Manifold Alignment by Feature Correspondence

Jay S. Stanley III
Comp. Bio. & Bioinfo. Prog.
Yale University
New Haven, CT, USA
jay.stanley@yale.edu

Scott Gigante
Comp. Bio. & Bioinfo. Prog.
Yale University
New Haven, CT, USA
scott.gigante@yale.edu

Guy Wolf *
Dept. of Math. and Stat.
Université de Montréal
Montreal, QC, Canada
guy.wolf@umontreal.ca

Smita Krishnaswamy * ✉
Depts. of Genetics & Comp. Sci.
Yale University
New Haven, CT, USA
smita.krishnaswamy@yale.edu

Abstract

We propose a novel framework for combining datasets via alignment of their intrinsic geometry. This alignment can be used to fuse data originating from disparate modalities, or to correct batch effects while preserving intrinsic data structure. Importantly, we do not assume any pointwise correspondence between datasets, but instead rely on correspondence between a (possibly unknown) subset of data features. We leverage this assumption to construct an isometric alignment between the data. This alignment is obtained by relating the expansion of data features in harmonics derived from diffusion operators defined over each dataset. These expansions encode each feature as a function of the data geometry. We use this to relate the diffusion coordinates of each dataset through our assumption of partial feature correspondence. Then, a unified diffusion geometry is constructed over the aligned data, which can also be used to correct the original data measurements. We demonstrate our method on several datasets, showing in particular its effectiveness in biological applications including data fusion and batch effect removal.

1 Introduction

Manifold alignment aims to map disparate datasets into a common representation, under the assumption that the datasets all originate from noisy sampling of a common manifold determined by the data generation process. Under this assumption, the intrinsic geometry of the data should be similar across datasets, and global differences between the geometry of different datasets are considered as noise or data collection artifacts. Therefore, a common representation, which aligns the datasets from their original measurements onto a shared data manifold, both eliminates such artifacts and recovers clean intrinsic relations across measured datasets. Furthermore, such alignment enables data fusion and transfer of knowledge between separate domains when datasets are taken from different data collection environments (e.g., different sensors, technologies, or subjects in biomedical data).

This problem is particularly relevant in biomedical data analysis nowadays, with the advent of many modalities of high throughput measurements generated in biomedical and cellular systems. For example, a population of cells can be measured using both single cell ATAC-sequencing (scATAC-seq)[1], which measures regions of open chromatin DNA in each cell, and single cell RNA-sequencing (scRNA-seq) [2], which reveals gene expression profiles of cells. However, since both measurements

*These authors contributed equally; ✉ Corresponding author.

Preprint. Under review.

are destructive, they are not measured on the same cells but rather from cells sampled from the same population. To integrate such datatypes it is important to register neighboring (i.e., likely matching) cells from both measurements. We note that often we can expect such registration to be feasible, even for different technologies, since their measurements reflect related properties capturing the underlying state of equivalent cell populations. For example, in the scATAC-seq/scRNA-seq case, if a gene is expressed then it *must* be in an open chromatin region, so the resulting measurements are different but not independent of each other.

Furthermore, even when using the same measurement technology, biomedical data is often systematically different based on machine calibration, day-to-day temperature variation and underlying background biological differences between individuals or staining and treatment differences [3]. For example, patients with kidney disease at two different hospitals may have creatinine readings in two different ranges, simply due to machine calibration differences [4], which are not always known a priori. As a result, studying real experimental differences, comparing outcomes across patient cohorts, and generalizing results to different hospitals is challenging, if not impossible, without proper alignment to make varied collection environments comparable while alleviating such systematic *batch effects*.

Here, we propose an alignment approach that explicitly takes advantage of the typical correspondence between the underlying features quantified by measurement and data collection systems that can be aligned. Indeed, related systems often observe similar "entities" (e.g., cells, patients) and aim to capture related properties in them. Our approach uses graph signal processing tools (see Sec. 3) to relate measured data features (seen as graph or manifold signals) to intrinsic coordinates over the intrinsic geometry of each dataset, which are revealed via diffusion maps [5]. Then, as explained in Secs. 4 and 5, we leverage feature correspondence to capture pairwise relations between the intrinsic diffusion map coordinates of the separate data manifolds (i.e., of each dataset). Finally, we use these relations to compute an isometric transformation that aligns the data manifolds on top of each other without distorting their internal structure.

We demonstrate the results of our method on artificial manifolds created from corrupted MNIST digits, and single-cell biological data for both batch effect removal and multimodal data fusion. In each case, our method successfully aligns the data manifolds to recover appropriate data neighborhoods both within and across the two datasets. Further, we show an application of our approach in transfer learning by applying a lazy classifier to one unlabeled dataset based on labels provided by another dataset (with batch effects between them), and compare lazy classification accuracy before and after alignment. Finally, comparisons with recently developed methods such as the MNN-based method from [3] and the GAN-based method from [6] show significant improvements in alignment and neighborhood recovery achieved by our harmonic alignment methods.

2 Related Work

Algorithms for semi-supervised and unsupervised manifold alignment exist in classical statistics [7, 8], deep learning [9, 10, 6] and manifold learning [3, 11, 12]. A classic method for aligning linear structures (such as ones captured by PCA) is canonical correlation analysis (CCA) [8], which can be used to project two datasets on a common basis formed by directions that maximize feature correlation between them. More recently, a linear manifold alignment method was presented in [12] based on embedding a joint graph built over both datasets to preserve local structure in both manifolds. This method provides a mapping from both original features spaces to a new feature space defined by the joint graph, which is shared by both datasets with no assumption of feature correspondence. Finally, in biomedical data analysis, mutual nearest neighbors (MNN) batch correction [3] focuses on families of manifold deformations that are often encountered in biomedical data. There, locally linear manifold alignment is provided by calculating a correction vector for each point in the data, as defined by the distances from the point to all points for which it is a mutual k-nearest neighbor. This correction vector is then smoothed by taking a weighted average over a Gaussian kernel.

Beyond manifold learning settings, deep learning methods have been proposed to provide alignment and transfer learning between datasets. For example, cycle generative adversarial networks (Cycle GANs) [9, 10] are a class of deep neural network in which a generative adversarial network (GAN) is used to learn a nonlinear mapping from one domain to another, and then a second GAN is used to map back to the original domain. These networks are then optimized to (approximately) satisfy

cycle consistency constraints such that the result of applying the full cycle to a data point reproduces the original point. Manifold Aligning GAN (MAGAN) [6] is a particular cycle GAN that adds a supervised partial feature correspondence to enforce alignment of two data manifolds over the mapping provided by the trained network.

In contrast, this work provides a nonlinear method for aligning two datasets using their diffusion maps under the assumption of a partial feature correspondence. However, unlike MAGAN, we do not need to know which features correspond. In doing so, we obtain more information from datasets with partial feature correspondence than methods that assume no correspondence, but without the burden of determining in advance which or how many features correspond. To evaluate out method, in Section 6 we compare our method to MAGAN, as a leading representative of deep learning approaches, and MNN, as a leading representative of manifold learning approaches. We note that to the best of our knowledge, the method in [12] is not provided with standard implementation, and our attempts at implementing the algorithm have significantly underperformed other methods. For completeness, partial comparison to this method is demonstrated in Appendix F.

3 Preliminaries

Manifold Learning High dimensional data can often be conceptually modeled as originating from an intrinsically low dimensional manifold that is mapped via nonlinear functions to observable high dimensional measurements; this is commonly referred to as the manifold assumption. Formally, given a dataset $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$ of high dimensional observations, manifold learning methods assume its data points originate from a sampling $Z = \{z_i\}_{i=1}^N \in \mathcal{M}^d$ of the underlying manifold mapped via a nonlinear function $x_i = \mathbf{f}(z_i)$, $i = 1, \ldots, n$ to the high dimensional feature space. Then, these methods aim to learn a low dimensional intrinsic representation that approximates the manifold geometry of $\mathcal{M}^d$ (see, for example, [13–16] and references within).

Diffusion Maps To learn a manifold geometry from collected data, we use the popular diffusion maps construction [5]. This construction starts by considering local similarities, which we quantify via an anisotropic kernel

$$\mathcal{K}(x, y) = \frac{\mathcal{G}(x, y)}{\|\mathcal{G}(x, \cdot)\|_1 \|\mathcal{G}(y, \cdot)\|_1}, \tag{1}$$

where $\mathcal{G}(x, y) = e^{-\frac{\|x-y\|^2}{\sigma}}$ is the Gaussian kernel with neighborhood radius $\sigma > 0$. As shown in [5], this kernel provide neighborhood construction that is robust to sampling density variations and enables separation of data geometry from its distribution. Next, the kernel $\mathcal{K}$ is normalized to define transition probabilities $p(x, y) = \frac{\mathcal{K}(x,y)}{\|\mathcal{K}(x,\cdot)\|_1}$ that define a Markovian diffusion process over the data. Finally, a diffusion map is defined by organizing these probabilities in a row stochastic matrix $\mathbf{P}$ (typically referred to as the diffusion operator) as $\mathbf{P}_{ij} = p(x_i, x_j)$, and using its eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ and (corresponding) eigenvectors $\{\phi_j\}_{j=1}^N$ to map each $x_i \in X$ to diffusion coordinates $\Phi_t(x_i) = [\lambda_1^t \phi_1(x_i), \ldots, \lambda_N^t \phi_N(x_i)]^T$. The parameter t in this construction represents a diffusion time, i.e., the number of transitions considered in the diffusion process. To simplify notations, we also use $\Phi_t = \{\Phi_t(x_i) : x_i \in X\}$ to denote the diffusion map of the entire dataset X. We note that in general, as t increases, most of the eigenvalue weights λ_j^t, $j = 1, \ldots, N$, become numerically negligible, and thus truncated diffusion map coordinates (i.e., using only non-negligible ones) can be used for dimensionality reduction purposes, as discusses in [5].

Graph Fourier Transform A classic result in spectral graph theory (see, e.g., [17]) shows that the discrete Fourier basis (i.e., pure harmonics, such as sines and cosines, organized by their frequencies) can be derived as Laplacian eigenvectors of the ring graphs. More recently, this result was recently used in graph signal processing [18] to define a *graph Fourier transform* (GFT) by treating eigenvectors of the graph Laplacian as generalized Fourier harmonics (i.e., intrinsic sines and cosines over a graph). Further, as discussed in [5, 19], diffusion coordinates are closely related to these Laplacian eigenvectors, and can essentially serve as geometric harmonics over data manifolds. Indeed, the kernel $\mathcal{K}$ can be considered as defining edge weights on a graph whose vertices are the data point in X. It can be verified that for this graph, its normalized graph Laplacian is given by $\boldsymbol{\mathcal{L}} = \mathbf{I} - \mathbf{D}^{1/2}\mathbf{P}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \|\mathcal{K}(x, \cdot)\|_1$. Therefore, the eigenvectors of $\boldsymbol{\mathcal{L}}$ can be written as $\psi_j = D^{1/2}\phi_j$ with corresponding eigenvalues $\omega_j = 1 - \lambda_j$. The resulting GFT of a signal (or

function) f over X can thus be written as $\widehat{f}(\omega_j) = \langle f, \psi_j \rangle = \langle f, D^{1/2}\phi_j \rangle$. We note that here we treat either ω_j or λ_j as providing a "frequency" organization of their corresponding eigenvectors ψ_j or ϕ_j (treated as intrinsic harmonics). In the latter case, eigenvectors with higher eigenvalues correspond to lower frequencies on the data manifold, and vice versa. This frequency-based organization, and more generally the duality between diffusion coordinates and harmonics, will be leveraged here to provide an isometric alignment between the intrinsic coordinates of data manifolds with (partially) corresponding features.

4 Harmonic alignment

Let $\{X^{(s)}\}_{s=1}^{S}$ be a collection of S samples with features $\{f_j^{(s)}\}_{j=1}^{n(s)}$ that also serve as the observed ambient-space coordinates of the data. We assume that at least a subset of these features aim to measure the same quantities in the data, but are also affected by sample-dependent artifacts. Further, while in general some features may be unique for specific samples (e.g., when collected by different technologies), for simplicity we focus here on overlapping features that conceptually correspond with each other across samples. As a result, we assume the number of features is independent of specific sample and can be denoted as $n = n(s)$. The entire data is thus given by $X = \bigcup_{s=1}^{S} X^{(s)}$.

For simplicity, we describe here our proposed approach for aligning two samples $X^{(1)}, X^{(2)}$ and processing data across them, but this can naturally be generalized to any number of samples. In order to find an isometric transformation between the diffusion geometries of the two samples, we compute a *harmonic correlation matrix* $C_{i_1 i_2} = \text{corr}(\psi_{i_1}^{(1)}, \psi_{i_2}^{(2)})$ between the harmonics that serve (up to appropriate weighting) as coordinates of the diffusion maps $\Phi_t^{(1)}$ and $\Phi_t^{(2)}$ of the two samples. Notice that since we do not have any correspondence between data points, the computation of the correlations in C cannot be done directly between the two sets of Laplacian eigenvectors of the two samples when expressed in terms of data points. However, since we assume feature correspondence between samples, we can compute these correlations by expressing these eigenvectors in terms of the GFT of the data features. Namely, for each sample s we construct an $N^{(s)} \times n$ matrix $\widehat{X}^{(s)}$ whose j-th column is $\hat{f}_j^{(s)}$. Then, C can be computed with correlations between rows of $\widehat{X}^{(1)}$ and $\widehat{X}^{(2)}$.

We note that in fact, not all correlations should be computed, since each harmonic $\psi_i^{(s)}$ and corresponding GFT coefficients in $(\widehat{X}_{ij}^{(s)})_{j=1}^{n}$ represent intrinsic variations at a specific frequency expressed by the corresponding Laplacian eigenvalue ω_i. Therefore, we only expect information between similar frequency components to be considered for correlation, while components with very different frequencies can be considered as uncorrelated to begin with. As explained in Sec. 5, we incorporate this understanding in the construction of C by only populating near-diagonal elements in this matrix with correlations that are computed using a sliding bandpass filter that, at each time, selects only a local band of coefficients to be correlated from $\widehat{X}^{(1)}$ and $\widehat{X}^{(2)}$

Given the cross-sample harmonic correlation matrix C, we use its singular value decomposition (SVD) $C = U\Sigma V^T$ to obtain its nearest orthogonal approximation $\mathbf{T} = UV^T$ (e.g., as shown

Algorithm 1 Harmonic Alignment

Require: Dataset $\mathbb{X} = \mathbf{X}^{(1)} \cup \mathbf{X}^{(2)}$ with n features and two samples (i.e., sub-datasets), where each sample $\mathbf{X}^{(s)}$ has $N^{(s)}$ observations

Ensure: Aligned diffusion map $x \mapsto \mathbf{\Phi}_t^{(1,2)}(x)$, $x \in \mathbb{X}$

1: **for** $s \in \{1, 2\}$ **do**
2: Compute the $N^{(s)} \times N^{(s)}$ anisotropic kernel $\mathcal{K}^{(s)}$ (Eq. 1) over the sample $X^{(s)}$.
3: Compute the diffusion operator $\mathbf{P}^{(s)}$, its eigendecomposition, and the corresponding diffusion map $\Phi_t^{(s)}$ (see Sec. 3)
4: **end for**
5: Compute the $N^{(1)} \times N^{(2)}$ bandlimited harmonic correlation matrix $\mathbf{C}$ (see Sec. 5)
6: Orthogonalize via SVD $\mathbf{C} = U\Sigma V^T$ to get $\mathbf{T} = UV^T$
7: Compute the unified diffusion map $\mathbf{\Phi}_t^{(1,2)}$ (Eq. 2)

in [20]), which defines an isometric transformation between the diffusion maps of the two samples. Finally, we now compute a unified diffusion map, which can be written in (block) matrix form as

$$\Phi_t^{(1,2)} = \begin{bmatrix} \Phi_0^{(1)} & \Phi_0^{(1)}\,\mathbf{T} \\ \Phi_0^{(2)}\,\mathbf{T}^T & \Phi_0^{(2)} \end{bmatrix} \begin{bmatrix} \Lambda^{(1)} & 0 \\ 0 & \Lambda^{(2)} \end{bmatrix}^t, \tag{2}$$

where $\Lambda^{(s)}$ are diagonal matrices with the diffusion eigenvalues $\{\lambda_i^{(s)}\}$ as their main diagonal. A succinct summary of the described algorithm, which we call harmonic alignment, is presented in Algorithm 1. While our presentation here is given in terms of two samples for simplicity, it naturally extents to multiple samples by considering $S \times S$ blocks in (2), instead of 2×2 blocks, with isometric transformations $T^{(s_1,s_2)}$ in each (s_1, s_2) block with $s_1 \neq s_2$.

To complete the alignment process, we construct a new kernel over the combined diffusion coordinates in $\Phi_t^{(1,\ldots,S)}$ and build a robust unified diffusion geometry over the entire multi-sample data $X = \bigcup_{s=1}^{S} X^{(s)}$ that is invariant to batch effects and also naturally denoises various sample-specific artifacts. This diffusion geometry can naturally be incorporated in diffusion-based methods for several data processing tasks, such as dimensionality reduction & visualization [21], denoising & imputation [22], latent variable inference [23, 24], and data generation [25]. In Sec. 6.3 we demonstrate the application of harmonic alignment to batch effect removal and multimodal data fusion, and in particular in single-cell data analysis.

5 Bandlimited correlation

As discussed in Sec. 3, diffusion coordinates are organized by frequency and thus, we can constrain the isometric transformation learned in harmonic alignment to maintain the general frequency structure between the two data manifolds. Indeed, if the two manifolds can be aligned, then intrinsic low-frequency trends in one dataset should map to low-frequency ones in the other, and similarly any frequency band should map to an equivalent one across the aligned datasets. Therefore, the frequency structure of diffusion maps already provides a coarse alignment, which we leverage here by only applying our alignment within local frequency bands of the diffusion operator spectrum instead of globally over all diffusion coordinates.

To take advantage of the described frequency structure, we propose to partition harmonic correlations using graph spectral wavelets [26]. Since these wavelets are defined as functions of the Laplacian eigenvalues, they provide a natural extension of the Fourier-based alignment we have proposed. The definition of spectral graph wavelets in the Fourier basis allows one to select a wavelet basis that guarantees (1) uniform frequency response (called a *tight frame*) and (2) smooth partitioning. In particular, we use here the iterated sine wavelets of [27] that are formed by translation and dilation of the generating kernel

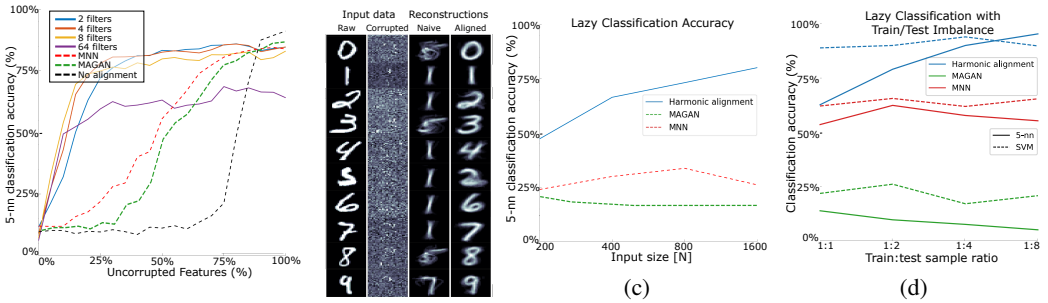
$$g(t) = \sin\left(0.5\pi \cos^2(\pi t)\right), t \in [-0.5, 0.5]. \tag{3}$$

This kernel yields a tight frame with a parameterized overlap between bands, as shown in the appendix (see Figure 5(a) there). In practice, we choose an overlap of $\frac{1}{2}$, which creates smooth frequency response transitions between each wavelet band.

These wavelets are applied in the Fourier domain via element-wise multiplication of the sample itersine kernel $g_\omega^{(s)} : \Lambda \mapsto \mathbb{R}$, which defines the frequency response of the wavelet at scales $\omega \in \Omega$, with the Fourier coefficients of an input signal. As there are $|\Omega|$ wavelet scales, then $|\Omega|$ spectrally partitioned signals are produced by this transform. A correlation matrix is then generated for each scale ω by correlating the GFT of wavelet transformed features (i.e., for the two samples), denoted $\hat{X}_\omega^{(1)}$ and $\hat{X}_\omega^{(2)}$. The resulting correlation is bandlimited and preserves the distribution of manifold harmonics. Finally, each correlation band smoothly transitions to its surrounding bands, and these partial correlations are combined together to obtain sparse inter-band correlations. This is done by summing all bands together to obtain a full "fuzzy-diagonal" correlation matrix $C = \sum_{\omega \in \Omega} \hat{X}_\omega^{(1)} \hat{X}_\omega^{(2)T}$, which is composed of overlapping bandlimited correlations (see Figure 5(c) in the appendix). For further discussion of this construction, and the robustness it provides to signal noise that degrades nonlimited correlations, we refer the reader to Appendix D .

6 Empirical results

6.1 Artificial feature corruption



(a) (b) (c) (d)

Figure 1: Recovery of k-nearest neighborhoods under feature corruption. Mean over 3 iterations is reported for each method. (a) At each iteration, two sets $X^{(1)}$ and $X^{(2)}$ of 1000 points were sampled from MNIST. $X^{(2)}$ was then distorted by a 784×784 corruption matrix $\mathbf{O}_p$ for various identity percentages p (see Section 6.1). Subsequently, a lazy classification scheme was used to classify points in $X^{(2)}\mathbf{O}_p$ using a 5-nearest neighbor vote from $X^{(1)}$. Results for harmonic alignment with different filterbank sizes (see Sec. 5), mutual nearest neighbors (MNN), and classification without alignment are shown. (b) Reconstruction of digits with only 25% uncorrupted features. Left: Input digits. Left middle: 75% of the pixels in the input are corrupted. Right middle: Reconstruction without harmonic alignment. Right: Reconstruction after harmonic alignment. (c) Lazy classification accuracy relative to input size with unlabeled randomly corrupted digits with 35% preserved pixels. (d) Transfer learning performance. For each ratio, 1K uncorrupted, labeled digits were sampled from MNIST, and then 1K, 2K, 4K, and 8K (x-axis) unlabeled points were sampled and corrupted with 35% column identity.

To demonstrate the alignment provided by our method, we assess its ability to recover k-nearest neighborhoods after random feature corruption, and compare it to MNN [3] and MAGAN [6], which are leading manifold- and deep-learning methods respectively, as discussed in Sec. 2. To this end, we drew two random samples $X^{(1)}$ and $X^{(2)}$ of $N^{(1)} = N^{(2)} = 1,000$ MNIST digit images. Then, for each trial, we drew 784^2 samples from a unit-variance Normal distribution to create a 784×784 random matrix. We orthogonalized this matrix to yield the corruption matrix $\mathbf{O}_0$. To vary the amount of feature corruption, we produced partial corruption matrices $\mathbf{O}_p$ (for several values of p) by randomly substituting $p\%$ of the columns in $\mathbf{O}_0$ with columns of the 784×784 identity matrix. Right multiplication of $X^{(2)}$ by these matrices yields corrupted images with only $p\%$ preserved pixels (see Figure 1(b), 'Corrupted'). To assess the recovery of k-nearest neighborhoods, we performed lazy classification on digits (i.e., rows) in $X^{(2)}\mathbf{O}_p$ by only using the labels of neighbors from $X^{(1)}$. The results of this experiment, performed for $p = \{0, 5, 10, \ldots 95, 100\}$, are reported in Figure 1(a). For robustness, at each p we sampled three different non-overlapping pairs $X^{(1)}, X^{(2)}$, and for each pair we sampled three $\mathbf{O}_p$ matrices, each with random identity columns, for a total of nine trials per p. It should be noted that while we report results in terms of mean classification accuracy, we do not aim to provide an optimal classifier here. Our evaluation merely aims to provide a quantitative assessment of neighborhood quality before and after alignment. We regard a lazy learner as ideal for such evaluation since it directly exposes the quality of data neighborhoods, rather than obfuscate it via a trained model. Results for harmonic alignment in conjunction with SVM classifier (in transfer learning settings) are discussed in Sec. 6.2 and demonstrated in Figure 1(d).

In general, none of the methods recovers k-nearest neighborhoods under total corruption, showing 10% accuracy for very small p, essentially giving random chance accuracy given that MNIST has ten classes. Note that in our case, it clearly violates our (partial) feature correspondence assumption. However, when using sufficiently many bandlimited filters, our harmonic alignment quickly recovers over 80% accuracy and consistently outperforms both MNN and MAGAN, except under under very

high correspondence (i.e., when $\mathbf{O}_p \approx \mathbf{I}$). The method proposed by [12] was excluded since it did not show improvement over unaligned classification, but is discussed in Appendix F for completeness.

Next, we examined the ability of harmonic alignment to reconstruct the corrupted data (see Figure 1(b)). We performed the same corruption procedure as before with $p = 25\%$ and selected ten examples of each MNIST digit. Ground truth from $X^{(2)}$ and corrupted result $X^{(2)}\mathbf{O}_{25}$ are shown in Figure 1(b). Then, reconstruction was performed by setting each pixel in a new image to the dominant class average of the $k = 10$ nearest neighbors from $X^{(1)}$. In the unaligned case, we see that most examples turn into smeared fives or ones; this is likely a random intersection formed by $X^{(1)}$ and $X^{(2)}\mathbf{O}_{25}$ (e.g., accounting for the baseline random chance classification accuracy). On the other hand, the reconstructions produced by harmonic alignment resemble their original input examples.

Finally, in Figure 1(c), we consider the affect of data size on obtained alignment. To this end, we fix $p = 35\%$ and vary the size of the two aligned datasets. We compare harmonic alignment, MNN, and MAGAN on input sizes range from 200 to 1600 MNIST digits, while again using lazy classification accuracy to measure neighborhood preservation and quantify alignment quality. The results in Figure 1(d) show that both MNN and MAGAN are not significantly affected by dataset size, and in particular do not improve with additional data. Harmonic alignment, on the other hand, not only outperforms them significantly – its alignment quality increases monotonically with input size.

6.2 Transfer learning

An interesting use of manifold alignment algorithms is transfer learning. In this setting, an algorithm is trained to perform well on a small (e.g., pilot) dataset, and the goal is to extend the algorithm to a new larger dataset (e.g., as more data is being collected) after alignment. In Figure 1(d) we explore the utility of harmonic alignment in transfer learning and compare it to MNN [3] and MAGAN [6]. In this experiment, we first randomly selected $1,000$ uncorrupted examples of MNIST digits, and constructed their diffusion map to use as our training set. Next, we took 65%-corrupted unlabeled points (see Section 6.1) in batches of $1,000$, $2,000$, $4,000$, and $8,000$, as a test set for classification using the labels from the uncorrupted examples. As shown in 1(d), with a 5-nearest neighbor lazy classifier, harmonic alignment consistently improves as the dataset gets larger, even with up to *eight* test samples for every one training sample. When the same experiment is performed with a linear SVM, harmonic alignment consistently outperforms other methods with performance being independent of test set size (or train-to-test ratio). This is due to the increased robustness and generalization capabilities of trained SVM. Further discussion of transfer learning is given in Appendix E . In addition to showing the use of manifold alignment in transfer learning, this example also demonstrates the robustness of our algorithm to imbalance between samples.

6.3 Biological data

6.3.1 Batch effect correction

To illustrate the need for robust manifold alignment in computational biology, we turn to a simple real-world example obtained from [28] (see Figure 2). This dataset was collected by mass cytometry (CyTOF) of peripheral blood mononuclear cells (PBMC) from patients who contracted dengue fever. Subsequently, the Montgomery lab at Yale University experimentally introduced these PBMCs to Zika virus strains.

The canonical response to dengue infection is upregulation of interferon gamma (IFNγ), as discussed in [29–31]. During early immune response, IFNγ works in tandem with acute phase cytokines such as tumor necrosis factor alpha (TNFα) to induce febrile response and inhibit viral replication [32]. We thus expect to see upregulation of these two cytokines together, which we explore in Figure 2.

In Figure 2(a), we show the relationship between IFNγ and TNFα without denoising. Note that there is a substantial difference between the IFNγ distributions of sample 1 and sample 2 (Earth Mover's Distance (EMD) = 2.699). In order to identify meaningful relationships in CyTOF data, it is common to denoise it first [16]. We used a graph low-pass filter proposed in [22] to denoise the cytokine data. The results of this denoising are shown in Figure 2(b). This procedure introduced more technical artifacts by enhancing the difference between batches, as seen by the increased EMD (3.127) between the IFNγ distributions of both patients. This is likely due to a substantial connectivity difference between the two batch subgraphs in the total graph of the combined dataset.

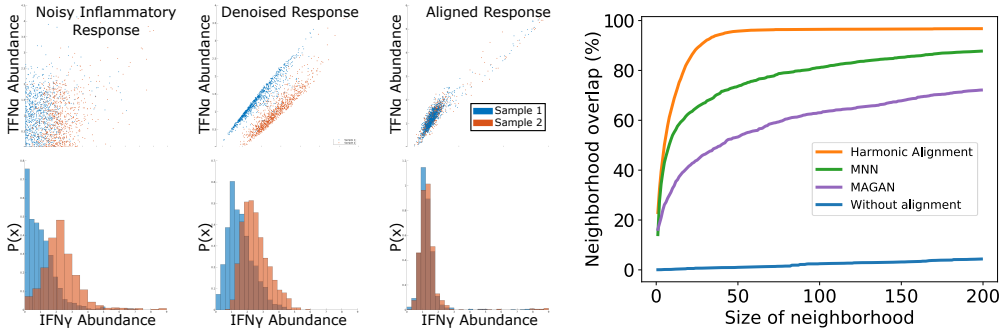


Figure 2: (a)-(c) *Batch effect removal.* 4K cells were subsampled from two single-cell immune profiles obtained via mass cytometry on blood samples of two patients infected with Dengue fever. *Top:* Both patients exhibit heightened IFNγ (x-axis), a pro-inflammatory cytokine associated with tumor necrosis factor alpha (TNFα, y-axis) *Bottom:* IFNγ histograms for each batch. (a) Data before denoising. (b) Denoising of unaligned data enhances a technical effect between samples in IFNγ. (c) Harmonic alignment corrects the IFNγ shift. (d) *Multimodal data fusion.* Percentage overlap of cell neighborhoods from joint gene expression and chromatin profiling of single cells. Harmonic alignment most accurately recovers the pointwise relationship between the two manifolds.

Next, we performed harmonic alignment of the two patient profiles. We show the results of this in Figure 2(c). Harmonic alignment corrected the difference between IFNγ distributions and restored the canonical correlation of IFNγ and TNFα (EMD=0.135). This example illustrates the utility of harmonic alignment for biological data, where it can be used for integrated analysis of data collected across different experiments, patients, and time points.

6.3.2 Multimodal Data Fusion

Since cells contain numerous types of components that are informative of their state (genes, proteins, epigenetics), modern experimental technologies are starting to obtain measurements of each of these components from different assays at the single cell level. Since most single-cell assays are destructive, it is challenging or impossible to obtain all desired measurements in the same cells. It is therefore desirable to perform each assay on a different subset of cells from a single sample, and align these datasets *in silico* in order to obtain a pseudo-joint profile of the multiple data types.

To demonstrate the utility of harmonic alignment in this setting, we use a dataset obtained from [33] of 11,296 cells from adult mouse kidney collected by a joint measurement technique named sci-CAR, which measures *both* gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) in the same cells simultaneously. The datasets are normalized separately as in [22], using a square root transformation for the scRNA-seq and a log transformation with a pseudocount of 1 for the scATAC-seq data, and finally the dimensionality of each dataset is reduced to 100 using truncated SVD. After randomly permuting the datasets to scramble the correspondence between them, we align the two manifolds in order to recover the known bijection between data modalities. Let $f(i) \in F$ be the scRNA-seq measurement of cell i, and $g(i) \in G$ be the scATAC-seq measurement of cell i. Figure 2(d) shows the average percentage overlap of the neighborhood of $f(i)$ in F with the neighborhood of $g(i)$ in F before and after alignment with: MAGAN, MNN and Harmonic Alignment. Harmonic Alignment most accurately recovers the cell neighborhoods, thereby allowing the generation of *in silico* joint profiles across data types and obviating the need for expensive or infeasible *in vitro* joint profiling.

7 Conclusion

We presented a novel method for aligning or batch-normalizing datasets, which involves learning and aligning their intrinsic geometries. Our method is based on the principle that corresponding features across samples or datasets should have similar "frequency" components on these intrinsic

data geometries, represented as manifolds. Our *harmonic alignment* leverages this understanding to compute cross-dataset similarity between manifold harmonics, which is then used to construct an isometric transformation that aligns the data manifolds. Results show that our method successfully aligns artificially misaligned samples, as well as biological data containing batch effects. Our method has the advantage of aligning data geometry rather than density, and thus, it is insensitive to sampling differences. Further, our method inherently denoises the data as it obtains alignment of significant manifold dimensions rather than noise. We expect future applications of harmonic alignment to include, for example, the use of multimodal data fusion to understand complex molecular processes through three or more different data modalities.

References

[1] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486, 2015.

[2] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[3] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.

[4] Pierre Delanaye, Etienne Cavalier, Jean-Paul Cristol, and Joris R Delanghe. Calibration and precision of serum creatinine and plasma cystatin c measurement: impact on the estimation of glomerular filtration rate. *Journal of nephrology*, 27(5):467–475, 2014.

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

[6] Matthew Amodio and Smita Krishnaswamy. Magan: Aligning biological manifolds. *arXiv preprint arXiv:1803.00385*, 2018.

[7] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[8] Bruce Thompson. *Canonical correlation analysis: Uses and interpretation*. Number 47 in Quantitative applications in the social sciences. Sage, 1984.

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251, 2017. doi: 10.1109/ICCV.2017.244. URL `https://doi.org/10.1109/ICCV.2017.244`.

[10] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1857–1865, 2017. URL `http://proceedings.mlr.press/v70/kim17a.html`.

[11] Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1120–1127. ACM, 2008.

[12] Chang Wang and Sridhar Mahadevan. Manifold alignment without correspondence. In *IJCAI*, volume 2, page 3, 2009.

[13] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.

[14] Alan Julian Izenman. Introduction to manifold learning. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(5):439–446, 2012.

[15] Binbin Lin, Xiaofei He, and Jieping Ye. A geometric viewpoint of manifold learning. *Applied Informatics*, 2(1):3, 2015.

[16] Kevin R Moon, Jay Stanley, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 2017.

[17] Robert Brooks, Carolyn Gordon, and Peter A Perry, editors. *Geometry of the Spectrum*, volume 173 of *Proc. 1993 Joint Summer Res. Conf. on Spectral Geometry, University of Washington, Seattle, July 17-23, 1993; Contemporary Mathematics*. American Mathematical Society, Providence, 1994.

[18] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[19] 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.

[20] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[21] Kevin R Moon, David van Dijk, Zheng Wang, Daniel Burkhardt, William Chen, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, Natalia B Ivanova, Guy Wolf, et al. Visualizing transitions and structure for high dimensional data exploration. *bioRxiv*, page 120378, 2017.

[22] David van Dijk, Roshan Sharma, Juoas Nainys, Kristina Yim, Pooja Kathail, Ambrose Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Krishnaswamy Smita, and Data Pe'er. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716 – 729.e27, 2018. doi: 10.1016/j.cell.2018.05.061.

[23] Laleh Haghverdi, Maren Buettner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845, 2016.

[24] Roy R Lederman and Ronen Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018.

[25] Ofir Lindenbaum, Jay S Stanley III, Guy Wolf, and Smita Krishnaswamy. Data generation based on diffusion geometry. *arXiv preprint arXiv:1802.04927*, 2018.

[26] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[27] Nathanaël Perraudin, Johan Paratte, David Shuman, Lionel Martin, Vassilis Kalofolias, Pierre Vandergheynst, and David K. Hammond. GSPBOX: A toolbox for signal processing on graphs. *ArXiv e-prints*, August 2014.

[28] Matthew Amodio, David van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, Anita Desai, Ravi V., Priti Kumar, Ruth Montgomery, Guy Wolf, and Smita Krishnaswamy. Exploring single-cell data with deep multitasking neural networks. *bioRxiv*, 2018. doi: 10.1101/237065.

[29] David A Chesler and Carol Shoshkes Reiss. The role of ifn-γ in immune responses to viral infections of the central nervous system. *Cytokine & growth factor reviews*, 13(6):441–454, 2002.

[30] Anita Chakravarti and Rajni Kumaria. Circulating levels of tumour necrosis factor-alpha & interferon-gamma in patients with dengue & dengue haemorrhagic fever during an outbreak. *Indian Journal of Medical Research*, 123(1):25, 2006.

[31] Elzinandes LA Braga, Patrícia Moura, Luzia MO Pinto, Sonia Ignácio, Maria José C Oliveira, Marly T Cordeiro, and Claire F Kubelka. Detection of circulant tumor necrosis factor-alpha, soluble tumor necrosis factor p75 and interferon-gamma in brazilian patients with dengue fever and dengue hemorrhagic fever. *Memorias do Instituto Oswaldo Cruz*, 96(2):229–232, 2001.

[32] Yoshihiro Ohmori, Robert D Schreiber, and Thomas A Hamilton. Synergy between interferon-γ and tumor necrosis factor-α in transcriptional activation is mediated by cooperation between signal transducer and activator of transcription 1 and nuclear factor κb. *Journal of Biological Chemistry*, 272(23):14899–14907, 1997.

[33] Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

[34] David I Shuman, Benjamin Ricaud, and Pierre Vandergheynst. Vertex-frequency analysis on graphs. *Applied and Computational Harmonic Analysis*, 40(2):260–291, 2016.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Pearson correlation as a measure of matrix diagonality

Given an $d \times d$ matrix A of frequencies, we can measure its diagonality by the sample correlation of the rows with the columns, noting that for a perfectly diagonal matrix, the rows and columns will be identical and hence the correlation will be 1. First note that we can write the sample correlation of two draws from a random distribution as

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2}\sqrt{n \sum y^2 - (\sum y)^2}} \tag{4}$$

We now adapt this definition to the setting where the samples are frequencies. Let j be the all ones vector, $r = (1, 2, \ldots, d)$ and $r_2 = (1^2, 2^2, \ldots, d^2)$. Then we define the corresponding expressions in Equation 4 as

$$\begin{aligned} n &= jAj^T \\ \sum x &= rAj^T \\ \sum y &= jAr^T \\ \sum x^2 &= r_2Aj^T \\ \sum y^2 &= jAr_2^T \\ \sum xy &= rAr^T \end{aligned}$$

which gives the sample correlation for rows and columns of a matrix as

$$r = \frac{jAj^T\, rAr^T - rAj^T\, jAr^T}{\sqrt{jAj^T\, r_2Aj^T - (rAj^T)^2}\sqrt{jAj^T\, jAr_2^T - (jAr^T)^2}} \tag{5}$$

We note that in Figure 5, this measure of diagonality correlates well with the ratio of offdiagonal elements to diagonal elements given by $\frac{\|\text{off}(T)\|}{\|\text{diag}(T)\|}$ when considering a range of signal to noise. In the following, we discuss this measure with respect to the number of filters $|\Omega|$.

B Filter count affects alignment diagonality

In Figure 5, we showed that for $|\Omega| = 16$ filters, bandlimited correlations outperform direct Fourier correlations robustly across a broad range of SNR. To identify the optimal number of filters for this example, we examined the relationship between diagonality (measured according to the off diagonal ratio and the Pearson Correlation (equation 5)). In this experiment (Figure 3), we performed a similar experiment of identical swiss rolls as D, instead fixing $\text{SNR}_{dB} = 0$ and dyadically varying the number of filters between $|\Omega| = 2$ and $|\Omega| = N = 400$.

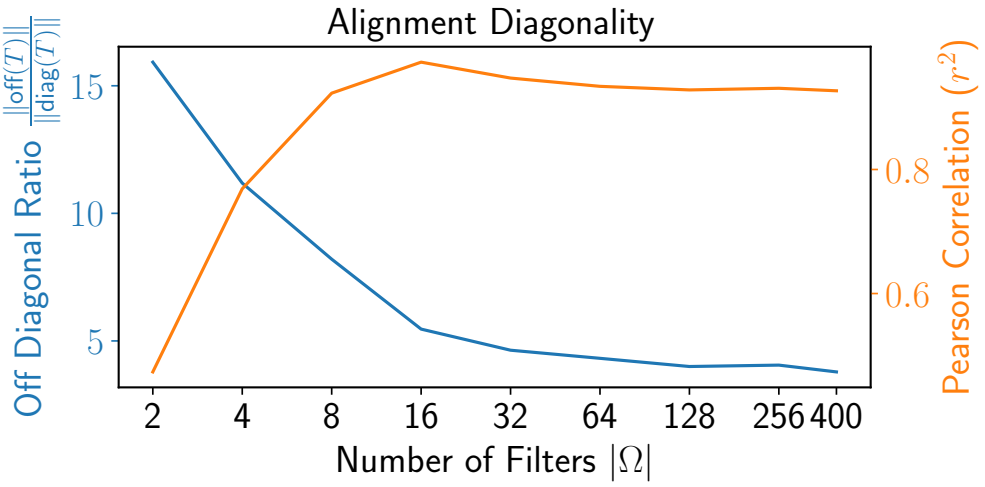


Figure 3: Filter Count and Alignment Diagonality. Two identical swiss roll graphs were generated with $N = 400$ vertices [27]. The ground truth mapping between these manifold is the identity matrix. Subsequently, $S = 4000$ signals were generated by sampling the identity matrix with random gaussian noise added at a signal to noise ratio of $0dB$. Next, a set of $|\Omega|$ itersine wavelets (x-axis) were used to obtain a harmonic alignment matrix and the off diagonal ratio (blue) and Pearson diagonal correlation (equation 5 , orange) was recorded. These metrics are inversely correlated and are different measurements of the diagonality of a matrix. While the off diagonal ratio decays as more filters are added, the Pearson diagonality peaks at $\Omega = 16$.

We note that, as discussed in Section A, the off diagonal ratio is inversely correlated with the Pearson correlation across the range of $|\Omega|$. However, while the off diagonal ratio continues to decay as $|\Omega| \to N$, the Pearson correlation reaches a maxima at $|\Omega| = 16$. This represents a tradeoff between the diagonal strictness of the correlations and the flexibility to find correlations in low SNR settings.

C Bandlimiting prevents spurious correlations caused by highly coherent eigenvectors

In Section 5, we mentioned that localization in the graph eigenbasis can lead to spurious correlations, which we proposed to eliminate by using spectral graph wavelets. To see an example of such a localized eigenbasis, we generated a random sensor graph of 20 nodes using `gspbox` [27]. Letting $\mathcal{L} = \Psi\Lambda\Psi^{-}1$ be the normalized Laplacian of this graph, we identified the vertex and eigenvector pair with the largest coherence for this graph using $\underset{i,j}{\text{argmax}}\langle\delta_i, \psi_j\rangle$ where δ_i is a dirac delta centered at the i-th vertex and ψ_j is the j-th eigenvector of $\mathcal{L}$. In matrix form, this is $\underset{i,j}{\text{argmax}}[\Psi I]_{(i,j)} = \underset{i,j}{\text{argmax}}\Psi_{(i,j)}$. For this graph, $i = 8$ and $j = 14$. We show this harmonic in Figure 4(c), which is very close a delta. Next, we generated a low frequency signal x localized on vertex 8, and added a small impulse to it (Figure 4(a)). This signal's Fourier transform is shown in Figure 4(d)); its two largest coefficients correspond to the second eigenvector ψ_2 (letting the trivial eigenvector be ψ_0; Figure 4(b)) and the

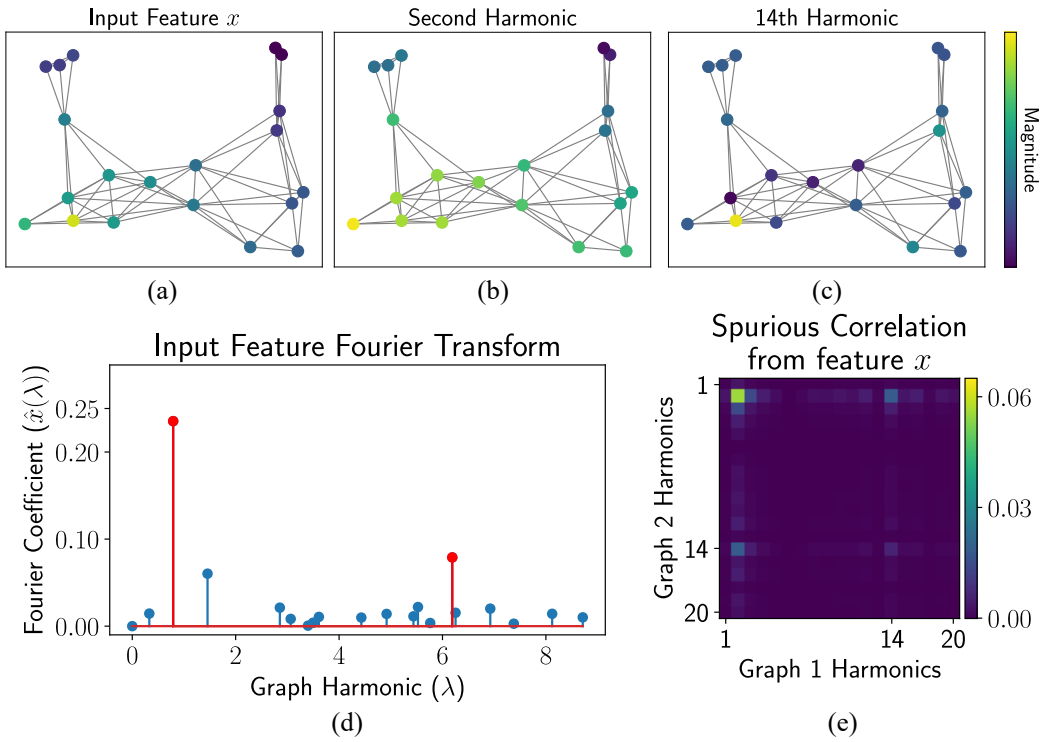


Figure 4: An example of a low-frequency feature incurring artificially high correlation with a highly coherent harmonic over a sensor network G. (a) Magnitude of the low-frequency input feature x; (b) Magnitude of the second harmonic of G; (c) Magnitude of the high-frequency 14th harmonic of G with high coherence around the vertex on which x is centered; (d) Magnitude of the Fourier transform of x over G. The second-largest Fourier coefficient corresponds to the 14th harmonic; (e) Correlation matrix of x with the harmonics of G contains spurious correlations with the 14th harmonic.

14th eigenvector. The contribution of these vectors to x is apparent when one considers the vertex domain representations in Figure 4 (a-c). However, when the Fourier Transform of x is used to generate correlations, it adds off diagonal correlations between the second eigenvector and the 14th eigenvector of $\mathcal{L}$. Such correlations are a deleterious product of coherence, which is a common feature of general graphs and signals [34].

D Demonstration of bandlimited correlations

To empirically measure the effect of partitioning on learned correlations, we demonstrate a simple example in Figure 5. We used a Swiss roll dataset (obtained from [27]) to generate the ground truth graph $\mathcal{G}_1$. Then, we produced a second graph $\mathcal{G}_2$ that is identical to $\mathcal{G}_1$, producing a ground truth mapping that is identity ($\mathbf{I}$). We can then measure how near an alignment is to ground truth by its "diagonality". To measure this, we used (1) the ratio of the norm of off diagonal elements to the norm of diagonal elements of the alignment matrix and (2) the correlation of the columns and the rows of the matrix (see Appendix A for a discussion of this correlation). Next, we sampled 10 copies of the identity matrix with normally distributed noise at signal to noise ratios (SNR) over a covering of $[-200,\ 100]dB$ to produce $S = 4000$ signals at each SNR. At high SNR, these signals are close to identity and transformations will converge to the ground truth mapping for both bandlimited and non-limited alignments. At low SNR, the signals that result are incompatible and alignment for small S is infeasible. Despite this, we observe that alignment with $|\Omega| = 16$ bandlimited filters (shown spectrally in Figure 5(a)) is robust to extreme noise regimes. In Appendix B we discuss the selection of $|\Omega| = 16$, which optimizes the Pearson correlation in this example. In Figure 5(c), we show the ground truth (identity), non-limited, and bandlimited alignments obtained for one realization at

$\text{SNR}_{dB} = -200$. This illustrates the patterns obtained by bandlimited alignment, which imposes mapping between compact ranges of the spectrum.

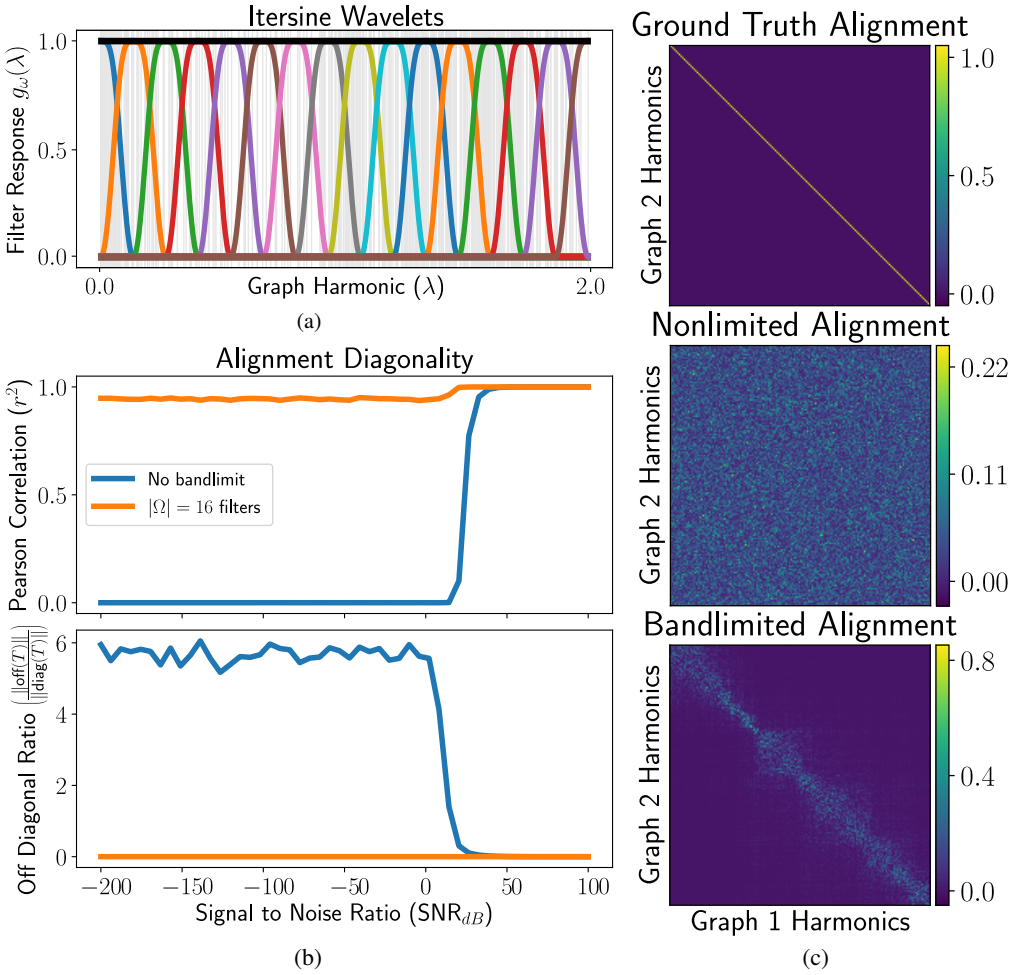


Figure 5: Bandlimited correlations restricted by itersine wavelets more accurately recover signals close to the diagonal. (a) Itersine wavelets are defined in the spectral domain as a translated kernel of the Laplacian eigenvalues (see Section 5 and Equation 3 for kernel). Each scale $\omega \in \Omega$ is represented as a different color in this figure. The frame bounds are shown in black; itersine wavelets form a tight frame. (b) Alignment diagonality is robust to noise when alignment is performed with $|\Omega| = 16$ itersine wavelets. Top: Pearson diagonality correlation, see Appendix Section A. Bottom: Ratio of norm of off diagonal elements to the norm of diagonal elements. Blue: Non-limited correlations. Orange: Bandlimited correlations. See Section D for experimental discussion. (c) *Top:* Ground truth mapping for a test case of aligning a swiss roll graph to itself. The mapping is identity. *Middle:* Non-limited correlation matrix of graph harmonics gives many spurious correlations and poorly approximates the ground truth. *Bottom:* Bandlimited correlation matrix more closely approximates the ground truth by limiting entries far off the diagonal.

E Transfer learning with alternative classifiers

Transfer learning classification was performed in `sklearn` [35] using default parameters for the 5-nearest neighbours, linear SVM and naive Bayes classifiers. Figure 6 shows results for all three classifiers. Although the classifiers' individual dependence on data imbalance varies, in all three cases harmonic alignment outperforms MAGAN and MNN. Specifically, k-nearest neighbour classification improves as the test set gets larger due to harmonic alignment's increased robustness with more data

(see Figure 1(c)), and approaches the performance of linear SVM as the test set increases in size. On the other hand, naive Bayes increasingly overfits to the training set as the test set increases in size. Further, naive Bayes is highly sensitive to changes in data density, which causes near-random performance with MNN and MAGAN, neither of which preserves the intrinsic density of the data. In contrast to this, harmonic alignment strictly preserves data density and hence achieves significantly improved performance in comparison.

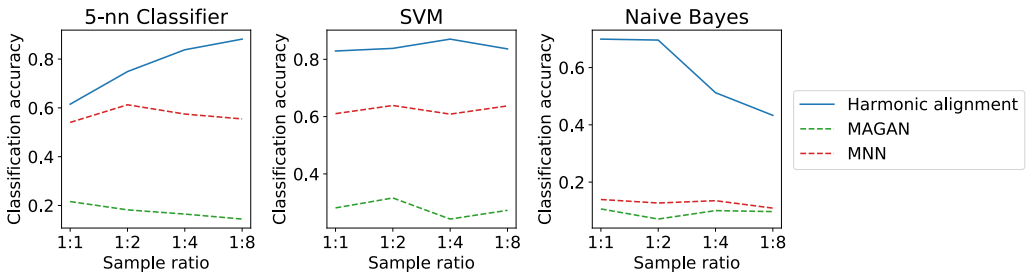
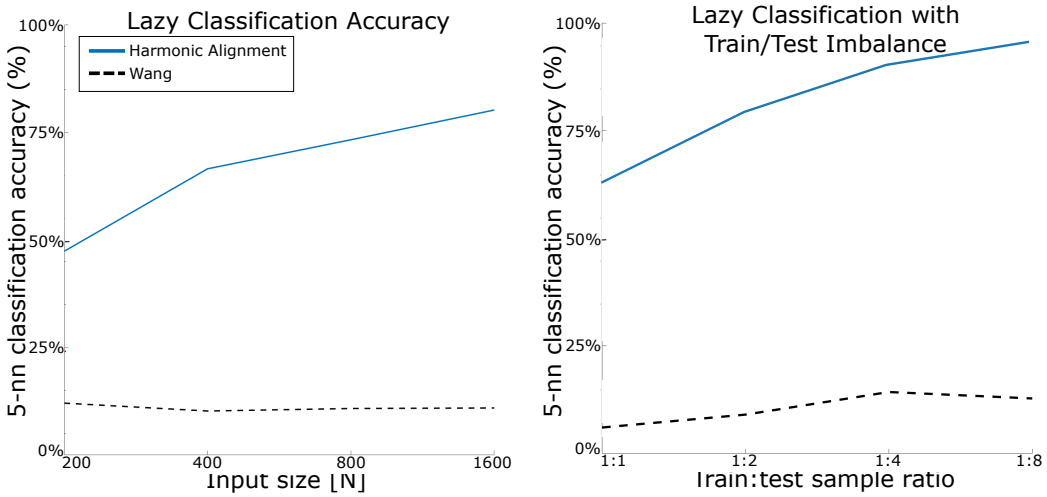


Figure 6: Transfer learning performance with various classifiers. For each ratio, 1K uncorrupted, labeled digits were sampled from MNIST, and then 1K, 2K, 4K, and 8K (x-axis) unlabeled points were sampled and corrupted with 35% column identity.

F Comparison to Wang and Mahadevan

Despite being a natural candidate for comparison to our method, unfortunately no standard implementation of the method proposed by Wang and Mahadevan [12] is available. Our implementation of their method performed extremely poorly (worse than random) on the comparisons and is extremely computationally intensive. The method is therefore not shown in the main comparisons; however, for completeness, the results are shown in Figure 7.



(a) (b)

Figure 7: Recovery of k-neighborhoods under feature corruption. Mean over 3 iterations is reported for each method. (a) Lazy classification accuracy relative to input size with unlabeled randomly corrupted digits with 35% preserved pixels. (b) Transfer learning performance. For each ratio, 1K uncorrupted, labeled digits were sampled from MNIST, and then 1K, 2K, 4K, and 8K (x-axis) unlabeled points were sampled and corrupted with 35% column identity.