Exploration of Latent Representations Generated by a Neural Network from Piano Music

Javier Campos, Rodrigo Cádiz, Marcelo Mendoza

Pontificia Universidad Católica de Chile

December 2024

Objective

- Investigate the latent space produced by Encodec for piano compositions.
- Analyze its capacity to distinguish musical styles by composers.

Dataset

Dataset: MAESTRO V3.0.0

- High-quality recordings (44.1-48 kHz, 16-bit PCM stereo).
- 1184 recordings (170 hours) of 430 compositions by various classical composers.
- Includes audio and MIDI transcriptions (MIDI not used in this study).
- Split: Training, validation, testing subsets.

Neural Network - Encodec

Architecture:

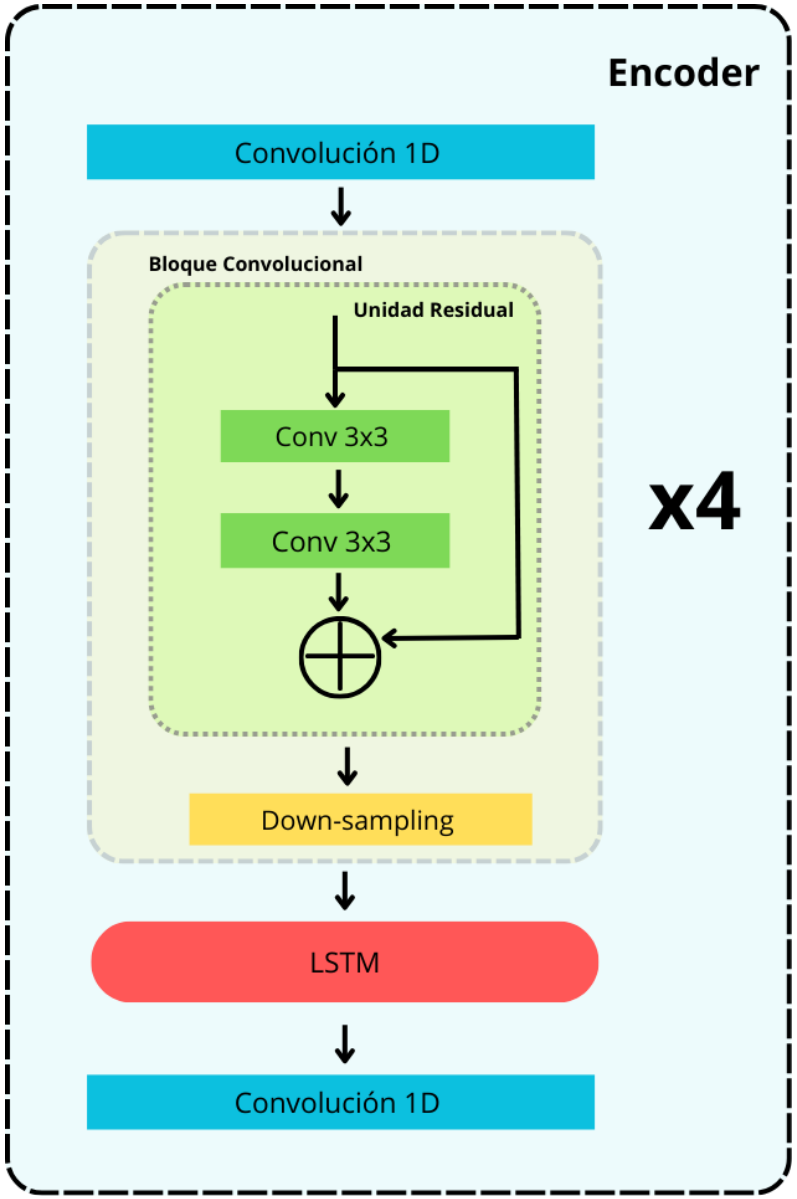
- Encoder-decoder model.
- Encoder creates latent representations; decoder reconstructs audio signals.
- Layers include 1D convolutional, residual units, LSTM, and final convolutional layers.

Model Used:

- 48 kHz stereo, non-causal variant, trained on music data.

Purpose: Use the encoder for feature extraction.

Neural Network - Encodec



Feature Extraction

Process:

- Extract features from all audio recordings.
- Fragments tested: 10 seconds and 1 minute.
- Time for extraction[1] :
  - 10 seconds: 30 minutes.
  - 1 minute: 2 hours 40 minutes.

[1] using NVIDIA GeForce RTX 4060 GPU

Dimensionality Reduction

Visualization Techniques:

- Principal Component Analysis (PCA): Reduces dimensionality to 50 components.
- t-SNE (t-Distributed Stochastic Neighbor Embedding): Projects data into 2D for visualization.
- Goal: Maintain local structure in high-dimensional space.

Steps:

- Flatten latent tensors.
- Apply PCA, then t-SNE.

Results - Visualization

Observations:

- Clusters formed in latent space indicate similarity in musical styles.
- Example: Compositions by composers with similar styles tend to group.
- Composers with diverse styles (e.g., Mozart, Beethoven) appear more spread out.
- Peripheral clusters often belong to composers with fewer recordings.

Challenges

Dimensionality Reduction:

- Loss of useful information during PCA.
- Noise introduced in visualization.
- t-SNE requires careful tuning (e.g., perplexity parameter).

Latent Space Interpretation:

- Lack of clarity on criteria used by the neural network for grouping.

Conclusions

Findings:

- Encodec can effectively group piano compositions based on learned latent space representations.
- Demonstrates potential for distinguishing musical styles.

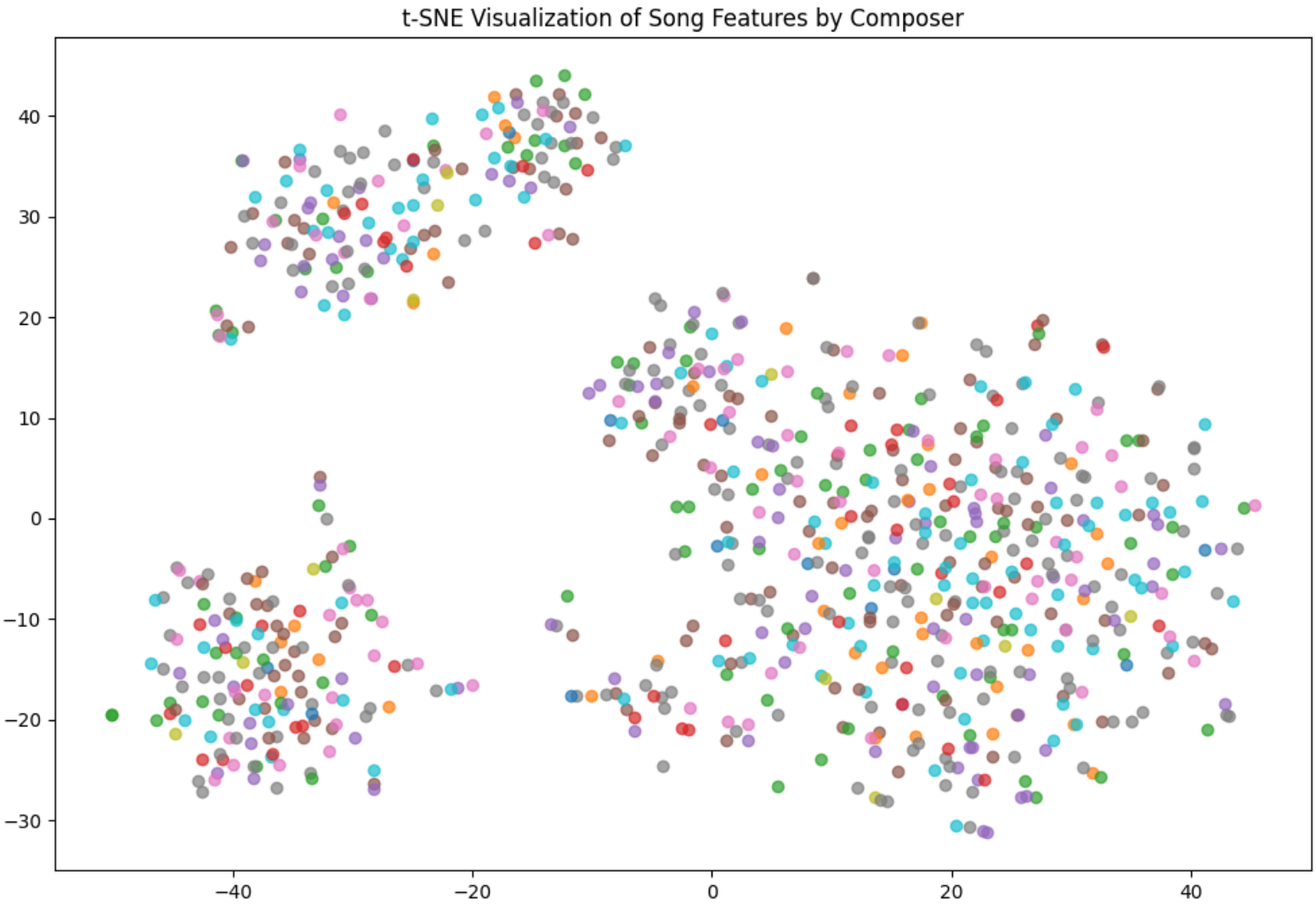
Future Work:

- Include contrasting data (e.g., non-musical audio).
- Fine-tuning Encodec for better latent space clarity.
- Probing latent features using classification tasks.

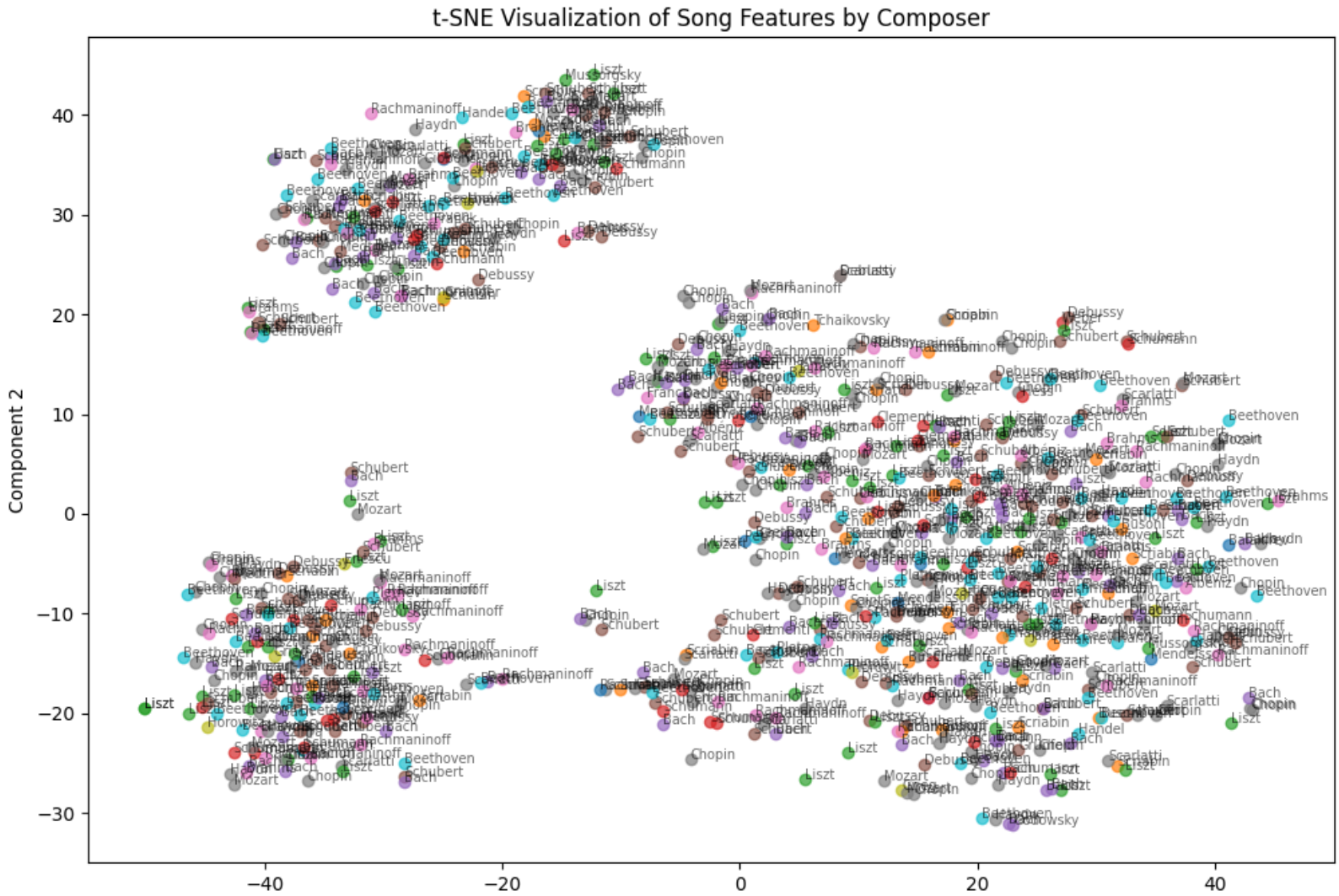
References

- Van den Oord et al., "WaveNet: A generative model for raw audio," 2016.
- D'efossez et al., "High fidelity neural audio compression," 2022.
- Hawthorne et al., "MAESTRO dataset," 2019.
- Van der Maaten & Hinton, "t-SNE," 2008.

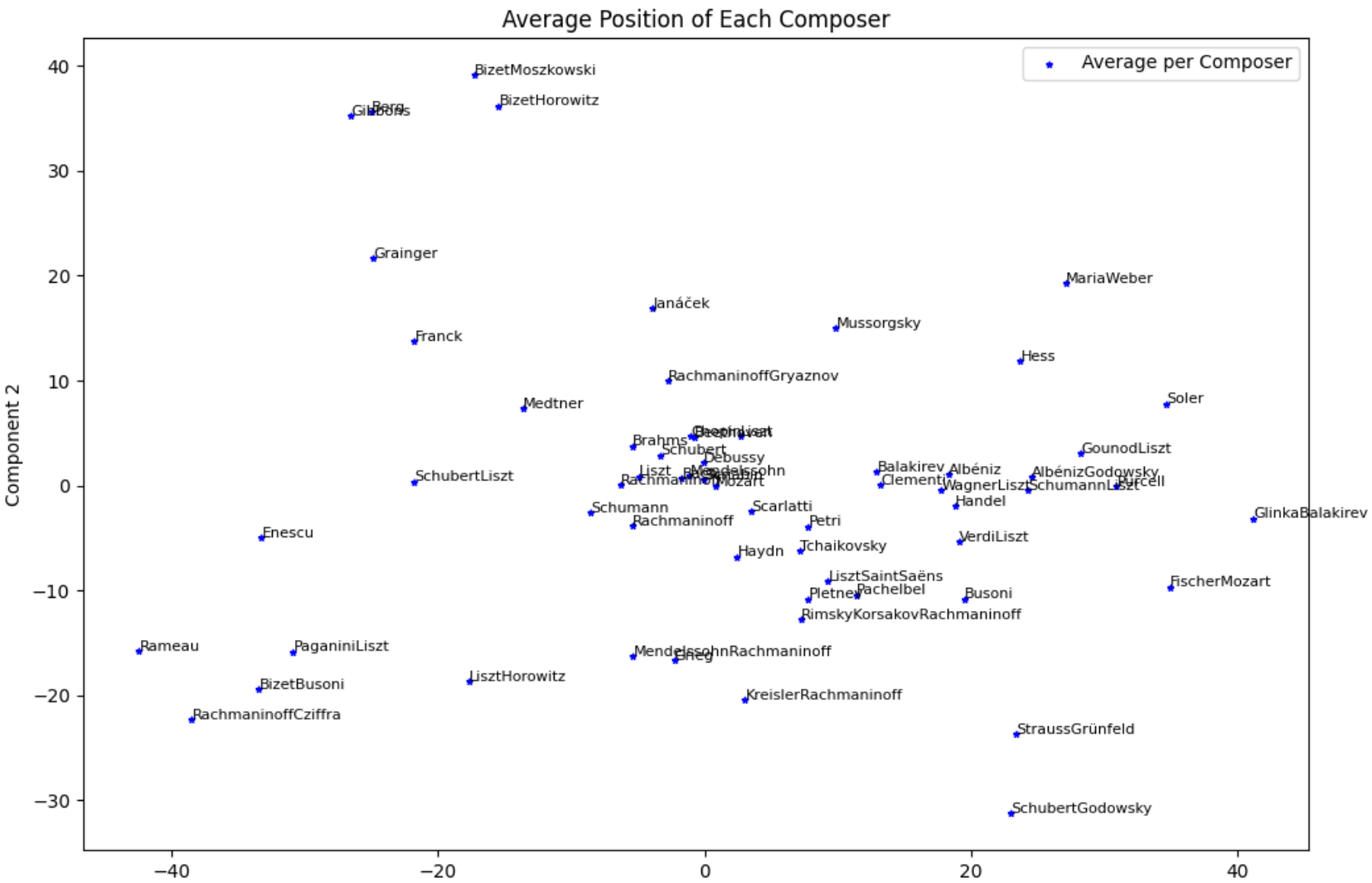
Visualizations



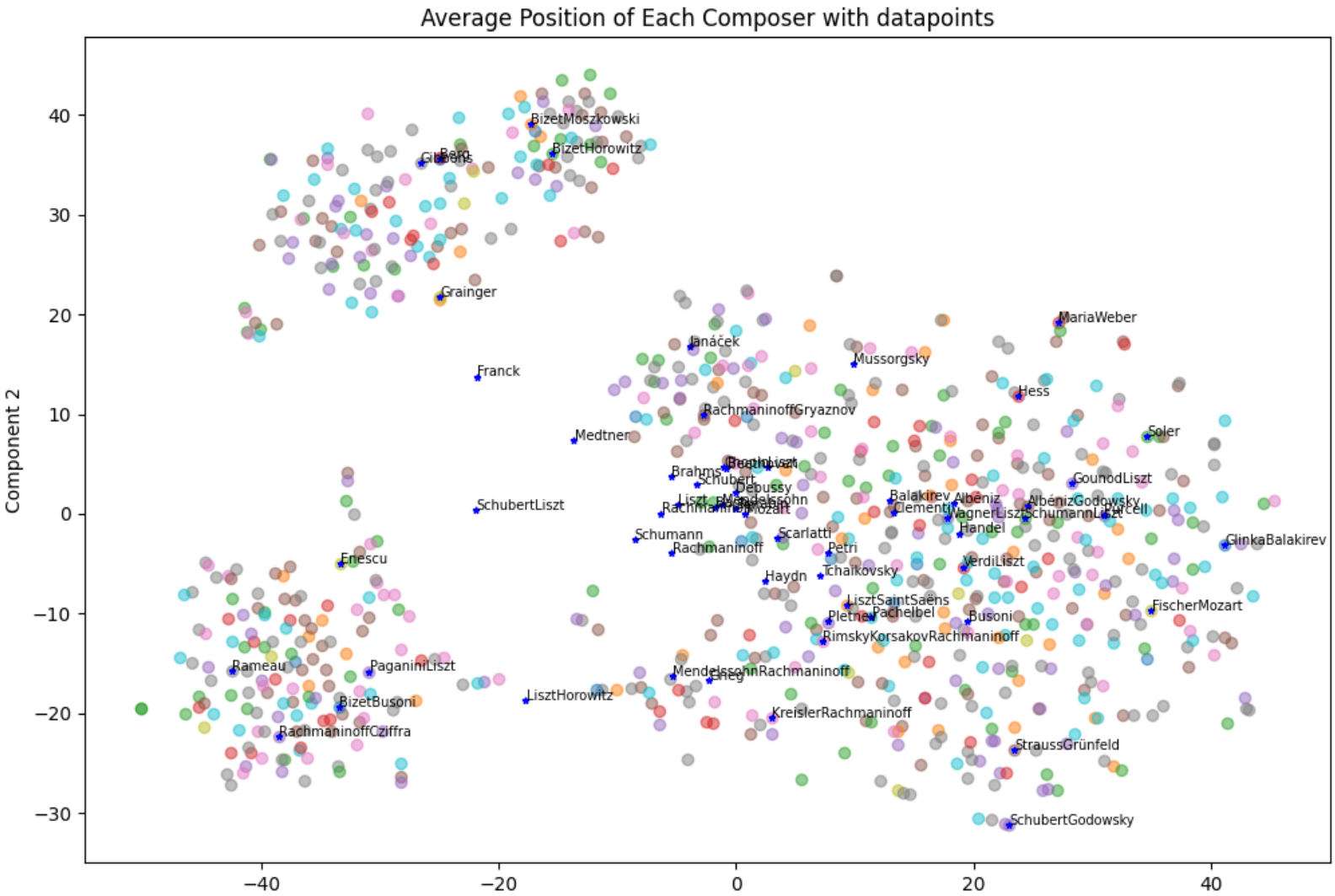
Visualizations



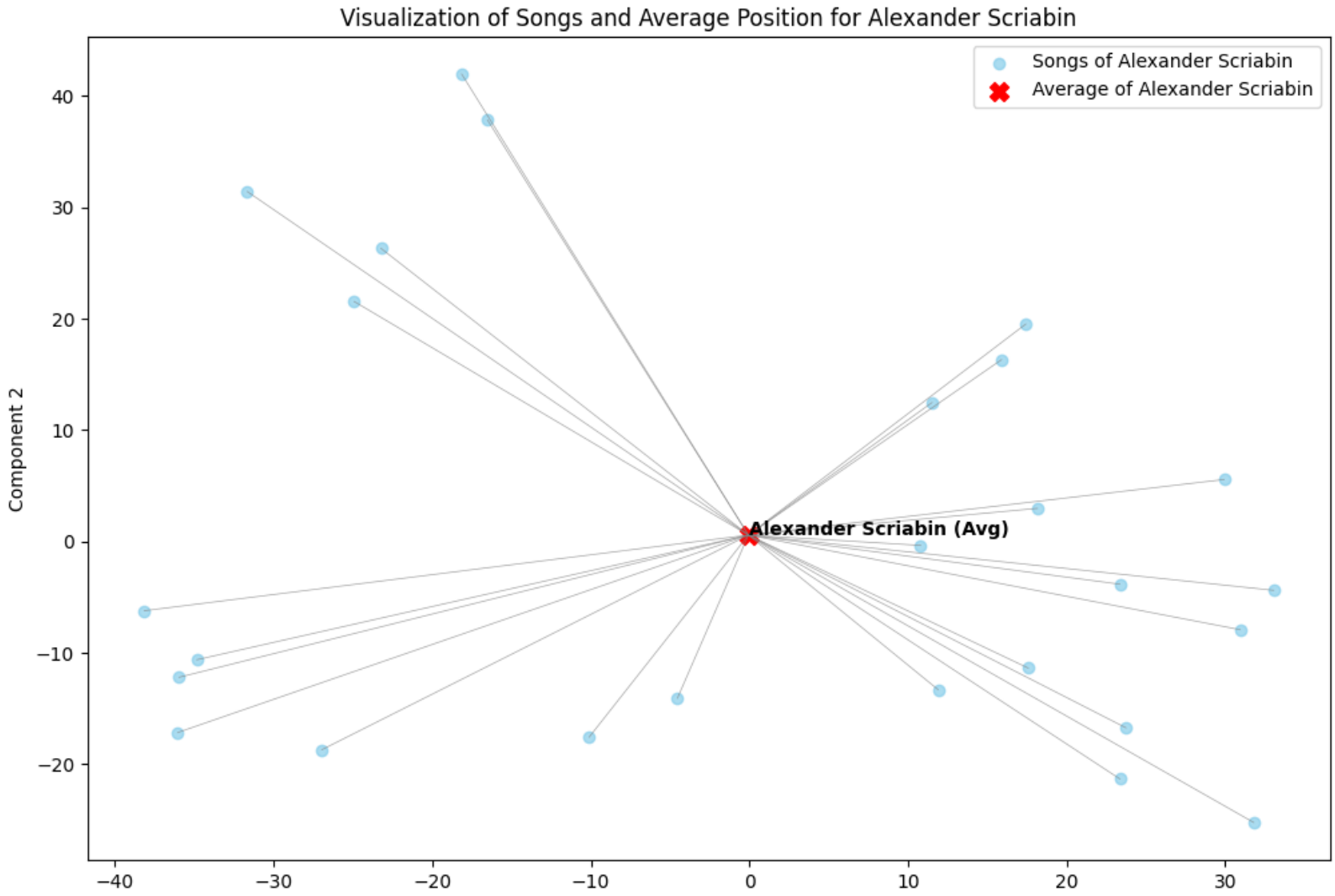
Visualizations



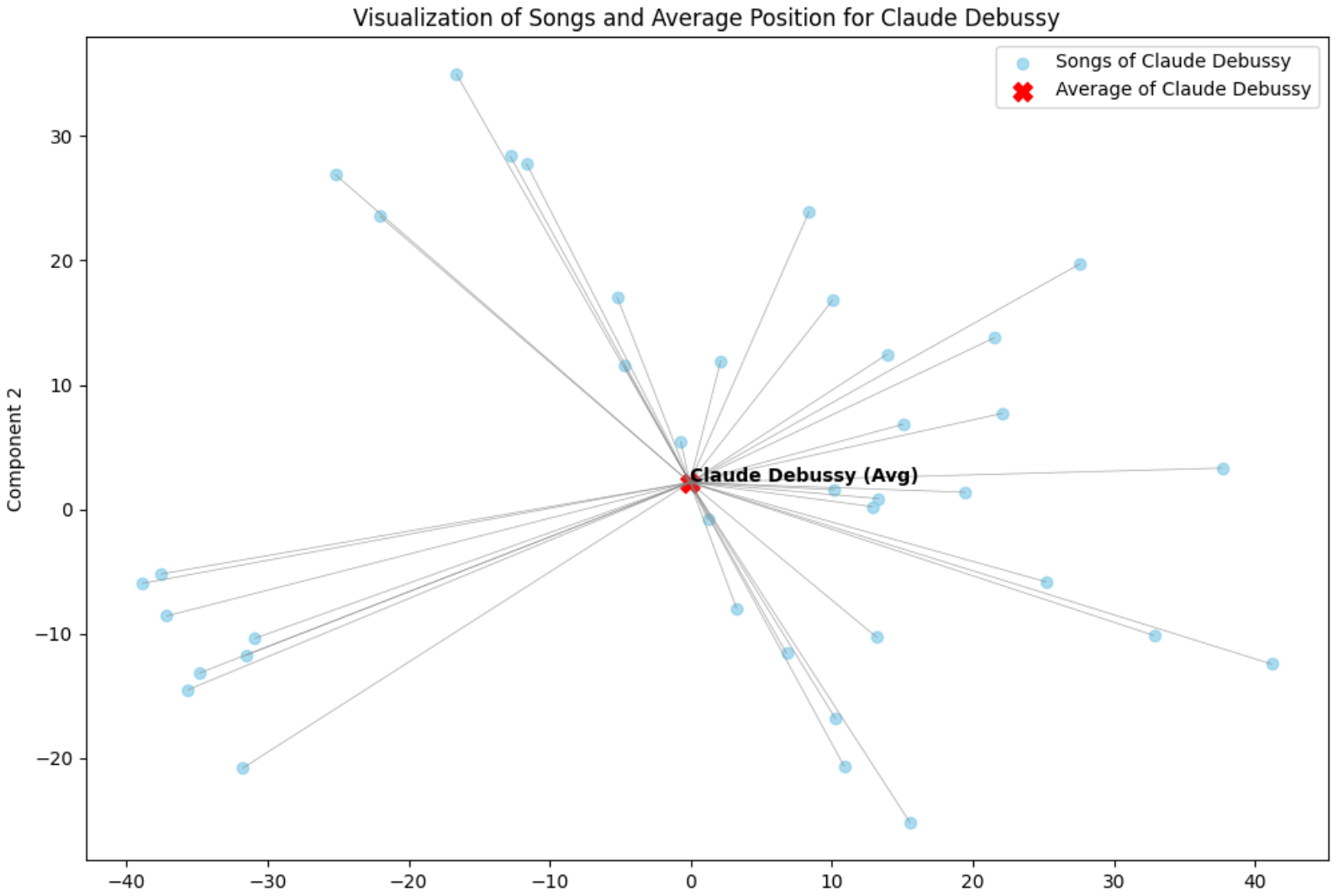
Visualizations



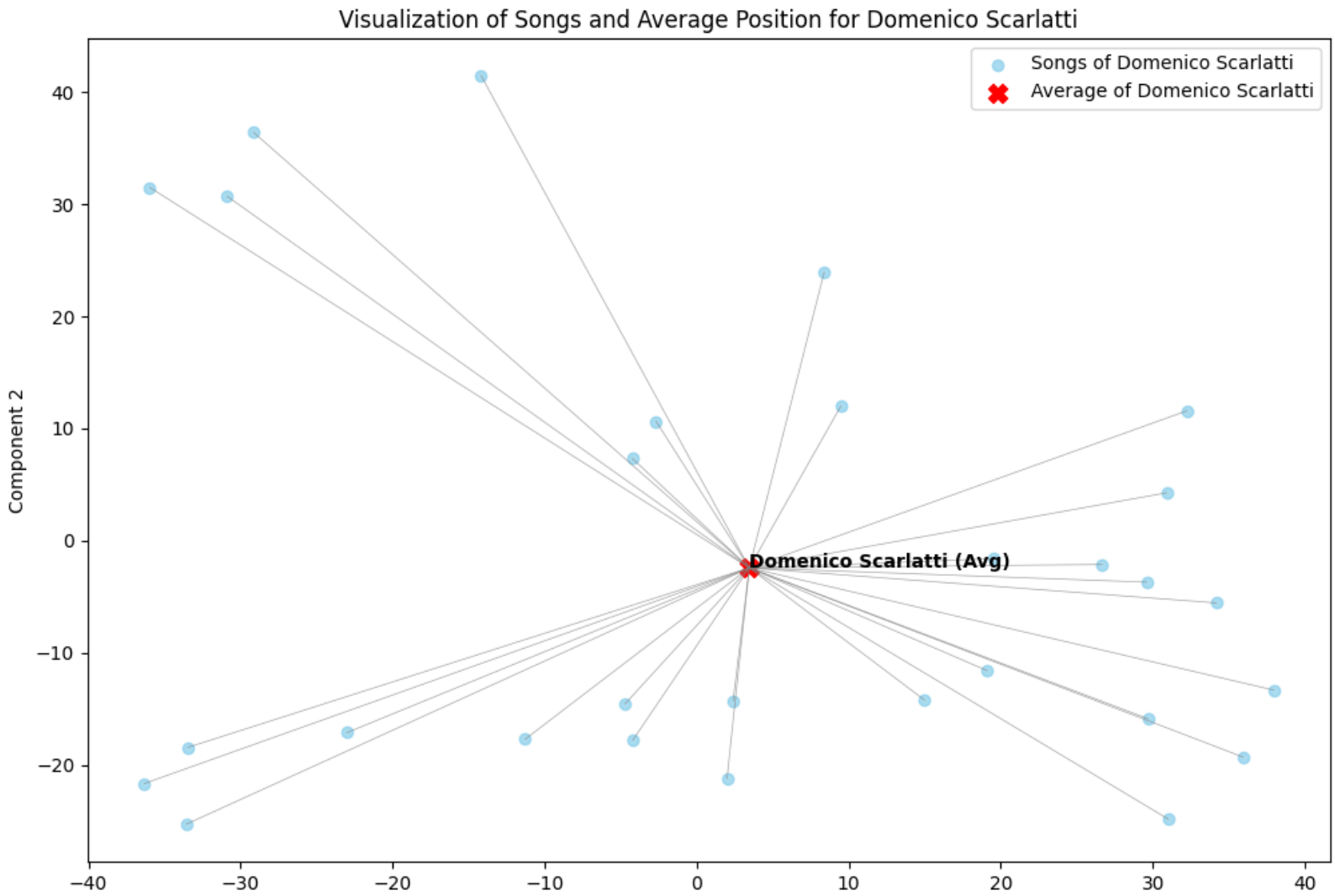
Visualizations



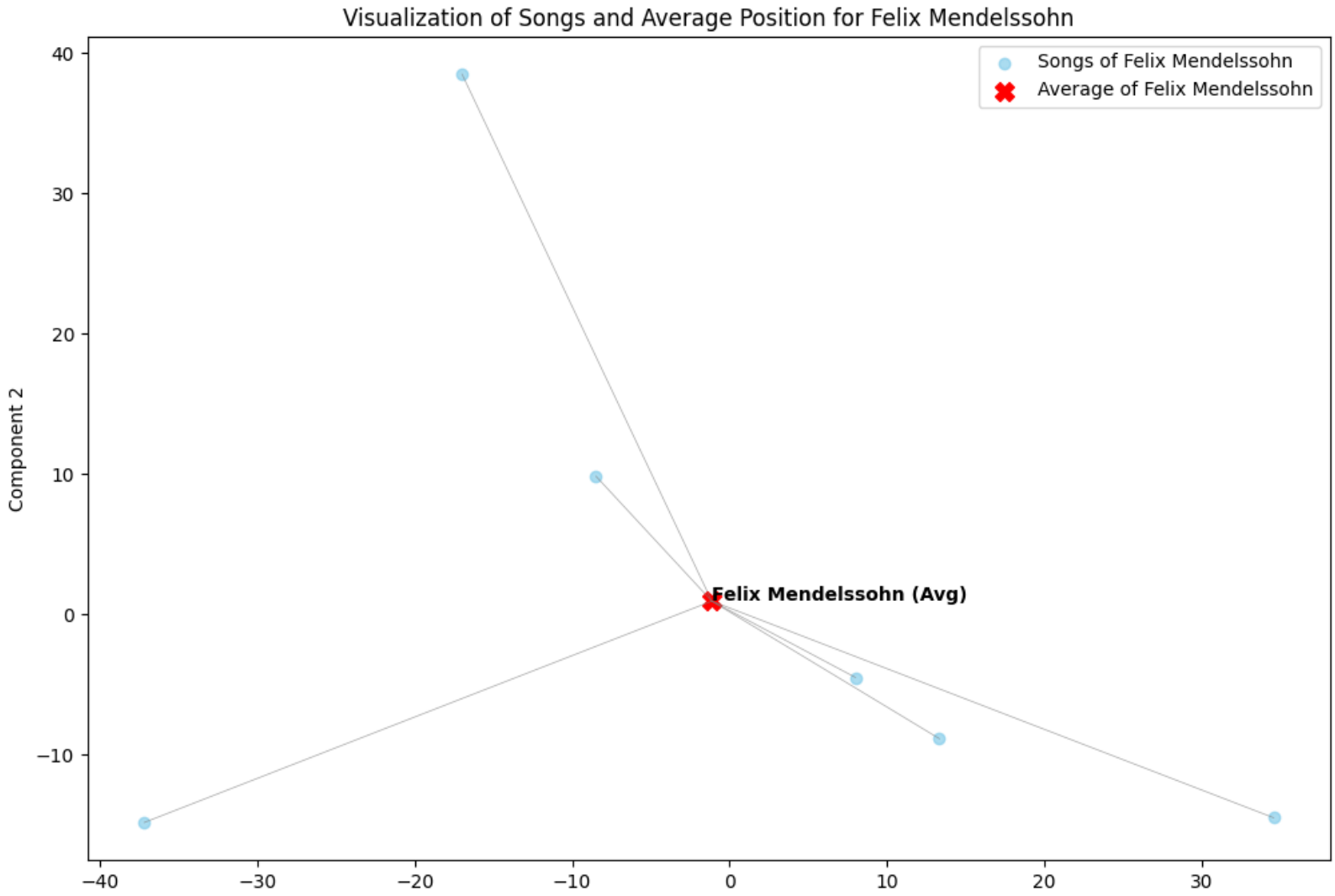
Visualizations



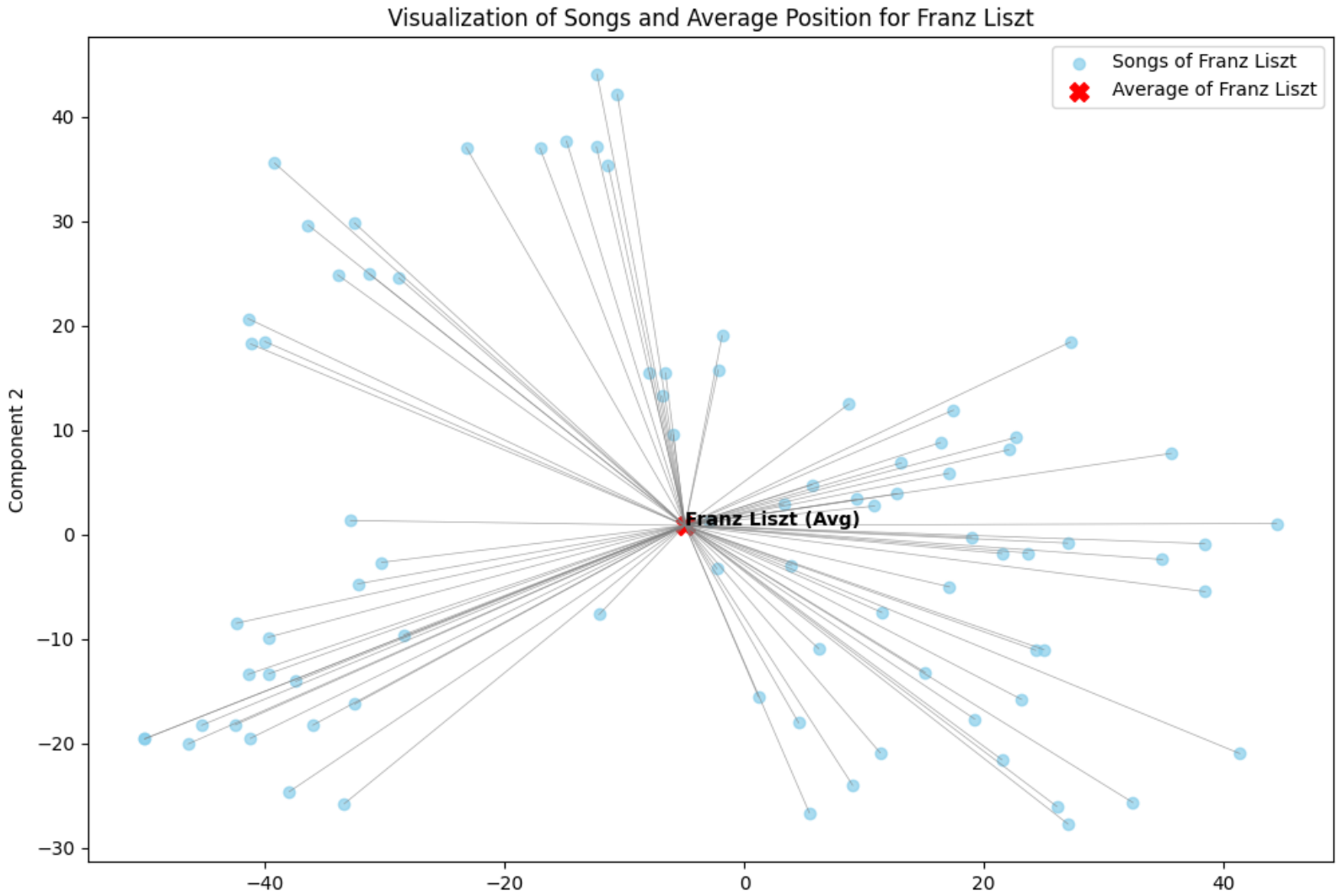
Visualizations



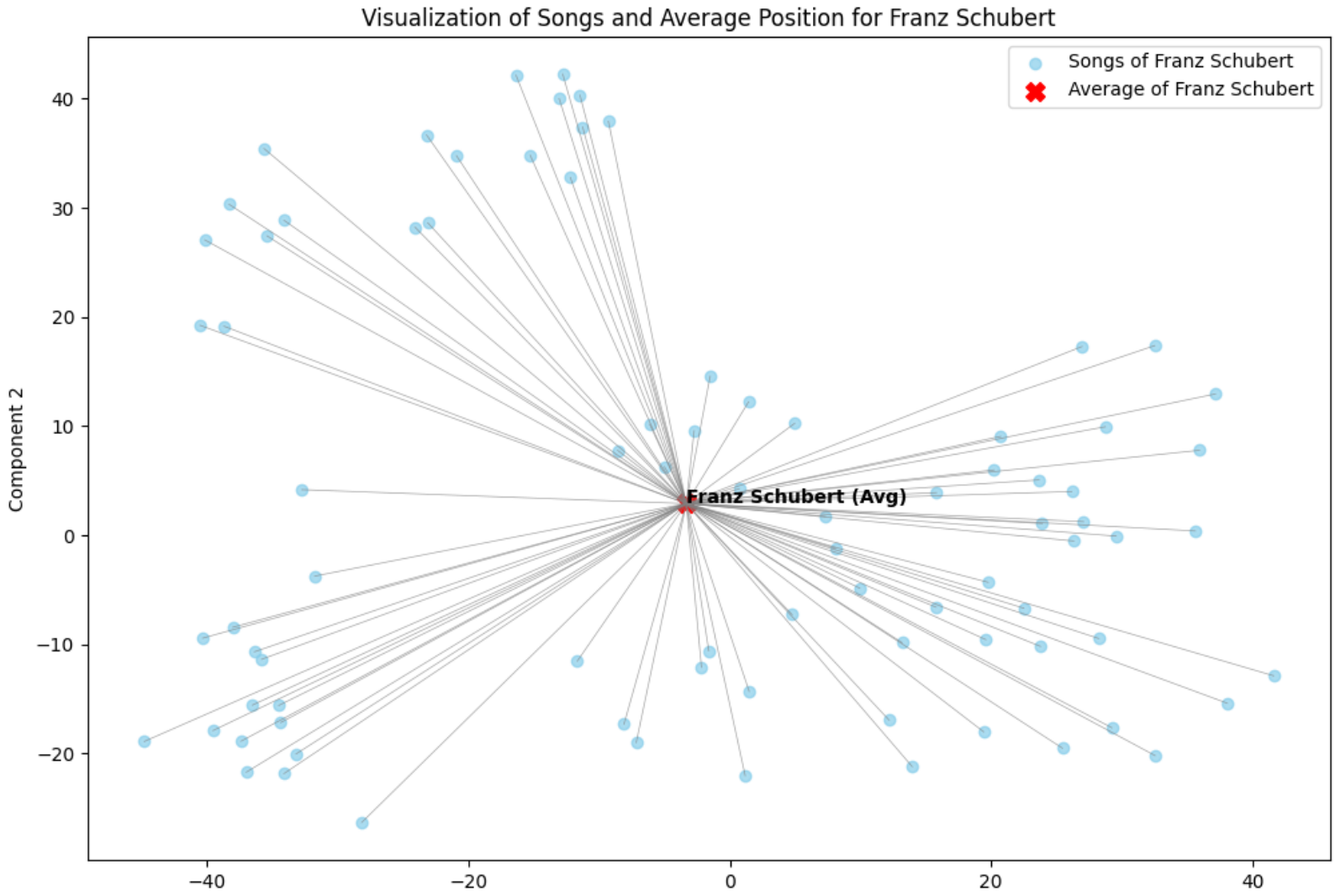
Visualizations



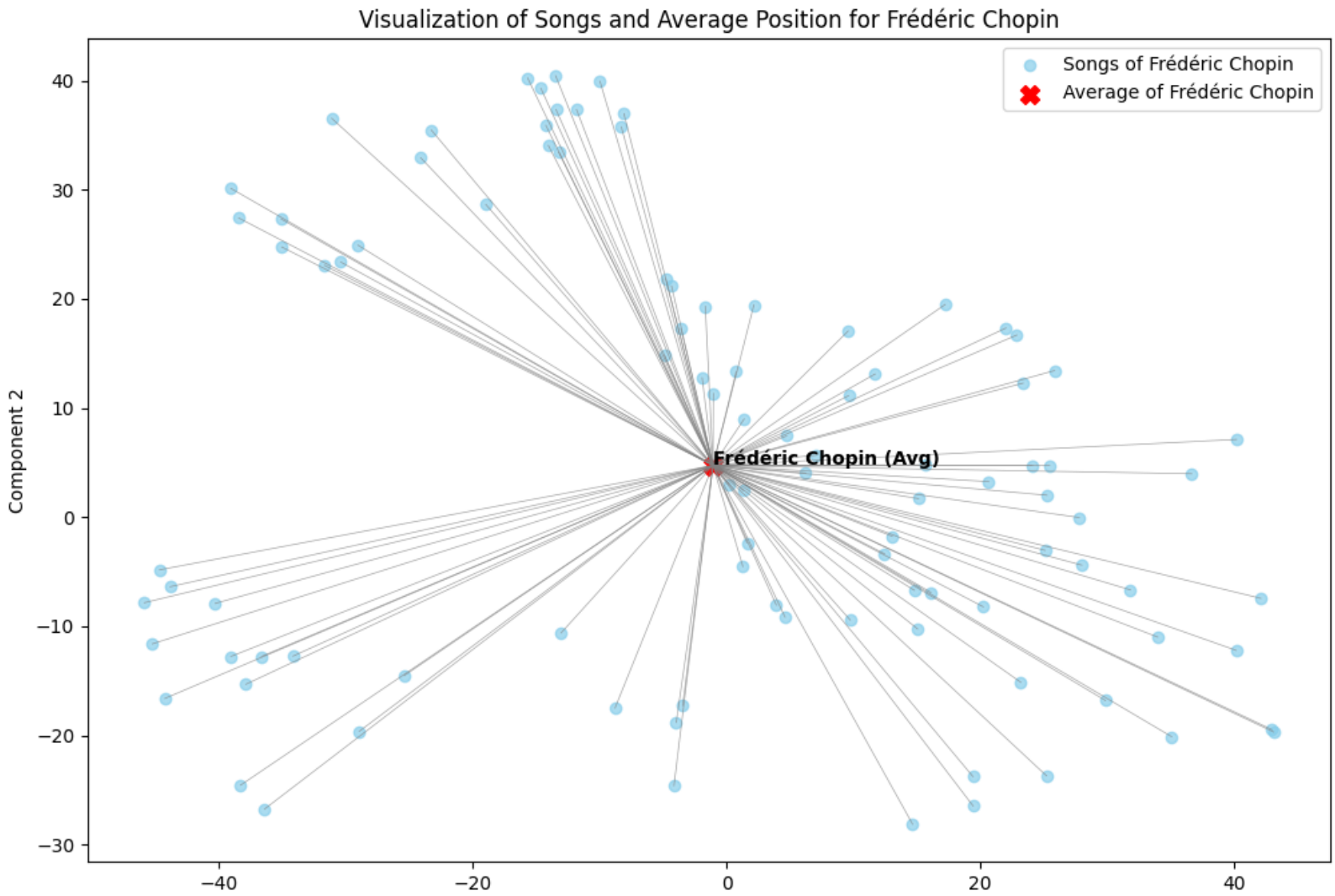
Visualizations



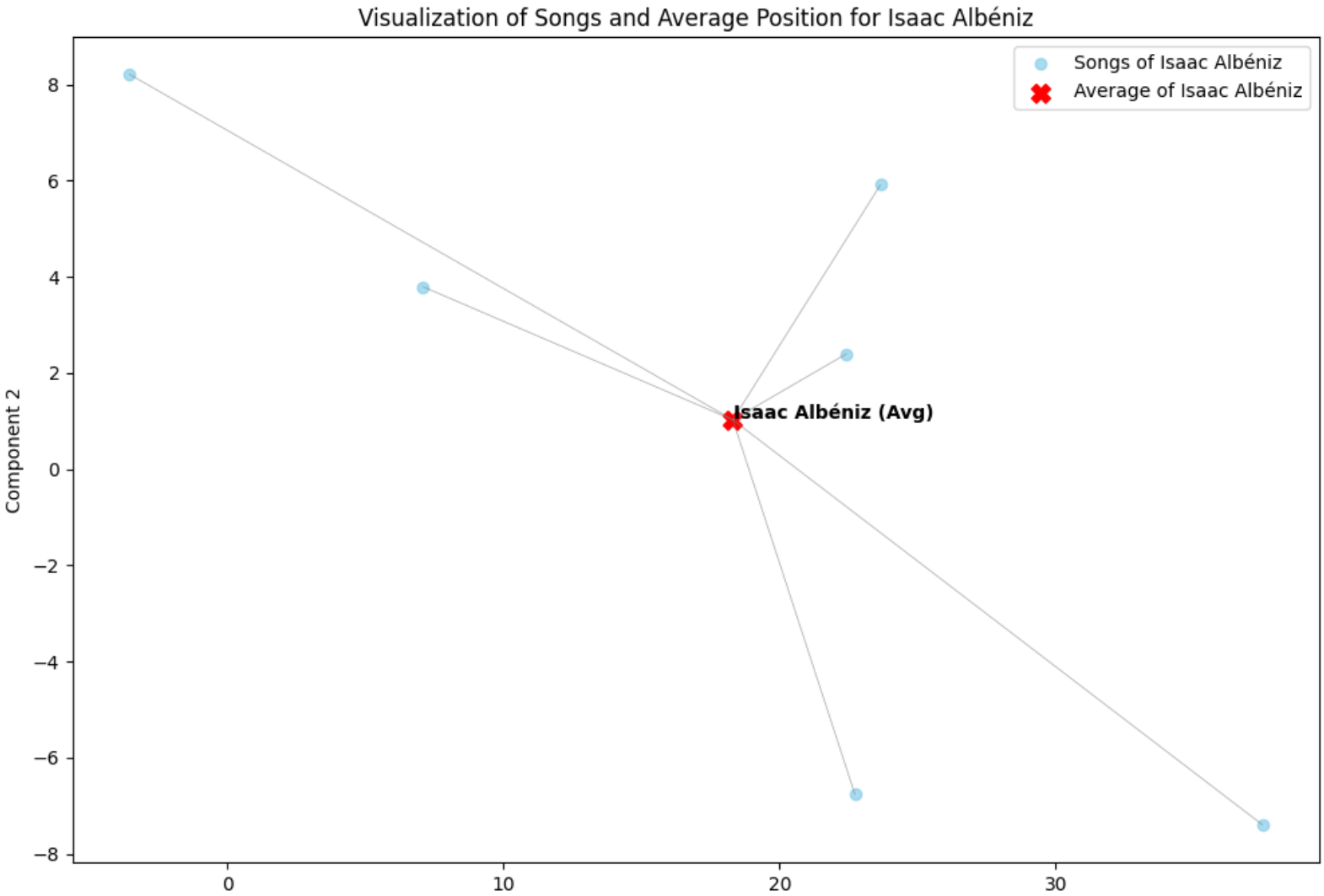
Visualizations



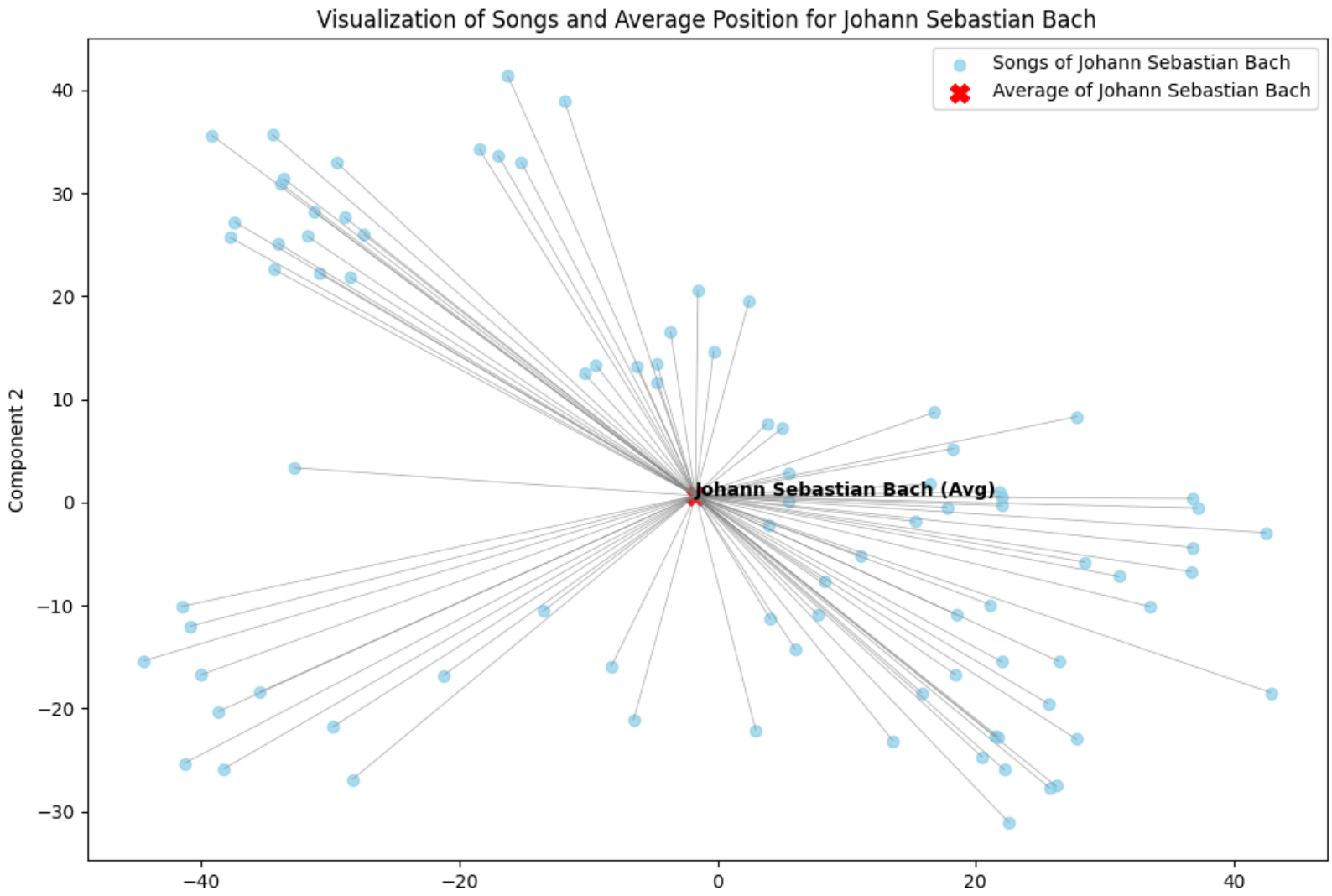
Visualizations



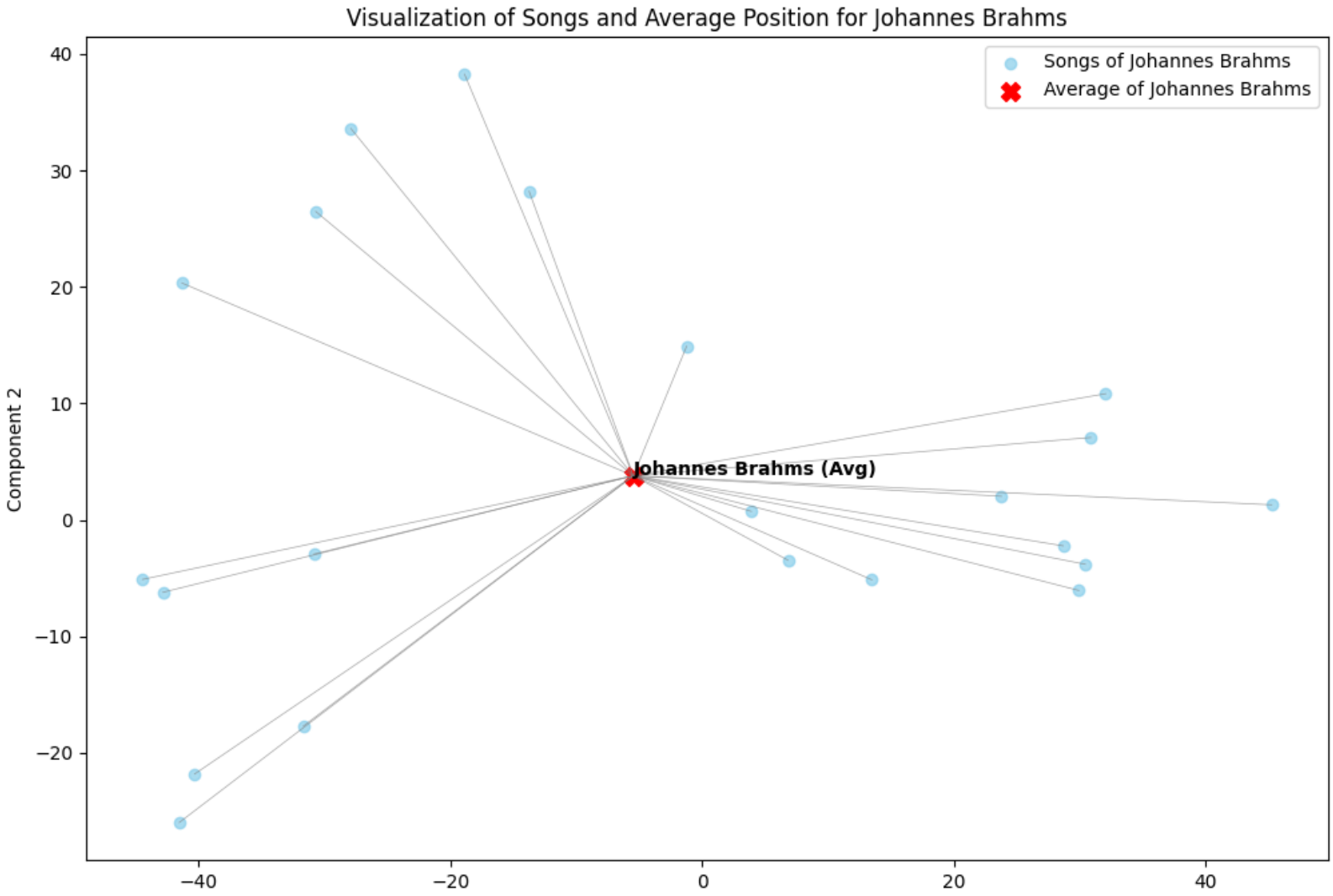
Visualizations



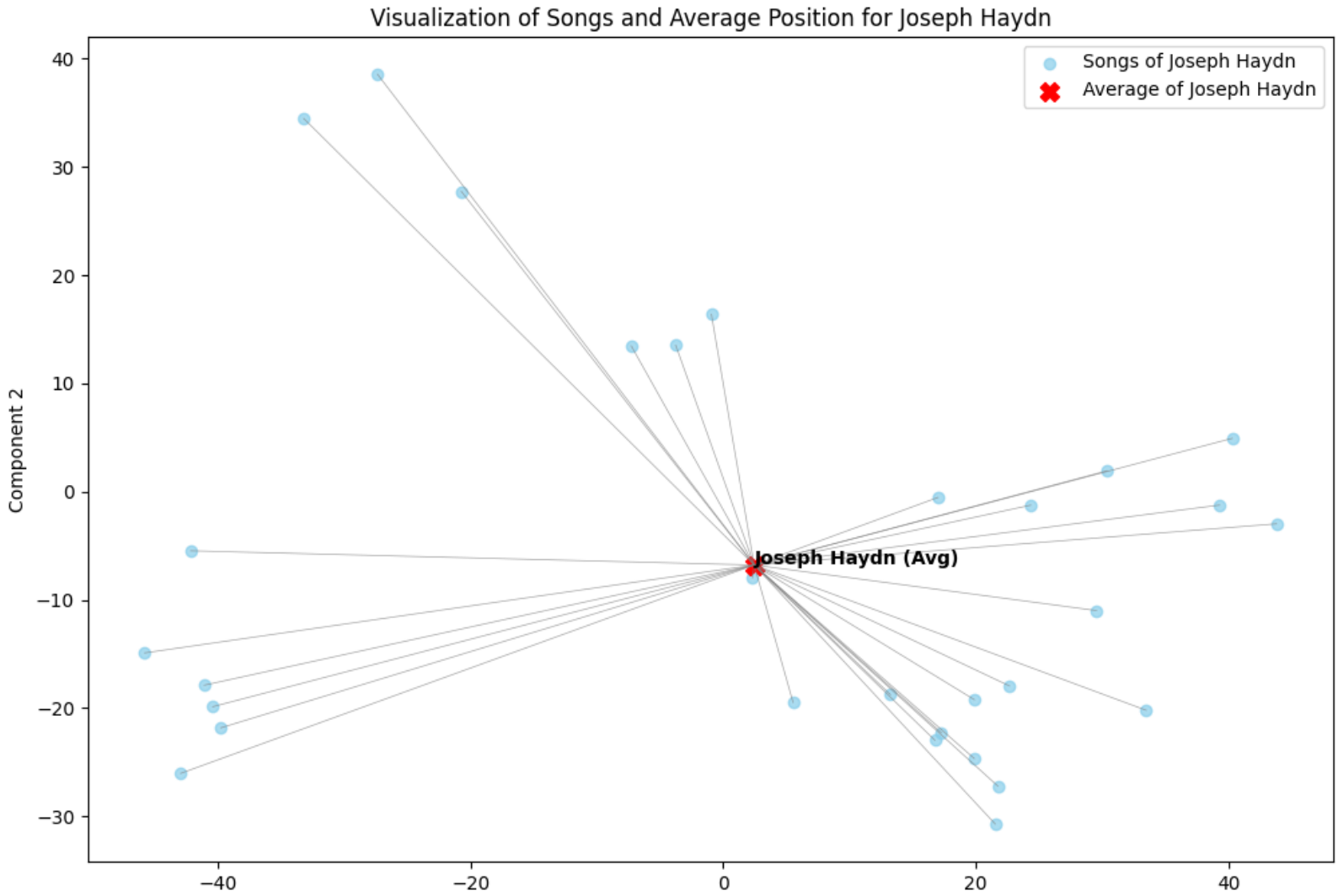
Visualizations



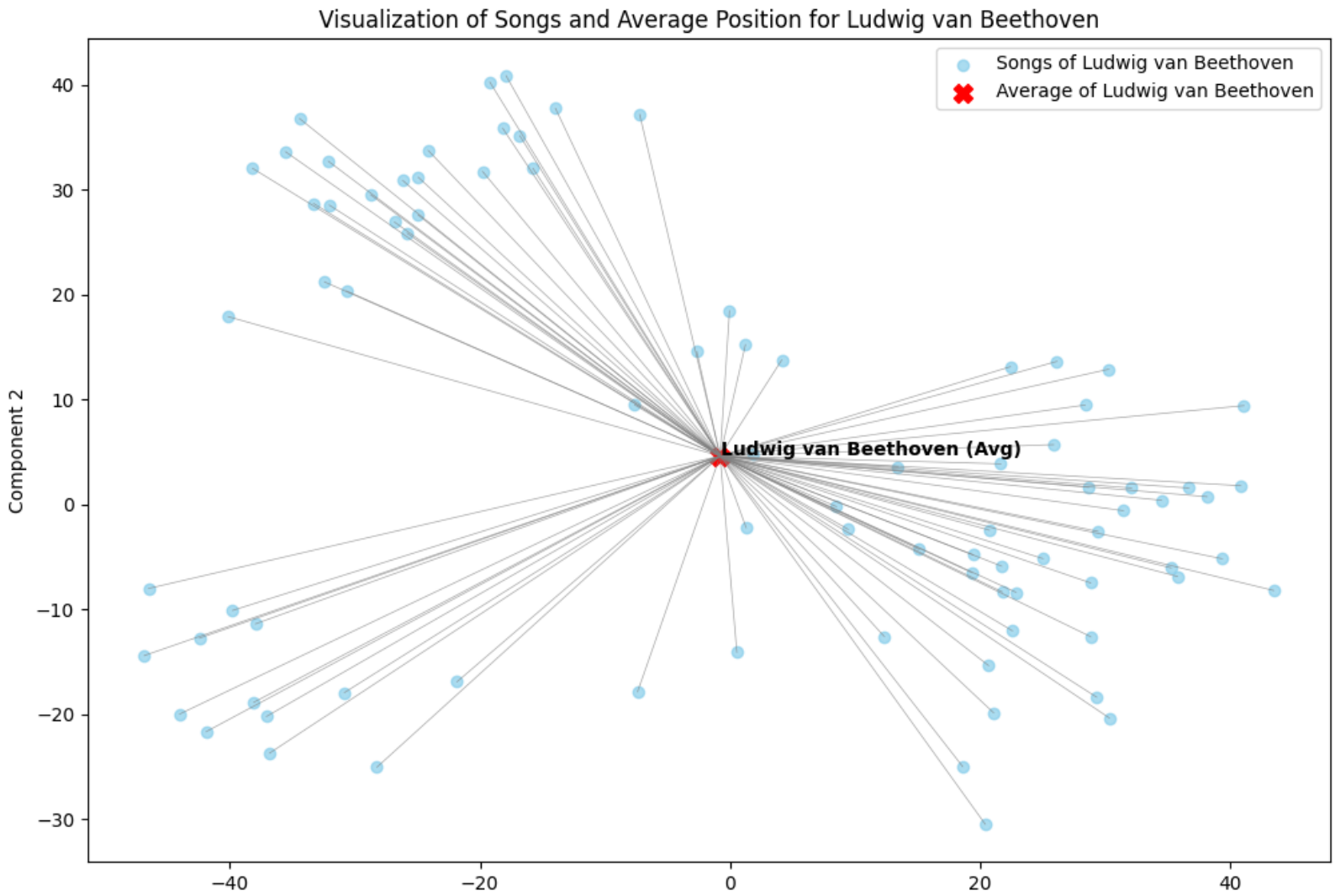
Visualizations



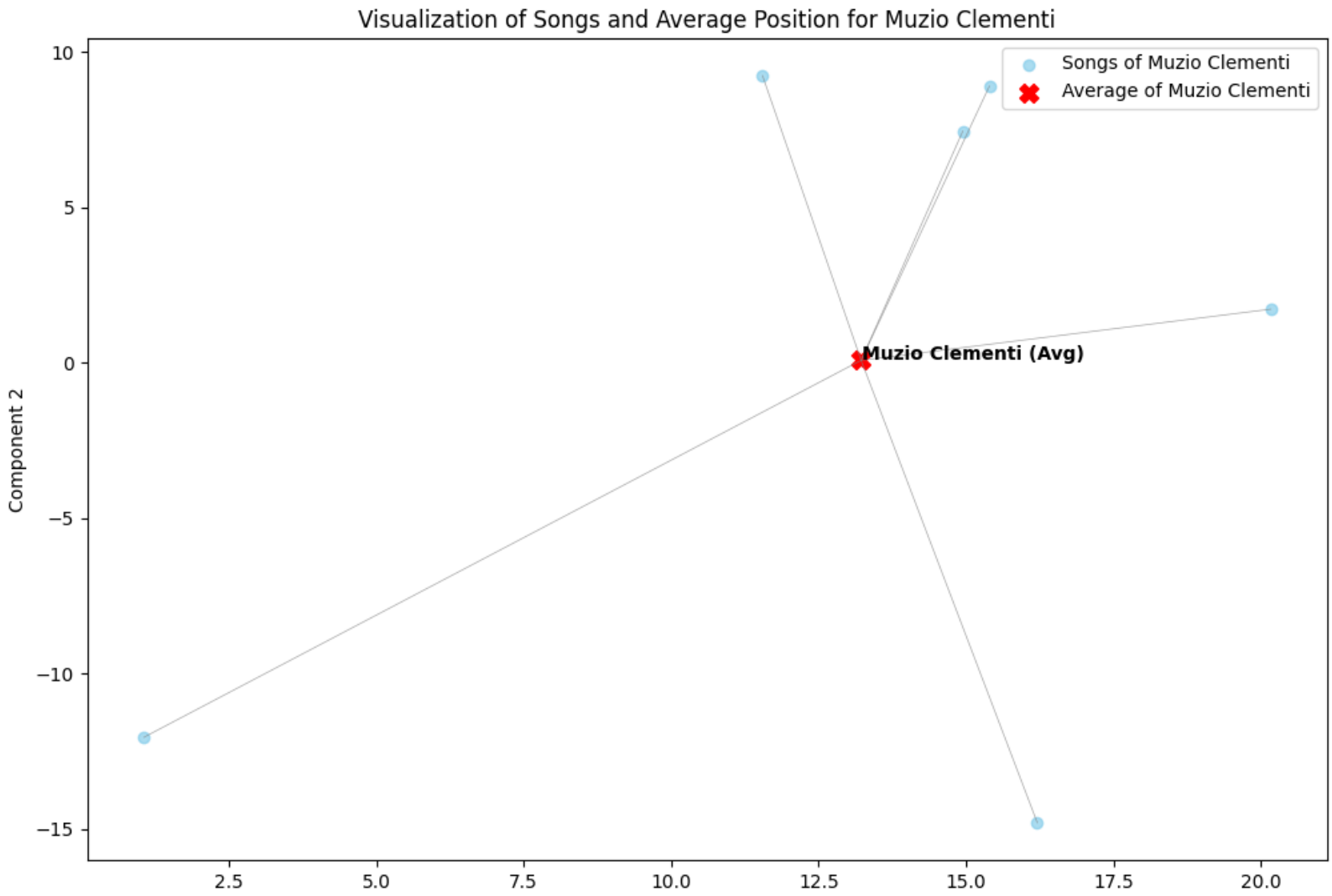
Visualizations



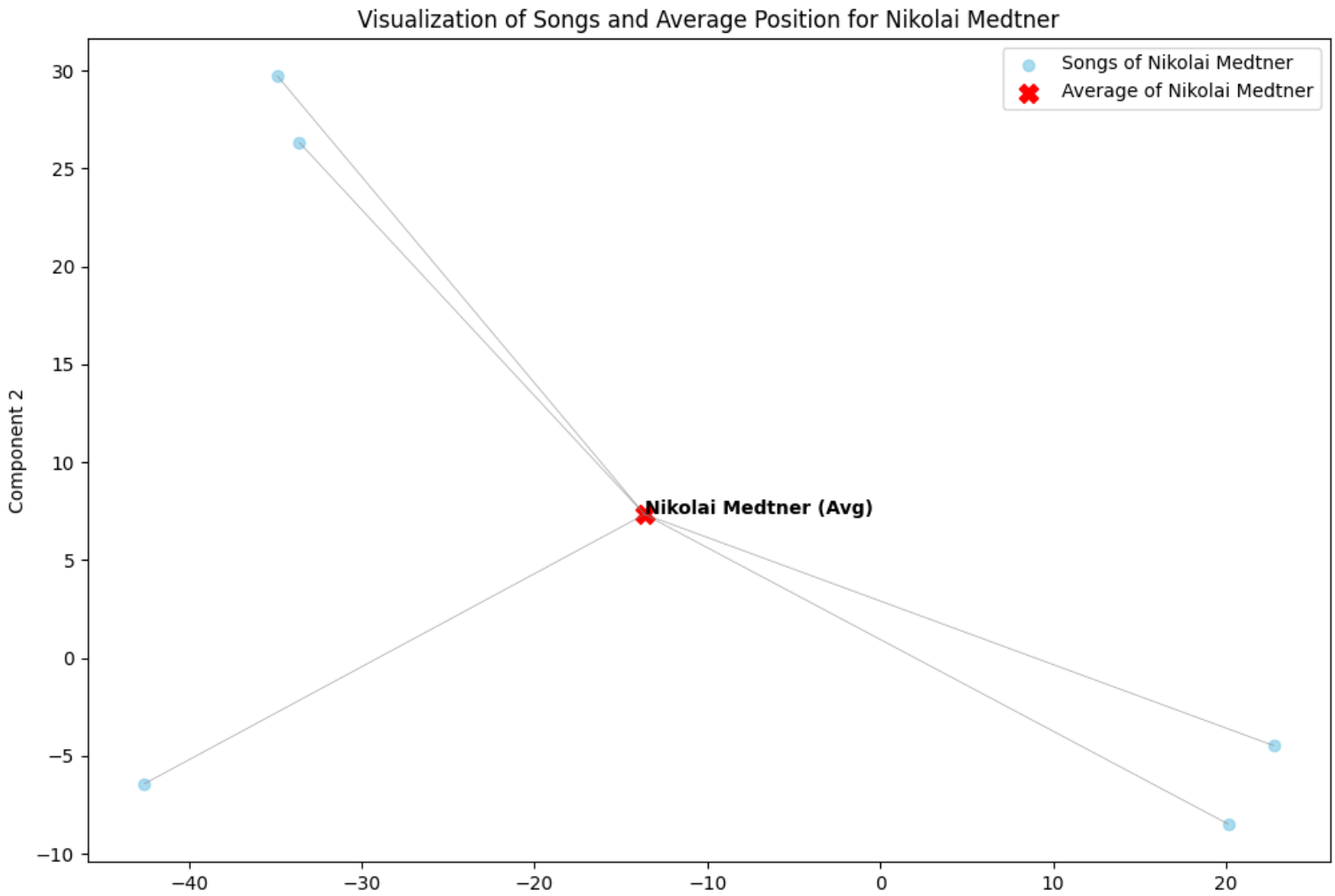
Visualizations



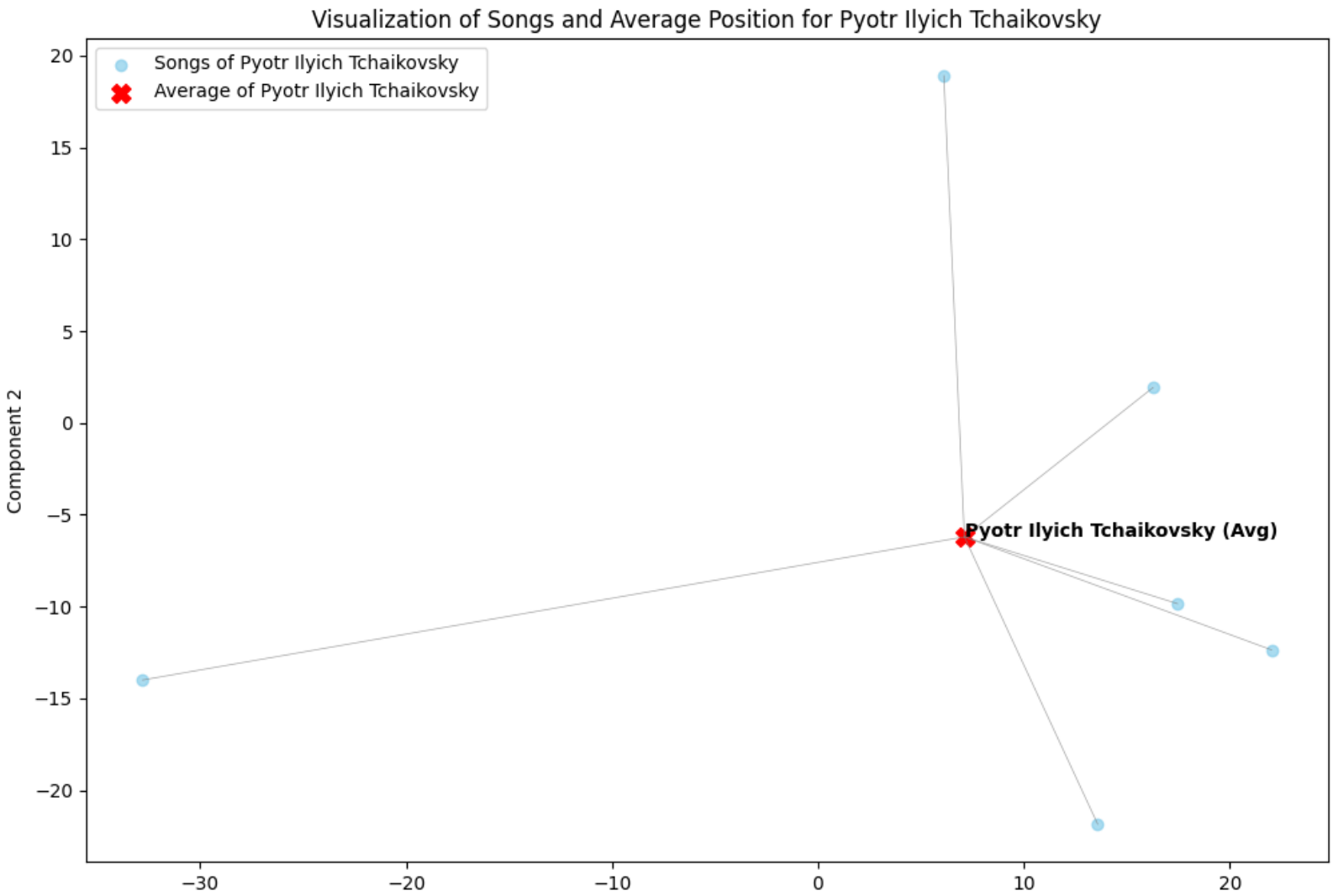
Visualizations



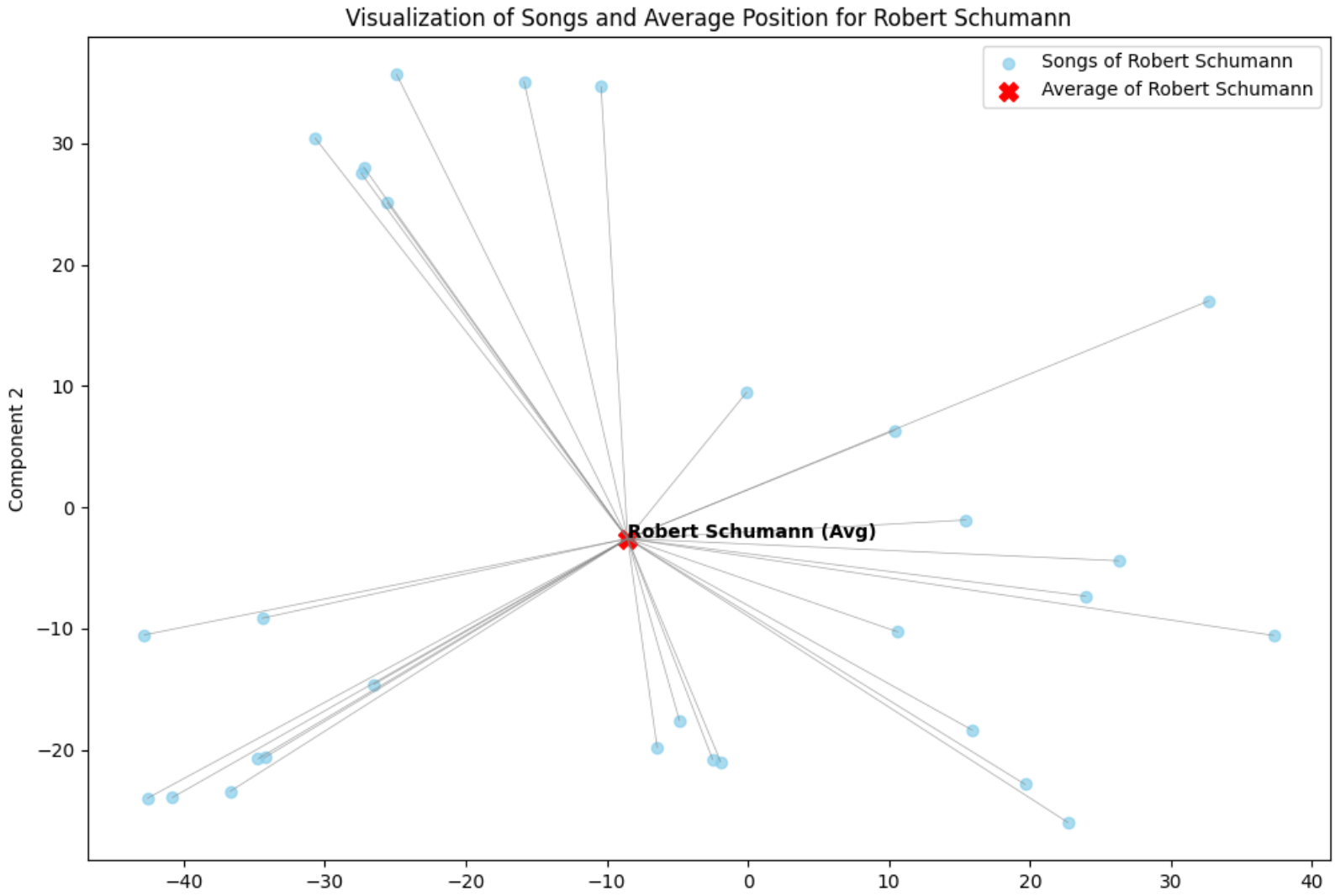
Visualizations



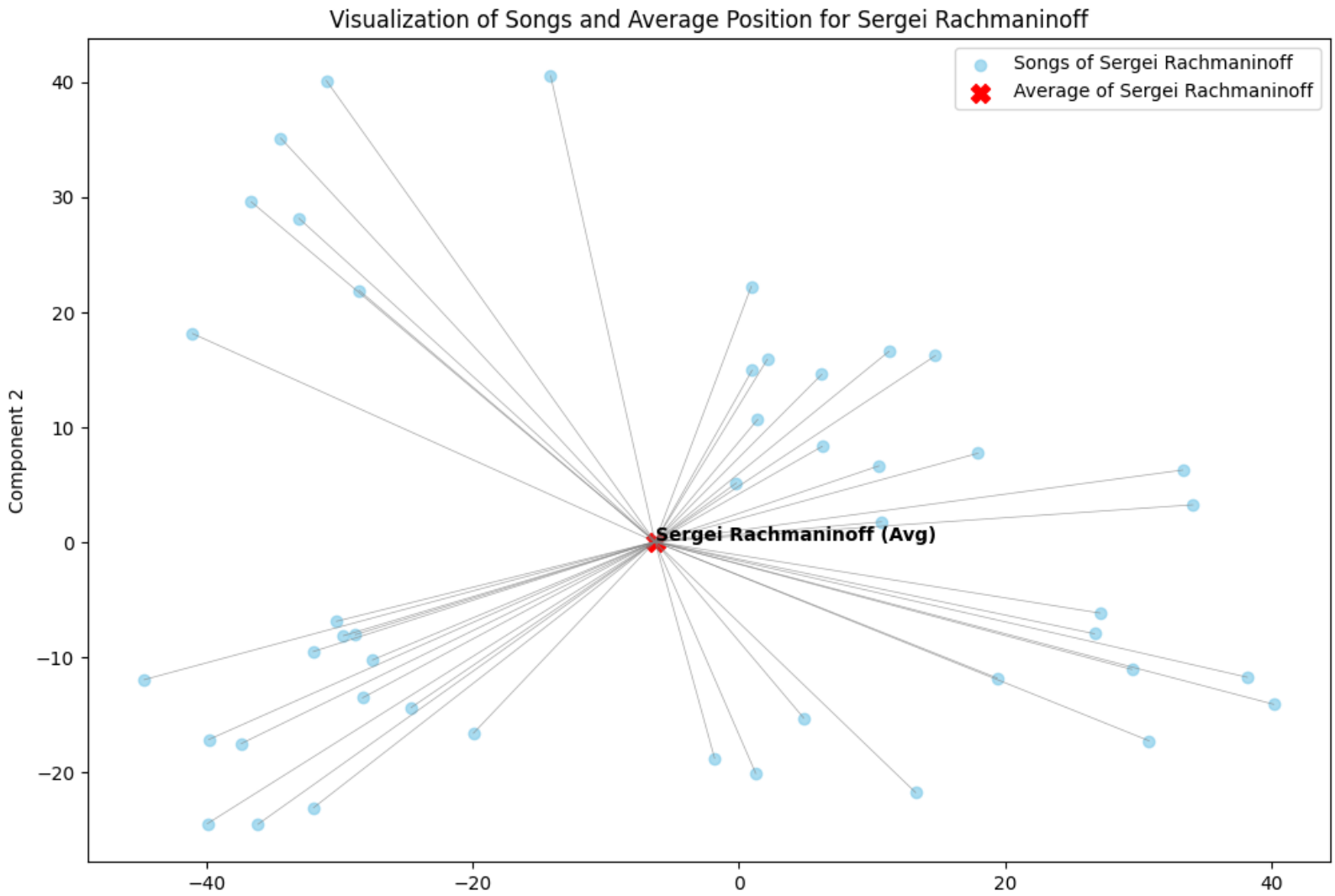
Visualizations



Visualizations



Visualizations



Visualizations