k_1, k_2)$ pairs. For the optimal alignment affinity, we can use length $N$ zero-padded FFT to efficiently find approximate solution, therefore the computational complexity for evaluating $S_t^{\mathrm{OA}}(i, j)$ is $O(N \log N)$. If $\mathcal{G} = \mathrm{SO}(3)$, the unitary irreps are the Wigner D-matrices $D_k(\omega)$ for $\omega \in \mathrm{SO}(3)$ [Wig32]. The dimensions of $D_k$ are $d_k = 2k + 1$, and $k_1 \bigotimes k_2 = \{|k_1 - k_2|, \ldots, k_1 + k_2\}$. The Clebsch–Gordan coefficients for all $(k_1, k_2)$ pairs can be numerically precomputed [Hal15]. The optimal alignment affinity can be efficiently approximated using the FFTs on rotation group [KR08].

## 5 Experiments

We evaluate our paradigm on several examples: (1) Nearest neighbor (In brevity: **NN**) search on 2-sphere $\mathrm{S}^2$ and 3-sphere $\mathrm{S}^3$; (2) Cryo-EM 2-D image classification; (3) Spectral clustering with $\mathrm{SO}(2)$ or $\mathrm{SO}(3)$ transformation. All the experiments are conducted in MATLAB on a computer with Intel i7 7th generation quad core CPU.

**Random Rewiring Model:** One advantage of our paradigm is the robustness to noise. We demonstrate this through simulated data under the following random rewiring model. We start with the clean neighborhood graph according to the $\mathcal{G}$-invariant distances, then build a noisy graph as following: with probability $p$, we keep the existing clean graph edge $(i, j)$, and with probability $1 - p$, we remove it and link $i$ to a vertex, drawn uniformly at random from the remaining vertices that are not already connected to $i$. For rewired edges, the alignment $g_{ij}$ is uniformly distributed over $\mathcal{G}$ according to the Haar measure. Therefore, the parameter $p$ controls the signal to noise ratio (SNR) of our graph where $p = 1$ indicates the clean case.

**Nearest Neighbor Search for $\mathcal{M} = \mathbf{SO(3)}$, $\mathcal{G} = \mathbf{SO(2)}$, $\mathcal{B} = \mathbf{S^2}$:** We simulate $n = 10^4$ points $x_i$ uniformly distributed over $\mathrm{SO}(3)$ according to the Haar measure. Each $x_i$ can be represented by a $3 \times 3$ orthogonal matrix $R_i = [R_i^1, R_i^2, R_i^3]$ whose determinant is equal to 1. Then the vector $v_i = R_i^3$ can be identified as a point on the unit 2-sphere $\mathrm{S}^2$. The first two columns $R_i^1$ and $R_i^2$ spans the tangent plane of the sphere at $v_i$. The angle $\alpha_{ij}$ optimally aligns the tangent bundles $[R_j^1, R_j^2]$ to $[R_i^1, R_i^2]$. We build the clean graph by connecting nodes with $\langle v_i, v_j \rangle \geq \epsilon = 0.97$, and add noise following the rewiring model with different SNRs $p$. For each node, we find its 50 NNs based on the affinities introduced in this paper and the vector diffusion maps (VDM)[SW12] affinity. In Fig. 2 we plot the histogram of $\arccos \langle v_i, v_j \rangle$ of founded neighbors. When $p = 0.08$ to $p = 0.1$ (over 90% edges are corrupted), bispectrum and optimal alignment outperform power spectrum and VDM. This indicates our proposed affinities are able to recover the underlying clean graph, at extremely high noise level.

**Nearest Neighbor Search for $\mathcal{M} = \mathbf{SO(4)}$, $\mathcal{G} = \mathbf{SO(3)}$, $\mathcal{B} = \mathbf{S^3}$:** Similarly, we simulate $n = 500$ points $x_i$ uniformly distributed over $\mathrm{SO}(4)$ according to the Haar measure. In our experiments we build the clean graph by connecting each nodes with its 20 nearest neighbors (distance of $x_i$ and $x_j$ is measured by $\langle x_i, x_j \rangle$), then noise is added based on the rewiring model. In Fig. 3 we show the 20 nearest neighbor search result at different noise levels. Due to the large computational complexity we do not perform the optimal alignment affinity in this example. It can be seen that both bispectrum and power spectrum could achieve similar results at different noise levels, they also outperform VDM [SW12], which only consider the transformation consistency at single representation. Again, this result demonstrates the robustness of our propose affinities.

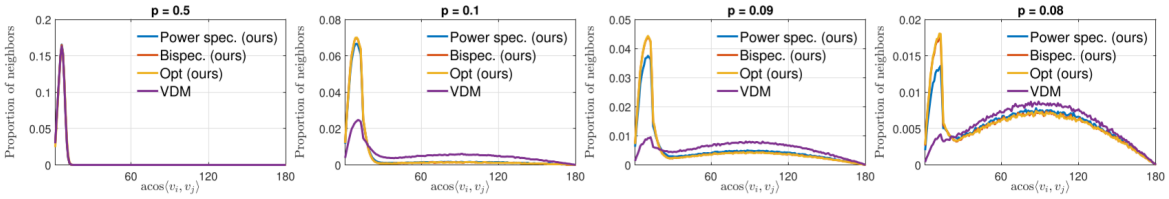


Figure 2: Histograms of $\arccos\langle v_i, v_j\rangle$ between estimated nearest neighbors on $S^2$, with different SNRs $p$. The clean histogram should peak at small angles. The lines of bispectrum and optimal alignment (Opt) almost overlap in all these plots. We set the maximum frequency $k_{\max} = 10$, the truncation $m_k = 50$ and the length $t = 1$.

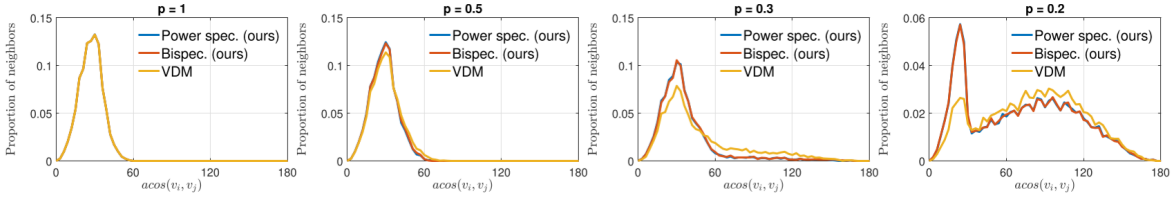


Figure 3: Histograms of $\arccos\langle v_i, v_j\rangle$ between estimated nearest neighbors on $S^3$, with different SNRs $p$. The clean histogram should peak at small angles. The lines of bispectrum and power spectrum almost overlap in all these plots. We set the maximum frequency $k_{\max} = 6$, the truncation $m_k = 10$ and the length $t = 1$.

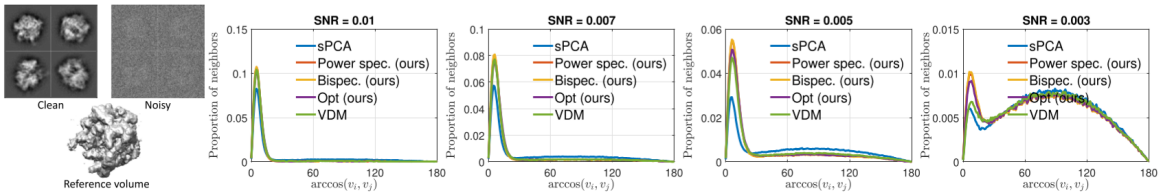


Figure 4: Cryo-EM 2-D image classification. *Left*: clean, noisy (SNR = 0.01) projections image samples, and reference volume of 70s ribosome. *Right*: Histograms of $\arccos\langle v_i, v_j\rangle$ between founded nearest neighbors. sPCA is the initial noisy input of our graph structure. The lines of power spectrum and bispectrum almost overlap in all these plots. We set the maximum frequency $k_{\max} = 20$, the truncation $m_k = 20$ and the length $t = 1$.

**Cryo-EM 2-D Image Classification:** An important application of the NN search above is the cryo-EM image analysis: Given a series of projection images of a macromolecule, with unknown random orientations, we aim to identify images with similar views, and perform local alignment and averaging to boost the image SNRs. Therefore, each projection can be viewed as a data point lying on the $S^2$ sphere, and the transformation is in-plane rotation of the image (i.e., $SO(2)$). In our experiments we simulate $n = 10^4$ projection images from a 3D electron density map of the 70S ribosome, the orientations of all projections are uniformly distributed over $SO(3)$ and the images are contaminated by additive white Gaussian noise with different SNRs, sample images are shown in Fig. 4. As a preprocessing step, we use fast steerable PCA (sPCA) [ZSS16] and rotationally invariant features [ZS14] to initially identify the images with similar views and the in-plane rotational angles according to [ZS14]. Then we compute our three proposed affinities and further improve the NN search result. In Fig. 4, we display the histograms of the angles (i.e., $\arccos\langle v_i, v_j\rangle$) between the estimated NN pairs. All our proposed methods outperform VDM [SW12]. Moreover, power spectrum and bispectrum affinity achieve similar accuracy, and outperform the optimal alignment affinity. This observation is different from the previous synthetic examples because of the different noise models in random rewiring model (independent noise on edges) and cryo-EM 2-D images (independent noise on nodes).

Table 1: Rand indices of spectral clustering results with SO(2) or SO(3) group transformation. We set the number of clusters *Left*: $K = 2$ and *right*: $K = 10$, for both cases the truncation $m_k = 10$ and maximum frequency $k_{\max} = 10$. For $K = 10$ and SO(3) case, each cluster has 25 points, otherwise each cluster has 50 points. We set the length $t = 1$ for all cases. See text for the method description. For SO(2) and SO(3) cases we use 50 and 10 trails respectively.

| | method | $K = 2$ clusters $p = 0.16$ | $p = 0.20$ | $p = 0.25$ | $K = 10$ clusters $p = 0.16$ | $p = 0.20$ | $p = 0.25$ |
|---|---|---|---|---|---|---|---|
| SO(2) | Scalar | $0.569 \pm 0.069$ | $0.705 \pm 0.092$ | $0.837 \pm 0.059$ | $0.868 \pm 0.010$ | $0.948 \pm 0.015$ | $0.981 \pm 0.013$ |
| | VDM | $0.526 \pm 0.036$ | $0.644 \pm 0.076$ | $0.857 \pm 0.057$ | $0.892 \pm 0.010$ | $0.963 \pm 0.011$ | $0.994 \pm 0.008$ |
| | **Power spec. (ours)** | $0.670 \pm 0.065$ | $0.899 \pm 0.051$ | $0.981 \pm 0.021$ | $0.975 \pm 0.010$ | $0.991 \pm 0.011$ | $0.998 \pm 0.006$ |
| | **Opt (ours)** | $\mathbf{0.687 \pm 0.011}$ | $\mathbf{0.912 \pm 0.009}$ | $\mathbf{0.986 \pm 0.007}$ | $\mathbf{0.976 \pm 0.012}$ | $0.994 \pm 0.008$ | $0.997 \pm 0.005$ |
| | **Bispec. (ours)** | $0.664 \pm 0.073$ | $0.901 \pm 0.062$ | $0.983 \pm 0.019$ | $0.967 \pm 0.014$ | $\mathbf{0.997 \pm 0.003}$ | $\mathbf{1 \pm 0.0003}$ |
| SO(3) | Scalar | $0.572 \pm 0.061$ | $0.666 \pm 0.095$ | $0.862 \pm 0.056$ | $0.840 \pm 0.003$ | $0.857 \pm 0.007$ | $0.910 \pm 0.019$ |
| | VDM | $0.600 \pm 0.048$ | $0.840 \pm 0.056$ | $0.974 \pm 0.023$ | $0.857 \pm 0.011$ | $0.912 \pm 0.013$ | $0.969 \pm 0.014$ |
| | **Power spec. (ours)** | $\mathbf{0.921 \pm 0.038}$ | $0.986 \pm 0.016$ | $\mathbf{1 \pm 0}$ | $\mathbf{0.884 \pm 0.011}$ | $\mathbf{0.945 \pm 0.011}$ | $0.977 \pm 0.017$ |
| | **Bispec. (ours)** | $0.911 \pm 0.043$ | $\mathbf{0.990 \pm 0.010}$ | $\mathbf{1 \pm 0}$ | $0.879 \pm 0.012$ | $0.938 \pm 0.009$ | $\mathbf{0.983 \pm 0.011}$ |

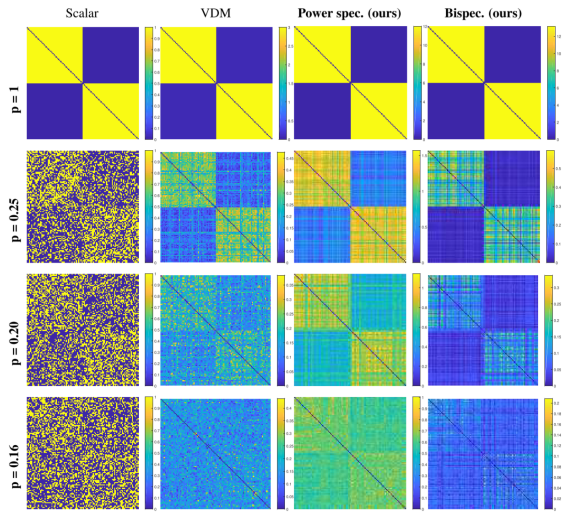


Figure 5: The affinity matrices for 2 clusters with SO(3) group transformation estimated by (1) the original scalar edge connections (Scalar), (2) vector diffusion maps (VDM), (3) power spectrum (Power spec.), and (4) bispectrum (Bispec.), at different SNRs. The underlying clean graph is corrupted according to the random rewiring model. The clusters are of equal size and form two diagonal blocks in the clean affinity matrix (see the scaler column at $p = 1$). Here we do not include the affinity of each node with itself and the diagonal entries are 0.

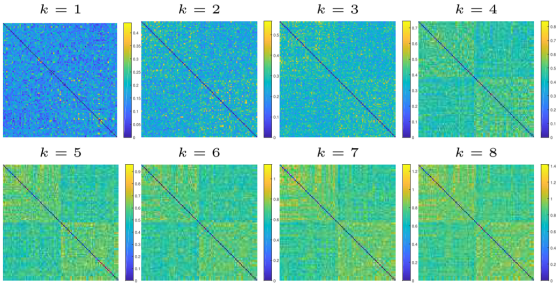


Figure 6: The affinity $|\mathrm{Tr}\left[P_{k,t}(i,j)\right]|$ for 2 clusters with SO(3) transformation at different frequency $k$. SNR $p = 0.16$.

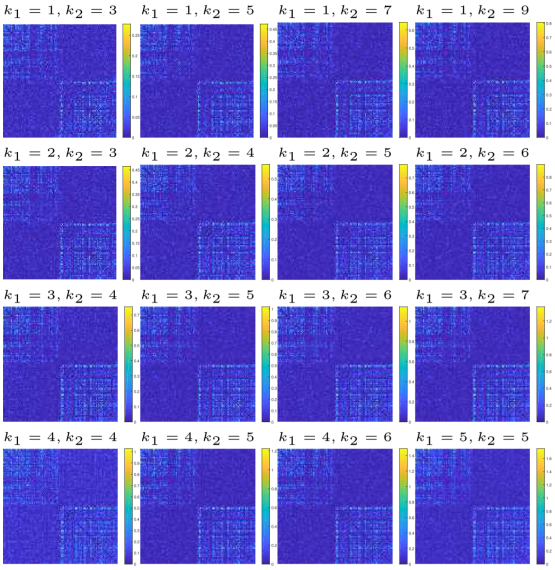


Figure 7: The affinity $|\mathrm{Tr}\left[B_{k_1,k_2,t}(i,j)\right]|$ for 2 clusters with SO(3) transformations at different frequencies $k_1, k_2$. SNR $p = 0.16$.

**Spectral Clustering with SO(2) or SO(3) Transformations:** We apply our unsupervised co-learning framework on spectral clustering: Given totally $n$ data points with $K$ equal sized clusters. For point $i$ we assign in-plane rotational angle $\alpha_i \in [0, 2\pi)$, or 3-D rotation $\omega_i \in \mathrm{SO}(3)$. Then the ground truth alignment

is $\alpha_{ij} = \alpha_i - \alpha_j$ (SO(2)), or $\omega_{ij} = \omega_i \omega_j^{-1}$ (SO(3)). We build the clean graph by fully connecting nodes within each cluster. The noisy graph is then built following the random rewiring model at a given SNR $p$. We perform clustering by using our proposed affinities as the input of spectral clustering, and compare with the traditional spectral clustering [NJW02, VL07], which only takes into account the scalar edge connection, and VDM [SW12], which defines affinity based on the transformation consistency at single representation. In Tab. 1, we use rand index [Ran71] to measure the performance (larger value is better). Our proposed three affinities achieve similar accuracy and they outperform traditional spectral clustering (scalar) and VDM. The values (mean and standard deviation) reported in Tab. 1 are evaluated over 50 trials for $\mathrm{SO}(2)$ and 10 trials for $\mathrm{SO}(3)$ respectively.

To a better visualization, in Fig. 5 we show the graph adjacency matrices using the VDM affinity, the power spectrum and bispectrum affinities, for the 2 clusters example with $\mathcal{G} = \mathrm{SO}(3)$ transformation ($\mathrm{SO}(2)$ case is similar). At low noise levels as $p > 0.2$, all the three affinities could reveal the 2 clusters. While at high noise levels as $p = 0.16$ or $0.2$, VDM is visually unable to reveal the clusters, whereas our proposed affinities still work. It visually demonstrates that our framework is more robust to noise than the state-of-the-arts.

To better understand the performance and the properties of the power spectrum and bispectrum affinity, in Fig. 6 and Fig. 7, we present the affinities $|\mathrm{Tr}\left[P_{k,t}(i,j)\right]|$ and $|\mathrm{Tr}\left[B_{k_1,k_2,t}(i,j)\right]|$ at individual $k$ (for power spectrum) and $(k_1, k_2)$ (for bispectrum) respectively. For the power spectrum affinity, we observe that with increasing frequency $k$, the 2-cluster structural property is gradually revealed. Similarly, for the bispectrum with $k_1$ and $k_2 > 0$, we find that individual $(k_1, k_2)$ components are able to reveal the underlying 2-cluster structure at high level of noise ($p = 0.16$). These two examples indicate the importance of using high frequencies to make the affinity robust to noise.

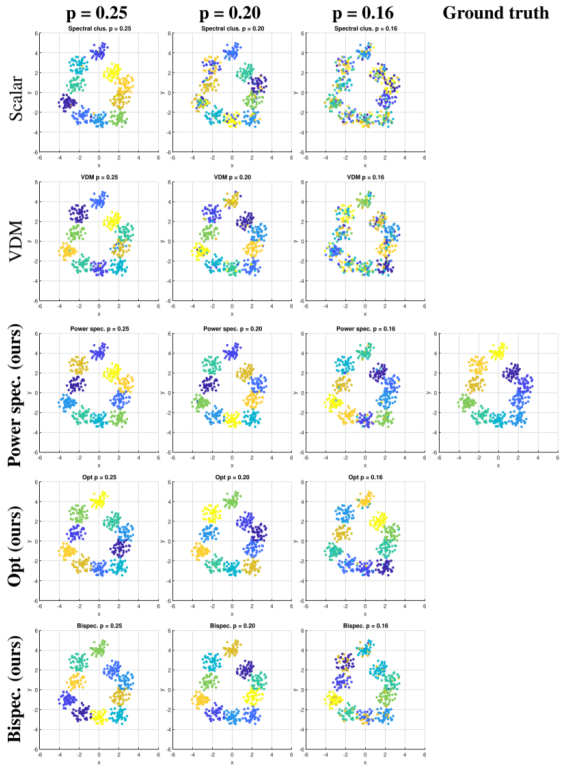


Figure 8: Spectral clustering result with $K = 10$ clusters and SO(2) group transformation, at different noise levels. We have $n = 500$ number of points and each cluster has 50 points.

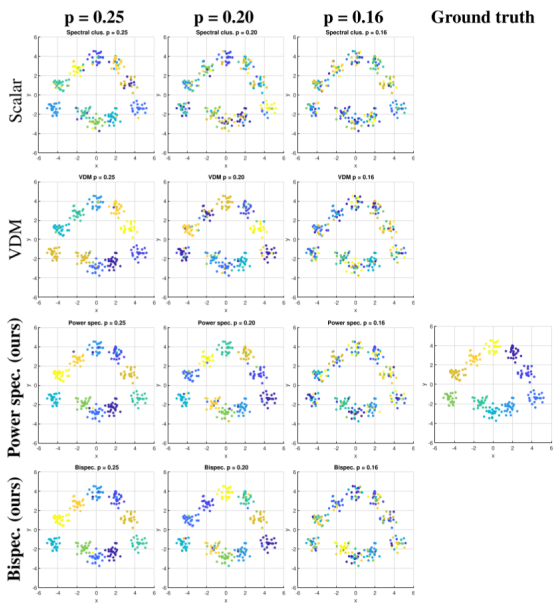


Figure 9: Spectral clustering result with $K = 10$ clusters and SO(3) group transformation, at different noise levels. We have $n = 250$ number of points and each cluster has 25 points.

To visualize the clustering results of $K = 10$, in Figs. 8 and 9, we show the coloring of the clusters in the

Table 2: Parameter tuning for maximum frequency $k_{\max}$: Rand indices of spectral clustering results with $\mathrm{SO}(2)$ group transformation, we set $p = 0.16$, $K = m_k = 10$.

| method | Maximum frequency cutoff $k_{\max}$ | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 5 | 10 | 20 | 50 | 100 |
| Scalar | 0.855 | | | | | |
| **Power spec. (ours)** | 0.897 | 0.944 | 0.965 | 0.978 | 0.982 | 0.984 |
| **Opt (ours)** | **0.898** | **0.948** | **0.970** | **0.981** | **0.989** | **0.993** |
| **Bispec. (ours)** | 0.887 | 0.922 | 0.933 | 0.960 | 0.980 | 0.986 |

2-D scatter plots of the data (clustering accuracies are shown in the main paper), with $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$ group transformations. We generate the scatter plots by first assigning the coordinates for each data point on the 2-D plane (these coordinates are used for visualization and not used for the initial clean graph generation). The top rows of Figs. 8 and 9 show the ground truth coloring of the points based on the clean graphs. For noisy graphs, we assign colors based on the clustering results by each affinity measure (the ordering of the colors may vary from plot to plot). Visually, our proposed affinities achieve more accurate clustering than vector diffusion maps (VDM) [SW12] and traditional spectral clustering [NJW02], with less mixing colors within individual clusters.

**Choice of Parameters:** We test the influence of the maximum frequency cutoff $k_{\max}$ on the performance of spectral clustering and also discuss the choice of $m_k$. In the clean case, the number of non-zero eigenvalues of the weight matrices $W_k$ is $d_k K$ for $K$ clusters and the dimension of the representation is $d_k$ at frequency $k$. Therefore, the matrix has a low-rank structure and we truncate at top $d_k K$ eigenvectors, i.e. $\boldsymbol{m_k = K}$. The noisy graph can be modeled as the clean weight matrices perturbed by random matrices $R_k$,

$$W_k = pW_k^{\text{clean}} + R_k, \tag{5.1}$$

where $R_k$ is a random matrix whose elements are independent and identically distributed (i.i.d) zero mean random variables with finite moments. The top $d_k K$ eigenvectors of $W_k$ have non-trivial correlation with the top eigenvectors of $W_k^{\text{clean}}$ as long as the 2-norm of $R_k$ is not too large. Therefore, we are able to use the top $d_k K$ eigenvectors for clustering. Using $m_k < K$ will lead to loss of information and using $m_k > K$ will include spurious information from noise.

For the maximum frequency cutoff $k_{\max}$, in Tab. 2, we show that the rand indices of clustering results get improved with increasing $k_{\max}$ for all three affinities proposed in the paper. However, using a larger $k_{\max}$ cutoff increases the computational complexities for all three affinity measures and the dimension of the irrep might increase with $k$ (e.g. the dimension of Wigner-D matrix at index $k$ is $2k + 1$), which is undesirable. Therefore, there is a tradeoff between the statistical accuracy and computational complexity. In the main paper, we use a moderate $k_{\max} = 10$.

# 6 Conclusion

In this paper, we propose a novel mathematical and computational framework for unsupervised co-learning on $\mathcal{G}$-manifolds across multiple unitary irreps for robust nearest neighbor search and spectral clustering. We have a two stage algorithm. At the first stage, the graph adjacency matrices are individually denoised through spectral filtering. This step uses the local cycle consistency of the group transformation. The second stage checks the algebraic consistency over different irreps and we propose three different ways to combine the information across all irreps. Using invariant moments bypass the pairwise alignment and are computationally more efficient than the affinity based on optimal alignment search. Experimental results show efficacy of the framework compared to the state-of-the-arts which do not take into account of the transformation or only use single representation. This general framework can be applied to many other applications, such as multi-frame alignment in computer vision [BGH+18]. For future work, we will take into account node-level noise and other noise models.

# References


[AA93] Andrey V Alekseevsky and Dmitry V Alekseevsky. Riemanniang-manifold with one-dimensional orbit space. *Annals of global analysis and geometry*, 11(3):197–211, 1993. 1

[BBEKY13] Derek Bean, Peter J. Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013. 1

[BBK08] Alexander Bronstein, Michael Bronstein, and Ron Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer Publishing Company, Incorporated, 1 edition, 2008. 1

[BGH+18] Chandrajit Bajaj, Tingran Gao, Zihang He, Qixing Huang, and Zhenxiao Liang. SMAC: Simultaneous mapping and clustering using spectral decompositions. In *International Conference on Machine Learning*, pages 334–343, 2018. 6

[BGV03] Nicole Berline, Ezra Getzler, and Michèle Vergne. *Heat Kernels and Dirac Operators (Grundlehren Text Editions)*. Springer, 1992 edition, 12 2003. 3

[BJZ+16] Katherine L Bouman, Michael D Johnson, Daniel Zoran, Vincent L Fish, Sheperd S Doeleman, and William T Freeman. Computational imaging for vlbi image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 913–922, 2016. 10

[BKSW18] Paul Breiding, Sara Kališnik, Bernd Sturmfels, and Madeleine Weinstein. Learning algebraic varieties from samples. *Revista Matemática Complutense*, 31(3):545–593, 2018. 1

[BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998. 1, 2

[BN02] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2002. 1, 2

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 2003. 1, 2

[BNS06] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006. 1, 2

[Bri65] David R Brillinger. An introduction to polyspectra. *The Annals of mathematical statistics*, pages 1351–1374, 1965. 10

[CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006. 1, 2, 4.2

[CLL+05a] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–31, may 2005. 1

[CLL+05b] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7432–7, may 2005. 1

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009. 1

[CSPW09] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Sparse and low-rank matrix decompositions. *IFAC Proceedings Volumes*, 42(10):1493–1498, 2009. 1

[CWB11] Minmin Chen, Kilian Q. Weinberger, and John C. Blitzer. Co-training for domain adaptation. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 2456–2464, USA, 2011. Curran Associates Inc. 2

[DG03] David L. Donoho and Carrie Grimes. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003. 1

[Feh10] Janis Fehr. Local rotation invariant patch descriptors for 3d vector fields. In *2010 20th International Conference on Pattern Recognition*, pages 1381–1384. IEEE, 2010. 10

[FHM+06] Jason Farquhar, David Hardoon, Hongying Meng, John S. Shawe-Taylor, and Sándor Szedmák. Two view learning: Svm-2k, theory and practice. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 355–362. MIT Press, 2006. 2

[FZ19a] Yifeng Fan and Zhizhen Zhao. Cryo-electron microscopy image analysis using multi-frequency vector diffusion maps. *arXiv preprint arXiv:1904.07772*, 2019. 1, 2

[FZ19b] Yifeng Fan and Zhizhen Zhao. Multi-frequency vector diffusion maps. In *ICML*, 2019. 1, 2, 4.3, 4.4

[Gao16] Tingran Gao. The diffusion geometry of fibre bundles: Horizontal diffusion maps. *arXiv preprint arXiv:1602.02330*, 2016. 1, 2

[GFZ19] Tingran Gao, Yifeng Fan, and Zhizhen Zhao. Representation theoretic patterns in multi-frequency class averaging for three-dimensional cryo-electron microscopy. *arXiv preprint arXiv:1906.01082*, 2019. 2

[GZ19] Tingran Gao and Zhizhen Zhao. Multi-frequency phase synchronization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2132–2141, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 2

[Hal15] Brian Hall. *Lie groups, Lie algebras, and representations: an elementary introduction*, volume 222. Springer, 2015. 4.7

[HS11a] R. Hadani and A. Singer. Representation theoretic patterns in three-dimensional cryo-electron microscopy II – the class averaging problem. *Foundations of Computational Mathematics*, 11(5):589–616, 2011. 2

[HS11b] Ronny Hadani and Amit Singer. Representation Theoretic Patterns in Three Dimensional Cryo-Electron Microscopy I: The Intrinsic Reconstitution Algorithm. *Annals of Mathematics*, 174(2):1219–1241, 2011. 2

[Ken89] David G. Kendall. A survey of the statistical theory of shape. *Statist. Sci.*, 4(2):87–99, 05 1989. 1

[Ket71] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 12 1971. 2

[KI11] Abhishek Kumar and Hal Daume III. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 393–400, USA, 2011. Omnipress. 2

[KM10] Ramakrishna Kakarala and Dansheng Mao. A theory of phase-sensitive rotation invariance with spherical harmonic and moment-based representations. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 105–112. IEEE, 2010. 10

[Kob12] Shoshichi Kobayashi. *Transformation groups in differential geometry*. Springer Science & Business Media, 2012. 1

[Kon07] Risi Kondor. A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum. *arXiv preprint cs/0701127*, 2007. 10

[Kon08] Imre Risi Kondor. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008. 10

[KR08] Peter J Kostelec and Daniel N Rockmore. Ffts on the rotation group. *Journal of Fourier analysis and applications*, 14(2):145–179, 2008. 4.7

[KRD11] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1413–1421. Curran Associates, Inc., 2011. 2

[LMQW18] Chen-Yun Lin, Arin Minasian, Xin Jessica Qi, and Hau-Tieng Wu. Manifold learning via the principle bundle approach. *Frontiers in Applied Mathematics and Statistics*, 4:21, 2018. 1, 3

[LS18] Boris Landa and Yoel Shkolnisky. The steerable graph laplacian and its application to filtering image datasets. *SIAM Journal on Imaging Sciences*, 11(4):2254–2304, 2018. 1, 3

[LYZ18] Y. Li, M. Yang, and Z. M. Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2018. 1, 2

[MBM16] Hà Quang Minh, Loris Bazzani, and Vittorio Murino. A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17(25):1–72, 2016. 2

[MFK94] David Mumford, John Fogarty, and Frances Kirwan. *Geometric invariant theory*, volume 34. Springer Science & Business Media, 1994. 1

[Mic08] Peter W Michor. *Topics in differential geometry*, volume 93. American Mathematical Soc., 2008. 1

[MMK06] Ion Muslea, Steven Minton, and Craig A. Knoblock. Active learning with multiple views. *J. Artif. Int. Res.*, 27(1):203–233, October 2006. 2

[MR97] David K Maslen and Daniel N Rockmore. Generalized ffts—a survey of some recent results. In *Groups and Computation II*, volume 28, pages 183–287. American Mathematical Soc., 1997. 4.3

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002. 5, 5

[NLKC06] Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald R Coifman. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In *Advances in neural information processing systems*, pages 955–962, 2006. 1

[OWNB17] Greg Ongie, Rebecca Willett, Robert D. Nowak, and Laura Balzano. Algebraic variety models for high-rank matrix completion. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2691–2700, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 1

[PZF96] Pawel A Penczek, Jun Zhu, and Joachim Frank. A common-lines based method for determining orientations for $n > 3$ particle projections simultaneously. *Ultramicroscopy*, 63(3-4):205–218, 1996. 1

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. 5

[Ric01] Ken Richardson. The transverse geometry of $g$-manifolds and riemannian foliations. *Illinois Journal of Mathematics*, 45(2):517–535, 2001. 1

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000. 1, 2

[RST09] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009. 4.6

[RWY12] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for high-dimensional sparse additive models over kernel classes. *Journal of Machine Learning Research*, 13:281–319, 2012. 1

[Sch08] Alexander HW Schmitt. *Geometric invariant theory and decorated principal bundles*, volume 11. European Mathematical Society, 2008. 1

[SH10] Shiliang Sun and David R. Hardoon. Active learning with extremely sparse labeled examples. *Neurocomputing*, 73(16):2980 – 2988, 2010. 10th Brazilian Symposium on Neural Networks (SBRN2008). 2

[SN05] Vikas Sindhwani and Partha Niyogi. A co-regularized approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005. 2

[SR08] Vikas Sindhwani and David S Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983. ACM, 2008. 1, 2

[Sun13] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7):2031–2038, Dec 2013. 1, 2

[SW12] Amit Singer and Hau-Tieng Wu. Vector Diffusion Maps and the Connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. 1, 2, 3, 5, 5, 5, 5

[SW16] Amit Singer and Hau-Tieng Wu. Spectral convergence of the connection Laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 12 2016. 1, 2, 3

[SZSH11] A. Singer, Z. Zhao, Y. Shkolnisky, and R. Hadani. Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences*, 4(2):723–759, 2011. 1, 2

[TSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000. 1, 2

[TWH15] Robert Tibshirani, Martin Wainwright, and Trevor Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015. 1

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018. 1

[VG18] Elif Vural and Christine Guillemot. A study of the classification of low-dimensional data with supervised manifold learning. *Journal of Machine Learning Research*, 18:1–55, 2018. 1

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 5

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. 1

[Wig32] Eugene Paul Wigner. Gruppentheorie und ihre anwendung auf die quantenmechanik der atomspektren. *Monatshefte für Mathematik und Physik*, 39(1):A51, 1932. 4.7

[WK00] Limin Wang and Marc Kamionkowski. Cosmic microwave background bispectrum and inflation. *Physical Review D*, 61(6):063504, 2000. 10

[XTX13] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013. 2

[ZS14] Zhizhen Zhao and Amit Singer. Rotationally invariant image representation for viewing direction classification in cryo-EM. *Journal of structural biology*, 186(1):153–166, 2014. 1, 10, 5

[ZSS16] Zhizhen Zhao, Yoel Shkolnisky, and Amit Singer. Fast steerable principal component analysis. *IEEE transactions on computational imaging*, 2(1):1–12, 2016. 5

[ZXXS17] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43 – 54, 2017. 2