SE AND S_NL DIAGRAMS: FLEXIBLE DATA STRUCTURES FOR MIR

Melissa R. McGuirl [1] **Katherine M. Kinnaird**[1] **Claire Savard**[2]
Erin H. Bugbee[3]
[1] Division of Applied Mathematics, Brown University, USA
[2] Department of Mathematics, University of Michigan, USA
[3] Department of Biostatistics, Brown University, USA
melissa_mcguirl@brown.edu

ABSTRACT

The matrix-based representations commonly used in MIR tasks are often difficult to interpret. This work introduces start-end (SE) diagrams and start(normalized)-length (S_NL) diagrams, two novel structure-based representations for sequential music data. Inspired by methods from topological data analysis, both SE and S_NL diagrams come equipped with efficiently computable and stable metrics. Utilizing SE or S_NL diagrams as input, we address the cover song task for score-based data with high accuracy. While both representations are concisely defined and flexible, S_NL diagrams in particular address issues introduced by commonly used resampling methods.

1. INTRODUCTION

Since Foote's introduction of the self-similarity matrix (SSM) in [8], matrix-based representations for music-based data streams have been commonly used in MIR literature. Both SSMs and self-dissimilarity matrices (SDMs) have been used as the starting point for a variety of tasks including the cover song task [2, 10, 13, 21], the chorus detection task [9], and segmentation task [14, 18, 19].

While straightforward to compute, these matrix-based representations are challenging to interpret, requiring extensive post-processing, such as smoothing and resampling techniques used in [10] or path enhancement applied in [15–17]. These post-processing steps can also introduce uncertainty or reduce some of the intuitive explanations for the resulting visualizations. The aligned hierarchies from [13] is an intuitive structure-based representation that is also the result of post-processing SDMs. However, this representation is rigid as it requires two songs to be the exactly the same length for comparisons. The aligned sub-hierarchies attempt to address this rigidity, but many songs do not have enough structure to have this collection of structure-based representations for sections of a song [12].

In this paper, we contribute two new structure-based visualizations for music-based data streams: *Start-end diagrams* (in Section 3) and *Start(normalized)-length diagrams* (in Section 4). With roots in topological data analysis, the presented methods are flexible, computationally efficient, and easily adaptable. Moreover, we present experiments applying these methods to a version of the cover song task. We discuss contributions of our novel methods (in Section 6) and share future directions (in Section 7).

2. MOTIVATION AND BACKGROUND

This work builds upon aligned hierarchies developed in [13]. The aligned hierarchies for a song encodes all possible hierarchical structure decompositions of that song on one common time axis. The aligned hierarchies representation is defined as a collection of three components: a binary onset matrix B_H, a length vector, and an annotation vector that acts as a key for B_H [13]. Each row of B_H corresponds to one kind of repetition, with entries equal to one denoting where instances of a repeat begins.

Aligned hierarchies have been used to compare songs under the fingerprint task by leveraging that this representation can be embedded into a classification space with a natural notion of distance. This distance computes the number of dissimilarities between start-times for repeats of each size and then totals those dissimilarities across all sizes. Using the aligned hierarchies as the basis of comparison yields precise results, yet the metric is both rigid with respect to the length of the songs and computationally expensive as it is based on a binary classification [13].

In this work, we produce novel methods of representing and comparing songs. Inspired by work in topological data analysis, our methods extend the aligned hierarchies while addressing their limitations. Moreover, we offer several variations of our method, which make our representations flexible and easily adaptable to many applications such as cover song and remix detection.

3. START-END DIAGRAMS

Aligned hierarchies represents repeated structures of music data. Similarly, topological data analysis (TDA), an emerging field of mathematics, aims to extract structural, or topological, information from complex data. The start-

end diagram is a transformation of the aligned hierarchies that is reminiscent of persistence diagrams from TDA.

In TDA, data are thresholded via a sequence of parameter values. Topological summaries, such as the number of loops in the data, are then computed for each parameter value in the sequence [4, 5, 7]. A common way to represent this topological information is with persistence diagrams. Briefly, a *persistence diagram* is a collection of points $\{(b_i, d_i)\}_{i=1}^{N} \subset \mathbb{R}^2_+$, such that (b_i, d_i) corresponds to a topological structure that appears at some parameter value b_i and disappears at parameter value $d_i \geq b_i$ [4,5,7].

Inspired by TDA, we transform aligned hierarchies into a *start-end (SE) diagram*. The *SE diagram* corresponding to aligned hierarchies with N repeated structures is defined as a collection of points $\{(s_i, e_i)\}_{i=1}^{N} \subset \mathbb{R}^2_+$, where s_i and e_i are the start and end times, respectively, of the i^{th} repeated structure. Under this transformation, we adjust the time scale such that time zero refers to the start of the first block of the aligned hierarchies and truncate the song to end where the last block of the aligned hierarchies ends.

SE diagrams are not inherently topological (in a mathematical sense), rather we are adapting data structures from TDA. While SE diagrams cannot delineate two different types of repeats of the same length, there are several advantages of using SE diagrams over aligned hierarchies. First, they are a more concisely defined structure, as each diagram is simply a finite collection of points. Second, leveraging theoretical results from TDA, there are easily adaptable metrics on the space of SE diagrams (Subsection 3.1). Third, these metrics are more flexible than those for the aligned hierarchies while maintaining accuracy and precision in the cover song task (Subsection 3.2).

3.1 Metrics for SE diagrams

In TDA there are two common metrics for persistence diagrams that can be extended to SE diagrams: *the bottleneck metric* and *the Wasserstein metric*. Both metrics measure the error of an optimal alignment of points in two persistence (or SE) diagrams. The metrics are stable, meaning small differences between aligned hierarchies will yield a small SE diagram distance [6, 7, 11]. Moreover, as shown in [11], these distances can be computed efficiently using k-dimensional trees. Thus, under either the Wasserstein or bottleneck notions, SE diagrams are equipped with stable and computable metrics which facilitate their ability to address the cover song task efficiently and accurately.

3.1.1 Intuitive definitions

Intuitively, the Wasserstein and bottleneck metrics attempt to find the best alignment of points between two SE diagrams and then measure cost of the alignment using an l^p metric. When aligning two diagrams for comparison, each diagram point must have a corresponding aligned point in the other diagram and no points can be aligned with more than one point. The aligned points thus form a pair.

Recall, the l^p norm for any point $\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is given by $||\vec{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ for $1 \leq p < \infty$, and the l^∞ norm is $||x||_\infty = \max_i(|x_i|)$. Norms naturally give rise to metrics. Specifically, the l^p metric $d_p : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ between any two points $\vec{x}, \vec{y} \in \mathbb{R}^n$ is defined for $1 \leq p \leq \infty$ as $d_p(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_p$. Note that in $\mathbb{R}^2$, d_2 is the the straight line distance between two points in the plane, while d_∞ is the maximum of the horizontal and vertical distance between two points in the plane.

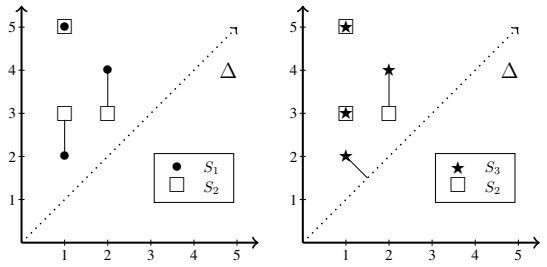

Figure 1. Optimal alignment of S_1 and S_2 without aligning with Δ (left), and optimal alignment of S_2 and S_3 while allowing for alignments with Δ (right). Note that $\Delta \sim (1, 2) \in S_3$ and the l^2 distances between the pairs are $\left\{\sqrt{\frac{1}{2}}, 0, 1, 0\right\}$, so $d_W^{2,1}(S_2, S_3) = \frac{1+\sqrt{2}}{\sqrt{2}}$, $d_B^2(S_2, S_3) = 1$.

For example, consider two SE diagrams $S_1 = \{(1, 2), (1, 5), (2, 4)\}$ and $S_2 = \{(1, 3), (1, 5), (2, 3)\}$. To find the optimal alignment we pair points in diagram S_1 with points in diagram S_2 in a way that minimizes the total distance between all pairs. The best alignment of S_1 and S_2 is given by $\{(1, 2) \sim (1, 3), (1, 5) \sim (1, 5), (2, 4) \sim (2, 3)\}$ (see Figure 1). The corresponding l^∞ distances of these pairs are $\{1, 0, 1\}$. The (∞,q)-Wasserstein and ∞-bottleneck metrics are then defined as the l^q and l^∞, respectively, norms of the l^∞ distances of the pairs in the optimal alignment.

In this example, the (∞,2)-Wasserstein distance between S_1 and S_2 is $d_W^{\infty,2}(S_1, S_2) = \sqrt{2}$, whereas the ∞-bottleneck distance between S_1 and S_2 is $d_B^\infty(S_1, S_2) = 1$. Note, the first superscript in $d_W^{\infty,2}$ corresponds to taking the l^∞ distances between the points in each pair (inner norm), and the second superscript denotes taking l^2 norm of those l^∞ distances (outer norm).

Thus far we have defined distances between two SE diagrams with the same number of points. By definition, computing the Wasserstein or bottleneck distance between two SE diagrams requires both diagrams to have the same number of points. In practice, however, we want to compare SE diagrams with any number of points, as songs have varying amounts of repeated structures.

To compare SE diagrams of differing numbers of points we find the optimal alignment of points in two SE diagrams, while also allowing diagram points to match to repeated structures existing for no time, meaning their start and end times are the same. Formally, we allow points to align with the *diagonal*, defined as $\Delta = \{(s, e) : s = e, s \geq 0\}$ (see Figure 1) [6, 7, 11].

The motivation for allowing points to align with repeated structures that exist for no time is two-fold. First, unlike arbitrary insertions or deletions of points in either SE diagrams, aligning points with Δ will give rise to a

metric that respects the triangle inequality. With the triangle inequality, it is impossible to have the case where songs B and C are both cover songs of song A so that $d(A, B)$ and $d(A, C)$ are small, but $d(B, C)$ is large.

Second, pairing unmatched points with Δ enforces that two songs will be considered dissimilar when their long-lasting repeated structures do not have a corresponding pair under the optimal alignment of points. To see this, observe that when a point $(s^1_*, e^1_*) \in S_1$ does not have a corresponding match in S_2 then $(s^1_*, e^1_*) \sim \Delta$ and this pairing contributes $d_p((s^1_*, e^1_*), \Delta) = 2^{\frac{1}{p}-1}|e^1_* - s^1_*|$ to the overall cost of the alignment. Thus, the cost for unmatched points aligning with Δ increases as the length ($|e^1_* - s^1_*|$) of the unmatched repeated structure increases.

In short, the (p,q)-Wasserstein and p-Bottleneck metrics measure the distances between pairs of points in the optimal alignment of two SE diagrams and Δ. When $q = 2$, the Wasserstein distance is the Euclidean norm of the distances between pairs in the optimal alignment. In contrast, the Bottleneck distance is to the maximum distance between pairs in the optimal alignment. In the following subsection we provide rigorous definitions of these metrics.

3.1.2 Rigorous Definitions

Let $S_1 = \{(s^1_i, e^1_i)\}_{i\in I}$ and $S_2 = \{(s^2_j, e^2_j)\}_{j\in J}$ be SE diagrams, and let ϕ be a bijection between subsets $\tilde{I} \subset I$ and $\phi(\tilde{I}) \subset J$. The *p-q penalty* of ϕ is defined as:

$$P^p_q(\phi) = \sum_{i\in\tilde{I}} d_p((s^1_i, e^1_i), (s^2_{\phi(i)}, e^2_{\phi(i)}))^q + \sum_{i\in I\setminus\tilde{I}} d_p((s^1_i, e^1_i), \Delta)^q + \sum_{j\in J\setminus\phi(\tilde{I})} d_p((s^2_j, e^2_j), \Delta)^q,$$

for $1 \leq q < \infty$, and the ∞-*q penalty* of ϕ is defined as:

$$P^p_\infty(\phi) = \max\left\{ \max_{i\in\tilde{I}} d_p((s^1_i, e^1_i), (s^2_{\phi(i)}, e^2_{\phi(i)})), \max_{i\in I\setminus\tilde{I}} d_p((s^1_i, e^1_i), \Delta), \max_{j\in J\setminus\phi(\tilde{I})} d_p((s^2_j, e^2_j), \Delta) \right\}.$$

These penalties define a cost function for aligning points in S_1 with points in S_2 (encoded in the first terms), and for aligning all unmatched points with Δ (encoded in the remaining terms). The *p-bottleneck distance* is then defined as $d^p_B(S_1, S_2) = \min_\phi P^p_\infty(\phi)$ and the *(p,q)-Wasserstein distance* is $d^{p,q}_W(S_1, S_2) = \min_\phi P^p_q(\phi))^{\frac{1}{q}}$ [6, 7, 11].

3.2 Applications of SE diagrams

Utilizing the metrics described in the previous section, there are several ways of comparing songs for the cover song task. This work explores the efficacy of using the pairwise p-bottleneck and (p,q)-Wasserstein distances as input for a mutual nearest neighbor search.

Noting that the presented methods take aligned hierarchies as input, we pre-process music-based data in three steps: 1) build audio shingles from the concatenated beat-synchronous chroma features, 2) compute the SDM, and finally 3) construct the aligned hierarchies for each song's SDM (see [13] for more details). After the aligned hierarchies are created, the procedure is as follows: [1]:

1. Transform each aligned hierarchies into the corresponding SE diagram as described in Section 3
2. Compute bottleneck or Wasserstein distances between pairs of SE diagrams using `Hera` [11] [2]
3. Mark a pair of songs as cover songs of each other if the songs are mutual nearest neighbors

See Figure 2 for a visual example of our method. To test our method we apply it to 52 Mazurka scores by

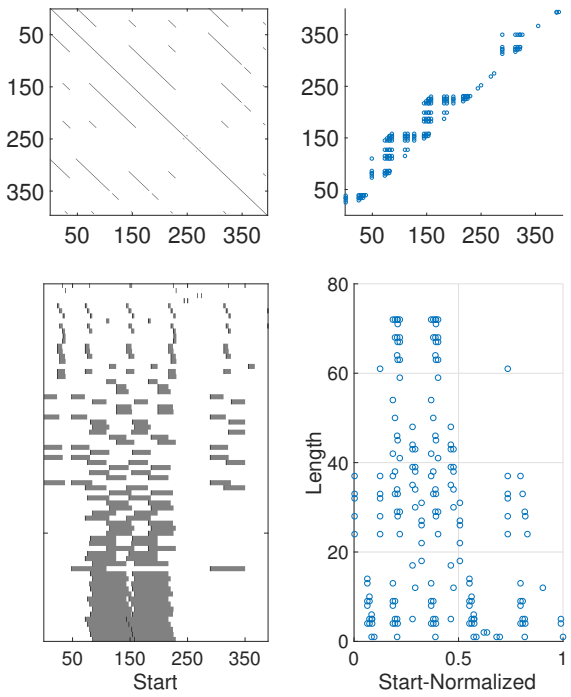

Figure 2. The thresholded SDM (top left), aligned hierarchies (bottom left), SE diagram (top right), and S_NL diagram with $\alpha = 1$ (bottom right) corresponding to Mazurka 52 expanded with threshold=0.02, shingle=12. Each dark block diagonal in the thresholded SDM represents two sections that are repeats of each other. Repetitions of all sizes are encoded in the aligned hierarchies as blocks, separated into rows. Each block in the aligned hierarchies is represented as a point in both the SE and S_NL diagrams. The smallest repeats are close to the diagonal in the SE diagram and are near the horizontal axis in the S_NL diagram. The tops of the peaks in the SE and S_NL diagrams represent the longest repetitions, which are the blocks at the bottom of the aligned hierarchies.

[1] Code and processed data are publicly available here: `https://github.com/MelissaMcguirl/SE_SNL_analysis.git`.

[2] `Hera` is publicly available here: `https://bitbucket.org/grey_narn/hera`.

	Threshold	0.01		0.02		0.03		0.04		0.05	
	Shingle	6	12	6	12	6	12	6	12	6	12
d_B^∞ Metric	Precision	**0.871**	0.622	0.848	0.718	**0.871**	0.800	0.818	0.725	0.824	0.757
	Recall	0.519	0.538	0.538	0.538	0.519	0.538	0.519	**0.558**	0.538	0.538
$d_W^{2,2}$ Metric	Precision	**0.966**	0.675	0.879	0.903	0.933	0.824	0.909	0.875	0.875	0.848
	Recall	0.538	0.519	0.558	0.538	0.538	0.538	**0.576**	0.538	0.538	0.538
$d_W^{\infty,2}$ Metric	Precision	**1.0**	0.683	0.909	0.909	0.933	0.882	0.906	0.906	0.879	0.853
	Recall	0.558	0.538	**0.577**	**0.577**	0.538	**0.577**	0.558	0.558	0.558	0.558

Table 1. Precision and recall values for the mutual nearest neighbor matching of SE diagrams for 104 Mazurka scores.

Chopin downloaded in **kern format from KernScore online database [3] (see [20]). Each score produces two data elements, an expanded version which includes all repeated sections as marked in the score, and a non-expanded version which has each repeated section played once. In the cover song task, the goal is to match the expanded and non-expanded versions of each song.

We construct SE diagrams for all 104 songs in our dataset and compute their pairwise distances for three metrics: d_B^∞, $d_W^{2,2}$, and $d_W^{\infty,2}$. We perform 10 experiment trials per metric, varying the number of chroma vectors per audio shingle and varying the threshold applied to the SDM. The precision and recall values are presented in Table 1.

These results show that SE diagrams can accomplish this challenging version of the cover song task with high precision and moderate recall values regardless of the metric. It is crucial and exciting to note that the SE diagrams achieved these results without any resampling to the diagrams. Moreover, as we will see in the next section, this method is easily adaptable to several useful variations.

4. START(NORMALIZED)-LENGTH DIAGRAMS

Since the length of repeats (e-s) is represented diagonally on SE diagrams, these representations can be difficult to interpret. In this section we describe a transformation of SE diagrams called *start-length (SL) diagrams*, along with normalizations. SL diagrams are more intuitive to read than their predecessor and yield stronger experimental results. While reminiscent of the constellation maps from [22], SL diagrams encode structural repeats instead of audio spectrogram peaks.

4.1 Start-Length Diagrams

Consider a SE diagram $S = \{(s_i, e_i)\}_{i=1}^N$. The associated *SL diagram* is $S' = \{(s_i, e_i - s_i)\}_{i=1}^N$, where the x-coordinate corresponds to the start time of a repeated structure and the y-coordinate denotes the length of that repeat. While SE and SL diagrams encode the same information, SL diagrams emphasize the lengths of repeats. This transformation has also been applied to persistence diagrams for TDA applications [1].

[3] `http://kern.humdrum.org/search?s=t\&keyword=Chopin`

4.2 Start(Normalized)-Length Diagrams

In most cases, normalizing SL diagrams before comparison proves to be more effective. We note that similar normalizations can also be applied to SE diagrams. The start(normalized)-length (S_NL) diagrams are defined as $S'_N = \{(\alpha(s_i/M), e_i - s_i)\}_{i=1}^N$, where α is a positive scaling factor and M is a normalization factor. Throughout this paper we will use $M = \max_i s_i$, but other normalizations may be applied. The vertical coordinate of S_NL diagrams are not normalized in order to maintain the emphasis on the lengths of the repeated structures.

Similar to SL diagrams, S_NL diagrams encode the lengths of repeated structures, except the start times in the normalized diagrams are proportional to the length of the song. This normalization acts as a kind of resampling by condensing start times of each song to be on the same scale, while also preserving the lengths of the found repetition patterns. The α parameter is a hyper-parameter that imbues a maximum tolerance on our comparisons. The impact of this parameter is left to Section 4.4.

4.3 Metrics for S_NL diagrams

As with SE diagrams, the bottleneck and Wasserstein metrics can be used to compare S_NL diagrams with one modification. Define $\tilde{\Delta}$ to be the set $\{(s,l) : s = 0, (s,l) \in \mathbb{R}^2_+\}$. The set $\tilde{\Delta}$ is the y-axis of SL and S_NL diagrams and it encodes repeats that start at zero. The S_NL p-bottleneck and S_NL (p,q)-Wasserstein metrics, $\tilde{d}_B^p$ and $\tilde{d}_W^{p,q}$, are then the same as the p-bottleneck metric and (p,q)-Wasserstein metric defined in Section 3.1, except Δ is replaced by $\tilde{\Delta}$ in the definitions of P_∞^p and P_q^p, respectively. We use a modified version of `Hera` to implement these metrics [11].

Pairing unmatched points with repeated structures that start at zero rather than repeated structures that exist for no time allows the user to control the penalty for having unmatched points while maintaining the emphasis on the length of the repeated structures. To better understand this, consider a point $(s_*^1, e_*^1 - s_*^1) \in S_1$ that does not have a corresponding pair in S_2 under the optimal alignment of points in S_1 and S_2. In this case we pair $(s_*^1, e_*^1 - s_*^1)$ with $\tilde{\Delta}$ so that $d_p((s_*^1, e_*^1 - s_*^1), \tilde{\Delta}) = |s_*^1|$. The penalty for unmatched points aligning with $\tilde{\Delta}$ consequently increases as the start time of the corresponding repeated structure increases. It is critical to note, however, that $s \in [0, \alpha]$

	Threshold	0.01		0.02		0.03		0.04		0.05	
	Shingle	6	12	6	12	6	12	6	12	6	12
$\bar{d}_B^\infty$ Metric	Precision	**1.0**	0.827	**1.0**	0.818	**1.0**	0.978	**1.0**	0.935	0.975	0.936
	Recall	0.788	0.827	0.769	**0.865**	0.788	**0.865**	0.750	0.827	0.750	0.846
$\bar{d}_W^{2,2}$, $\bar{d}_W^{\infty,2}$ Metric	Precision	**1.0**	0.833	0.974	0.975	**1.0**	0.976	0.976	**1.0**	0.976	**1.0**
	Recall	0.731	0.769	0.731	0.750	0.731	**0.788**	**0.788**	0.769	**0.788**	**0.788**

Table 2. Precision and recall values for mutual nearest neighbor matching using the distance between S_NL diagrams with $\alpha = 1$ corresponding to 104 Mazurkas. Note, we observe $\bar{d}_W^{\infty,2}$ and $\bar{d}_W^{2,2}$ to be equivalent when the optimal alignment only requires shifts in the start component so that $(|s_2 - s_1|^p + |e_2 - e_1|^p)^{\frac{1}{p}} = |s_2 - s_1| = \max(|s_2 - s_1|, |e_2 - e_1|)$.

for all start times s under the S_NL normalization. Thus, the penalty for having an unmatched diagram point is bounded above by α for S_NL diagrams with $\tilde{d}_B^p$ or $\tilde{d}_W^{p,q}$. We further explain the choice of α in the following section.

4.4 Choosing α

The hyper-parameter α is a positive scaling factor in the normalization. For small α, the cost of aligning points with $\tilde{\Delta}$ is low. For example, if $\alpha = 0.5$, then the cost of aligning a S_NL diagram point with $\tilde{\Delta}$ is at most 0.5. Consequently, when comparing two S_NL diagrams, points within each diagram will be paired only when the difference between their lengths is negligible. Otherwise, it will be more effective to pair both points with $\tilde{\Delta}$. Thus, a small value for α induces a strict matching criterion, where repeated structures are mostly shifted in the start coordinate to pair with a repeated structure of similar or equal length, and structures with unmatched lengths get paired to $\tilde{\Delta}$.

The penalty for matching points with $\tilde{\Delta}$ increases as α increases. For a large value of α, two S_NL diagram points of slightly different lengths are more likely to be matched with each other than with $\tilde{\Delta}$. Consequently, a larger α value yields a more flexible length-based matching system.

The choice of α depends on the importance of the length of the repeated structures. One might consider $0 < \alpha \leq 1$ if repeated structures of different lengths are considered significantly dissimilar, or $\alpha \gg 1$ to allow for flexibility in length-based matchings. The inclusion of this parameter further adds to the flexibility of S_NL diagrams.

4.5 Applications of S_NL diagrams

We apply the same algorithm to the same dataset defined in Section 3.2 with SL and S_NL diagrams using the adapted bottleneck and Wasserstein metrics for the cover song task. Again, 10 experiment runs are performed per metric, varying the number of beats per audio shingle and the threshold applied to the SDM. For SL diagrams, the mean precision values across the 10 experiments are 0.791, 0.803, and 0.791 with $\bar{d}_B^\infty$, $\bar{d}_W^{2,2}$, and $\bar{d}_W^{\infty,2}$, respectively. The corresponding mean recall values are 0.596, 0.581, and 0.577.

Experiments on S_NL diagrams yield a significant increase in accuracy over both SL and SE diagrams on the cover song task, suggesting that a strict matching criterion in the length coordinate and flexibility in the start coordinate is the most accurate way to approach the cover song task with these diagram representations. Across the 10 experiments, the ∞-Bottleneck metric yields mean precision and recall values of 0.947 and 0.808, respectively. The (2,2) and (∞,2)-Wasserstein metrics yield a mean precision value of 0.971 and a mean recall value 0.763.

The experimental results for S_NL diagrams with $\alpha = 1$ are presented in Table 2. Separate analyses show that precision and recall remain constant for $0 < \alpha \leq 1$, and decrease monotonically as α increases above 1 for this data.

S_NL diagrams are not restricted to the standard start-normalization presented here. Applying the same method under the (2,2)-Wasserstein metric with a Chebyshev-normalized start component yields comparable results with slightly lower precision values and higher recall values. This further demonstrates the robustness of S_NL diagrams.

To push the limits of S_NL diagrams, we applied this representation to audio-based data without making any modifications in the preprocessing steps. We extracted beat-synchronous chroma feature vectors using `librosa` toolbox [4] for a collection of performances of Mazurka [5] and constructed the corresponding S_NL diagrams. Following the evaluation method in [2], we ranked the songs based on their pairwise-distances for the cover song task. While the mean average precision values were less than 0.1 across a range of metrics and α values, these results demonstrate that audio-specific preprocessing must be done in order to mitigate noise and other artifacts on tracks. Since there exist theoretical guarantees of stability of the S_NL diagrams, we are confident that with the appropriate preprocessing methods S_NL diagrams are suitable for a range of both score-based and audio-based music.

5. COMPARISON TO PREVIOUS WORK

Previous experiments on this Mazurka score dataset were done with the aligned hierarchies [13] and with the aligned sub-hierarchies (AsH) [12]. The metric for aligned hierarchies only allowed for pairwise comparisons between songs that were of the same length, meaning that it could only be used for the fingerprint task [13].

Initial experiments using AsH to address the cover song task were completed in [12]. Following the same exper-

[4] `https://librosa.github.io/librosa/`

[5] The CHARM Project Discography website maintains a list of commercially available Mazurka recordings at `http://www.mazurka.org.uk/info/discography/?`

		Mean	Median	Min.	Max.
SE	Precision	0.847	0.873	0.622	1.0
	Recall	0.545	0.538	0.519	0.577
S_NL	Precision	**0.963**	0.976	0.818	**1.0**
	Recall	**0.778**	**0.779**	**0.731**	**0.865**
AsH	Precision	0.9511	**0.9808**	**0.840**	**1.0**
	Recall	0.771	0.754	0.692	0.882

Table 3. Summary statistics of the precision and recall values for mutual nearest neighbor matching of the Mazurka score data using SE diagrams, S_NL diagrams, and aligned sub-hierarchies (AsH). The AsH statistics are computed over the 10 combinations of thresholds and shingles, whereas the SE and S_NL statistics are computed over the 30 combinations of thresholds, shingles, and metrics. The S_NL experiments apply to a more complete dataset than the AsH and still attain similar high precision-recall values.

imental design varying the thresholds between 0.01 and 0.05 and testing shingle widths of 6 and 12, these experiments produced high precision rates (between 0.840 and 1.0) and modest recall rates (between 0.692 and 0.882).

Table 3 presents the summary statistics of the precision-recall values for mutual nearest neighbor matching using SE diagrams, S_NL diagrams, and AsH on the Mazurka score data. The AsH results are comparable to the precision-recall values of the S_NL experiments, while the SE experiments yield slightly lower precision-recall rates than both AsH and S_NL diagrams. However, the AsH post-processing technique requires repetitions to have enough repeated structure within them to build a smaller aligned hierarchies for these song sections. Consequently, as many as 68 songs (depending on the shingle size and threshold) do not have an AsH representation.

An advantage of the presented methods is that if a song has an aligned hierarchies representation then it has a SE diagram and S_NL diagram. Thus, the S_NL experiments work with a more complete dataset than the AsH experiments and still attain similar high precision-recall values.

6. DISCUSSION

Both the SE diagrams and the S_NL diagrams offer exciting contributions to the representation of music-based sequential data streams. These diagrams offer a clear representation of the relationships between repeated structural elements and have advantages over previous structure-based methods. By allowing for two recordings of different lengths to be directly compared without altering the beat-synchronized lengths of structural repeats, the S_NL diagram addresses an issue created by current resampling methods for music-based data streams.

The S_NL diagram provides a new method for resampling music-based data streams. The goal of resampling is to ensure that all matrix representations are the same size, which eases comparisons between music-based data streams. Current resampling methods compress all musical structures to represent a proportion of the length of the song, which results in comparing sections of a song that are proportionally the same length but not actually the same number of beats. In [10], the proportional comparisons of structural elements had issues comparing versions of Mazurka Op. 68, No. 4, which could be mitigated using the kind of resampling offered by S_NL diagrams.

Structure-based comparisons on resampled representations of a piece of music are then between proportions of the song instead of the true lengths of the repeats. This is especially an issue in cases where one artist plays a song once through, while another plays the piece through twice in its entirety. In such an example, a section of 100 beats long in the piece will look twice as long in the first representation when compared to the second representation.

The S_NL diagrams balance resampling all representations to be of the same length with maintaining the lengths of repeated structures. To accomplish this, the S_NL diagrams resample only the starting position of each repeated structure, while leaving the lengths alone. What is exciting about this innovative approach is that it not only allows for uniform comparisons – as desired by traditional resampling – but it also allows for comparisons between sections of the same length of time (or beats) instead of sections of the same proportional length of the song.

7. CONCLUSION

In this paper we present SE and S_NL diagrams, two novel, concisely defined, and flexible representations for music-based data. Leveraging theory from TDA, these diagrams come equipped with stable metrics which allow us to apply a mutual nearest neighbor search for the cover song task.

Experimental results demonstrate that SE and S_NL diagrams address the cover song task with high accuracy for score-based data, and these results are robust with respect to the choice of metric. Moreover, SE diagrams avoid resampling all together, while S_NL diagrams resample only the starting positions of repeated structures.

Overall, S_NL diagrams yield the highest accuracy in addressing the cover song task and they are more flexible. In addition to the choice of normalization, S_NL diagrams include the hyper-parameter α which allows the user to directly control the rigidity of the length-matching criterion.

In future work we plan to apply SE and S_NL diagrams to preprocessed audio data, and to extend these diagram representations so that they are suitable for machine learning tasks. Theoretical guarantees provide strong evidence that SE and S_NL diagrams will be applicable to both score and audio data after appropriate preprocessing. Beyond the method presented here, SE and S_NL diagrams can be mapped into spaces that are more suitable for machine learning tasks, just as, for example, persistence diagrams have been transformed to sequences of piecewise-linear functions and vectors in Euclidean space [1, 3]. Thus, SE and S_NL diagrams open up a range of new opportunities for applying machine learning methods through the lens of TDA to music-based tasks.

8. ACKNOWLEDGEMENTS

The first author is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1644760. The third and fourth authors are supported by the National Science Foundation under Grant No. DMS-1439786 and The Karen T. Romer Undergraduate Teaching and Research Awards while the authors were in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the Summer@ICERM 2017 program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Arnur Nigmetov for his generous help with adapting the Hera software for the $S_N L$ diagrams.

9. REFERENCES

[1] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[2] J. Bello. Measuring structural similarity in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2013–2025, 2011.

[3] P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015.

[4] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[5] G. Carlsson, A. Zomorodian, A. Collins, and L. J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(02):149–187, 2005.

[6] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Memoli, and S.Y. Oudot. Gromov-Hausdorff Stable Signatures for Shapes using Persistence. *Computer Graphics Forum*, 2009.

[7] H. Edelsbrunner and J. L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[8] J. Foote. Visualizing music and audio using self-similarity. *Proc. ACM Multimedia 99*, pages 77–80, 1999.

[9] M. Goto. A chorus-section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1783–1794, 2006.

[10] P. Grosche, J. Serrà, M. Müller, and J.Ll. Arcos. Structure-based audio fingerprinting for music retrieval. *Proc. of* 13^{th} *ISMIR Conference*, pages 55–60, 2012.

[11] M. Kerber, D. Morozov, and A. Nigmetov. *Geometry Helps to Compare Persistence Diagrams.* 2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX).

[12] K. M. Kinnaird. *Aligned Hierarchies for Sequential Data*. PhD thesis, Dartmouth College, 2014.

[13] K. M. Kinnaird. Aligned hierarchies: A multi-scale structure-based representation for music-based data streams. *Proc. of* 17^{th} *ISMIR Conference*, 2016.

[14] B. McFee and D. P. W. Ellis. Learning to segment songs with ordinal linear discriminant analysis. In *International conference on acoustics, speech and signal processing*, ICASSP, 2014.

[15] M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[16] M. Müller, P. Grosche, and N. Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. *Proc. of* 12^{th} *ISMIR Conference*, pages 615–620, 2011.

[17] M. Müller and F. Kurth. Enhancing similarity matrices for music audio analysis. *Proc. of ICASSP*, 2006.

[18] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. *Proc. of* 11^{th} *ISMIR Conference*, pages 625–636, 2010.

[19] C. Rhodes and M. Casey. Algorithms for determining and labelling approximate hierarchical self-similarity. *Proc. of* 8^{th} *ISMIR Conference*, 2007.

[20] C.S. Sapp. Online database of scores in the humdrum file format. *Proc. of* 6^{th} *ISMIR Conference*, pages 664–665, 2005.

[21] C. J. Tralie. Early MFCC and HPCP fusion for robust cover song identification. *Proc. of* 18^{th} *ISMIR Conference*, 2017.

[22] A. L. Wang. An industrial-strength audio search algorithm. In *Proc. of* 4^{th} *ISMIR Conference*, 2003.