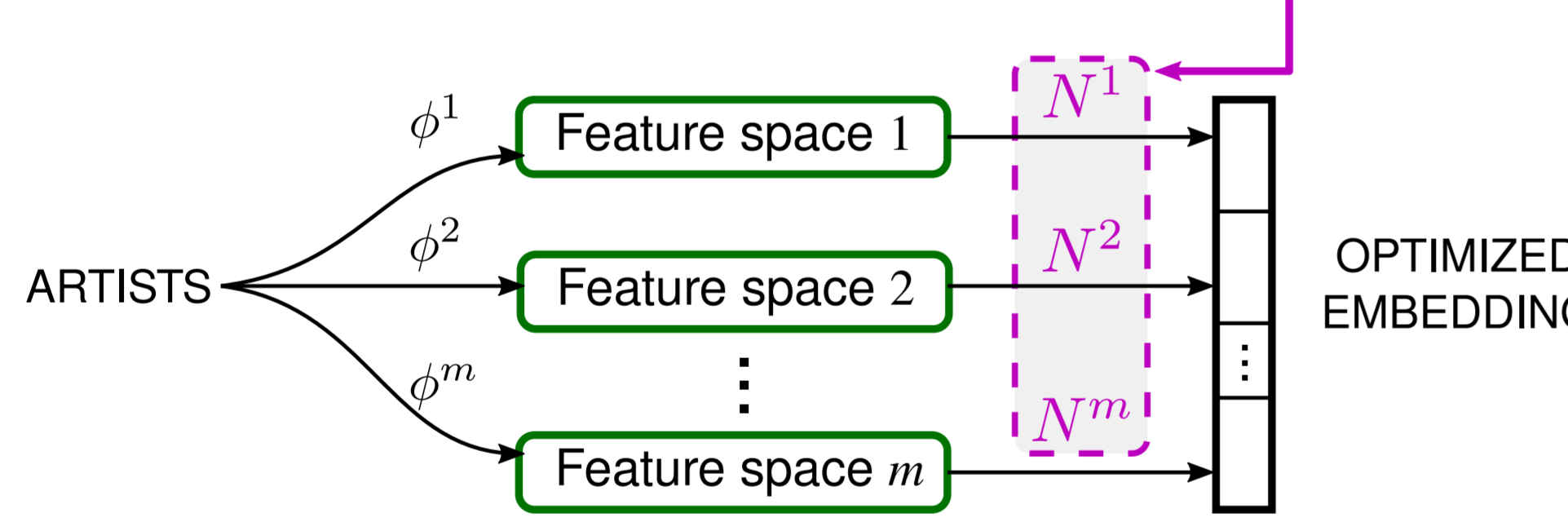


Heterogeneous Embedding for Subjective Artist Similarity


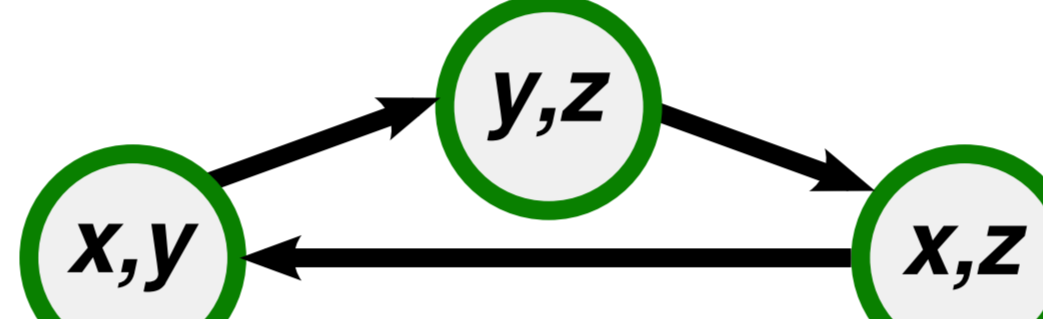
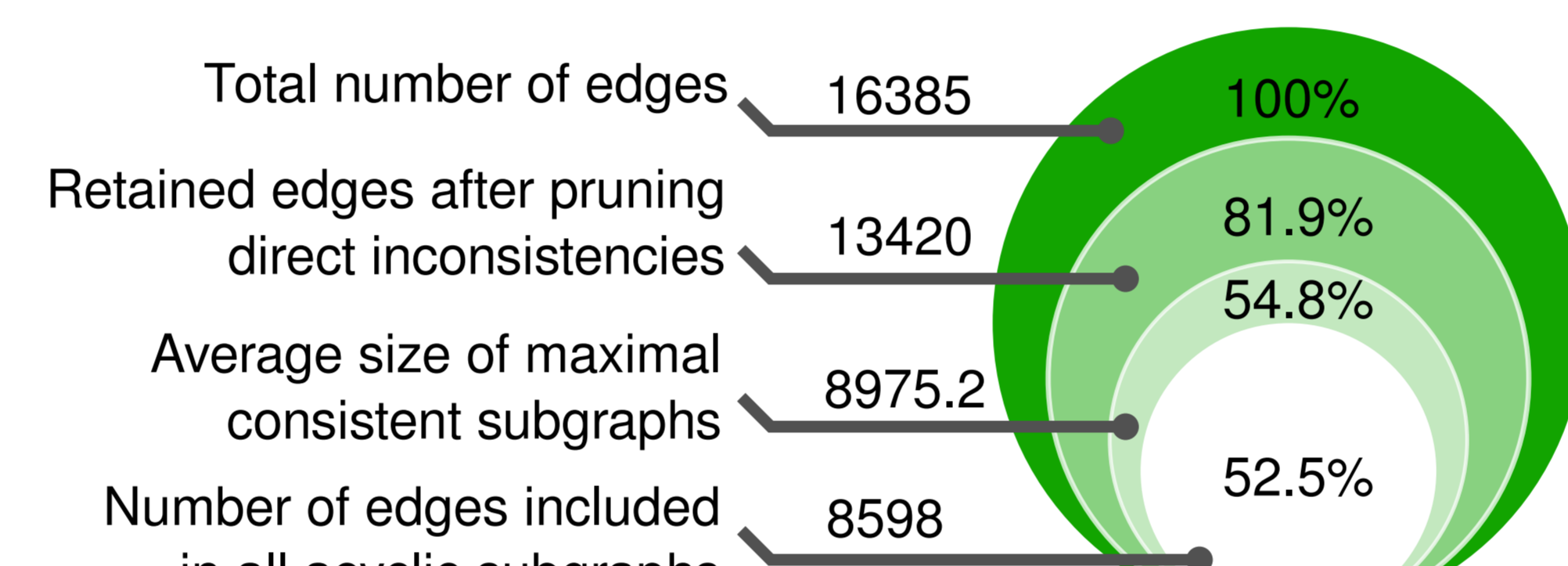
Overview

- Goal** Construct a similarity measure between artists that agrees with human perception.
- Similarity** Similarity between artists is expressed by **relative comparisons**:
 $(x,y,z) \leftrightarrow$ Artist x is **more similar** to y than to z .
- Consistency** Artist similarity is inherently *subjective*, and may vary from person to person.
How can we quantify consistency?
- Embedding** Our similarity measure is defined by the Euclidean distance between artists.
Given the variety of features available, what is the best way to combine them?
Human feedback will help us construct an optimal embedding from the input features.


Embedding Algorithm

- Idea** View each artist in heterogeneous **feature spaces** by using **multiple kernels**:
 $x \mapsto \{\phi^i(x)\} \quad K_{xy}^i = \langle \phi^i(x), \phi^i(y) \rangle$
- Problem** Features may disagree with human perception
Features are not all equally informative
- Solution** Construct an **optimal embedding** from the feature spaces by learning **projections**
- 
- Human perception** The optimization is constrained to match human perception measurements by **Partial Order Embedding** [1]:
 $(x,y,z) \leftrightarrow d^2(x,y) + 1 \leq d^2(x,z)$

Quantifying Consistency

- Data** We use *aset400* [2]:
412 popular artists
16385 similarity measurements
- Direct inconsistency** Disagreement on the direction of an edge

- General inconsistency** Higher-order disagreements can be removed by finding **maximal acyclic subgraphs**

- Results**
- 

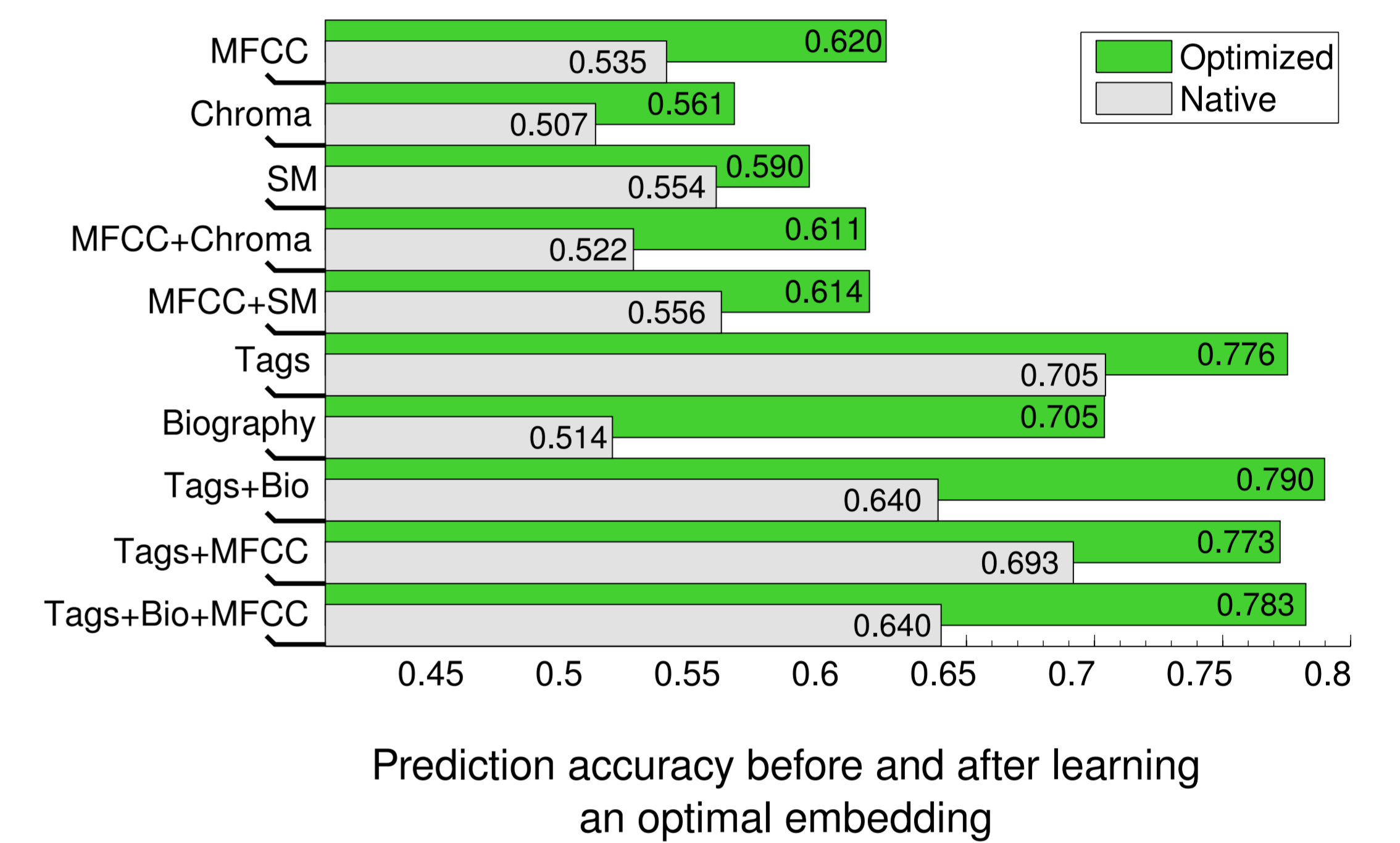
Input Features

- TEXT**
- Tags** 7737 tags from last.fm
TF-IDF cosine kernel 
 - Biography** 16753 words from artist biographies
TF-IDF cosine kernel
- ACOUSTIC**
- MFCC** 13 MFCCs + first and second derivatives
Modeled by Gaussian mixtures
Probability product kernel
 - Chroma** 12-d summary of pitch distribution
Modeled by full-covariance Gaussian
Symmetrized KL-divergence kernel
- Semantic Multinomial** Distribution over 149 auto-tags
Derived from MFCCs [3]
Probability product kernel

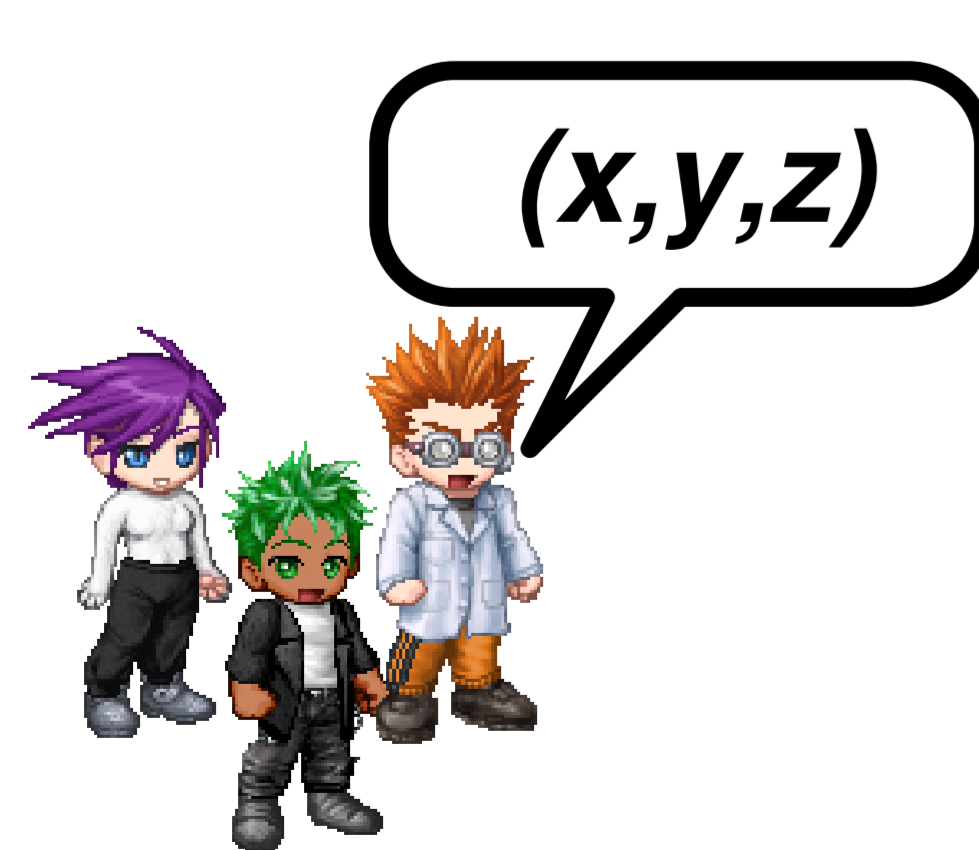
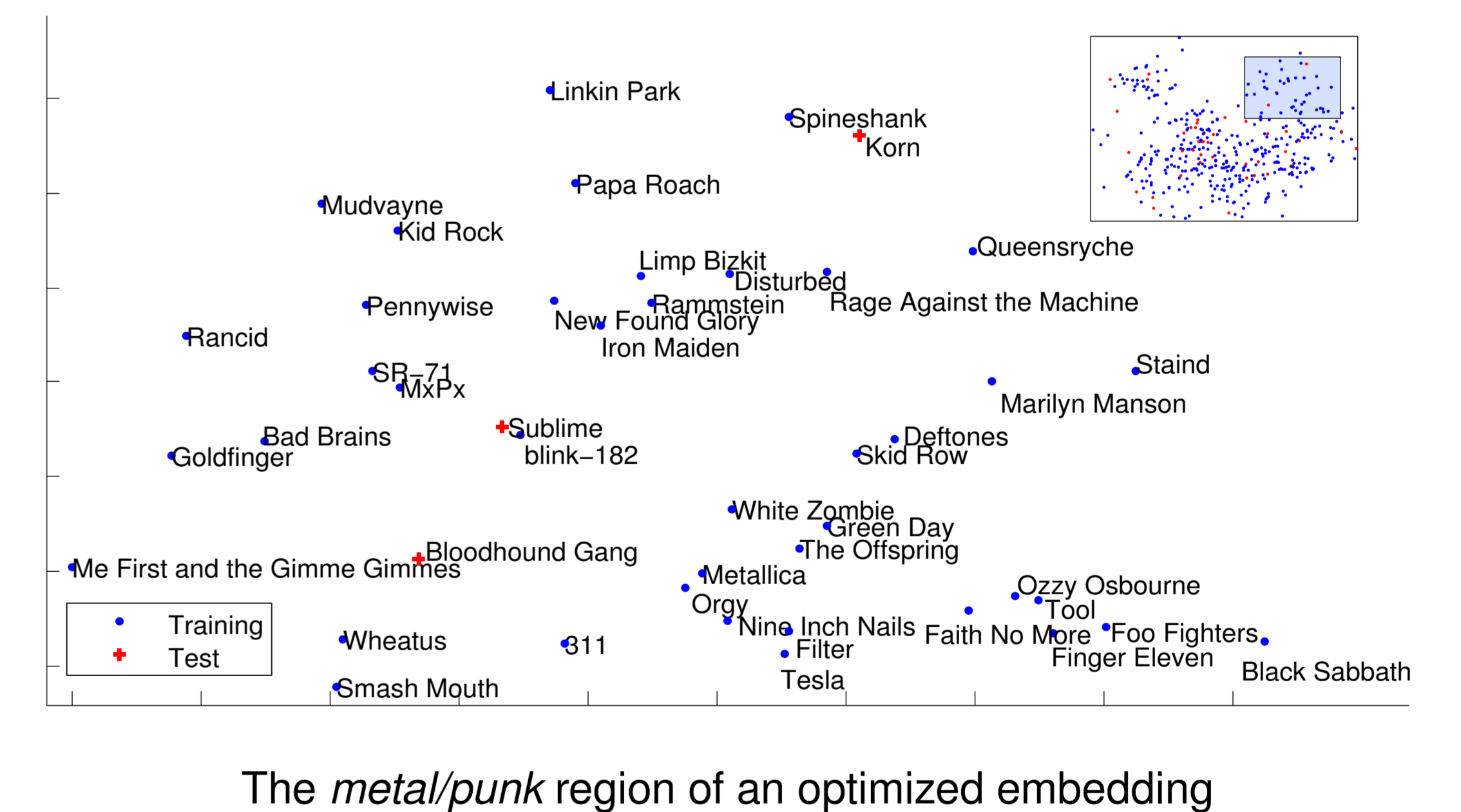
Similarity Prediction

- Evaluation** Direct inconsistencies are pruned and the data is split for 10-fold cross-validation.
Average training set size: 6252.7 edges
Average test set size: 1149.6 (x,y,z)
- Prediction task**
- Embed the training set (**learn projections**)
 - Map unseen artists into the space
 - Use distance to predict similarities (x,y,z) where x is unseen.
- Note that the test set has not been processed for internal consistency, so 100% accuracy is not possible.

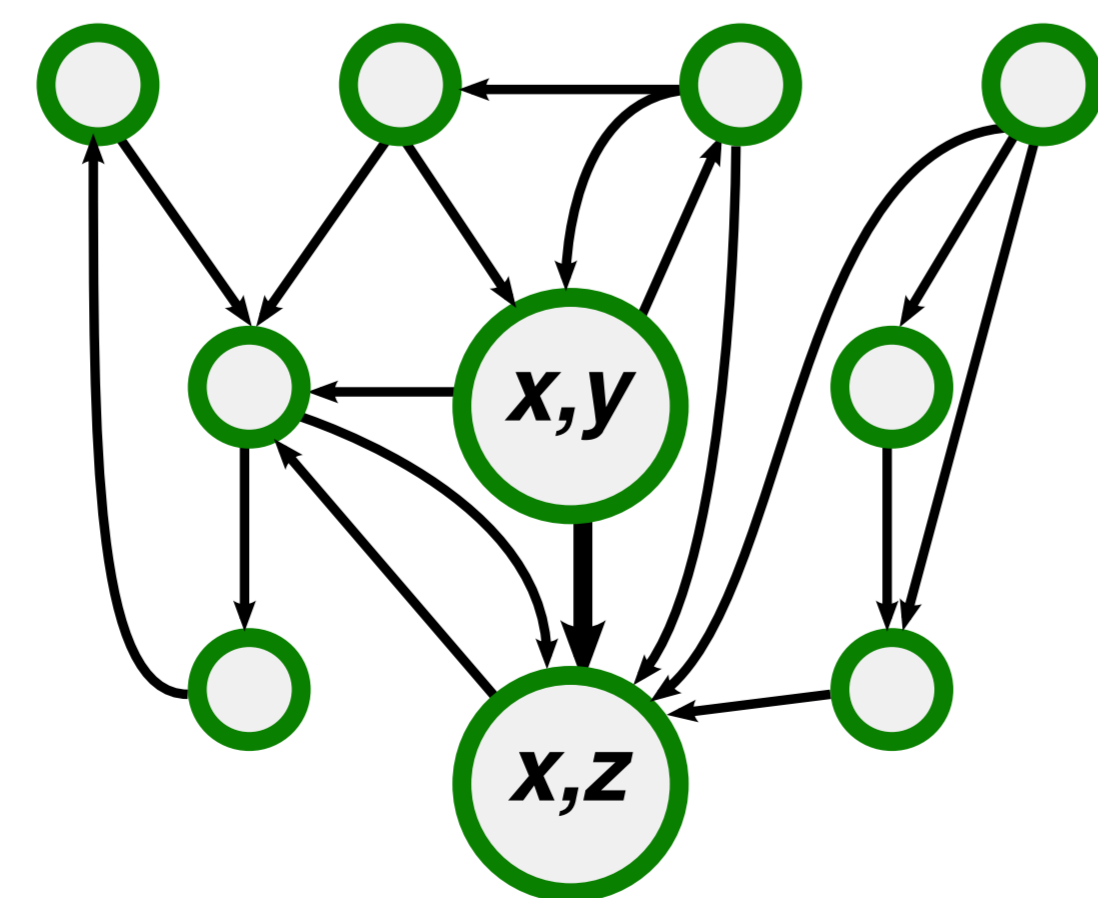
Results



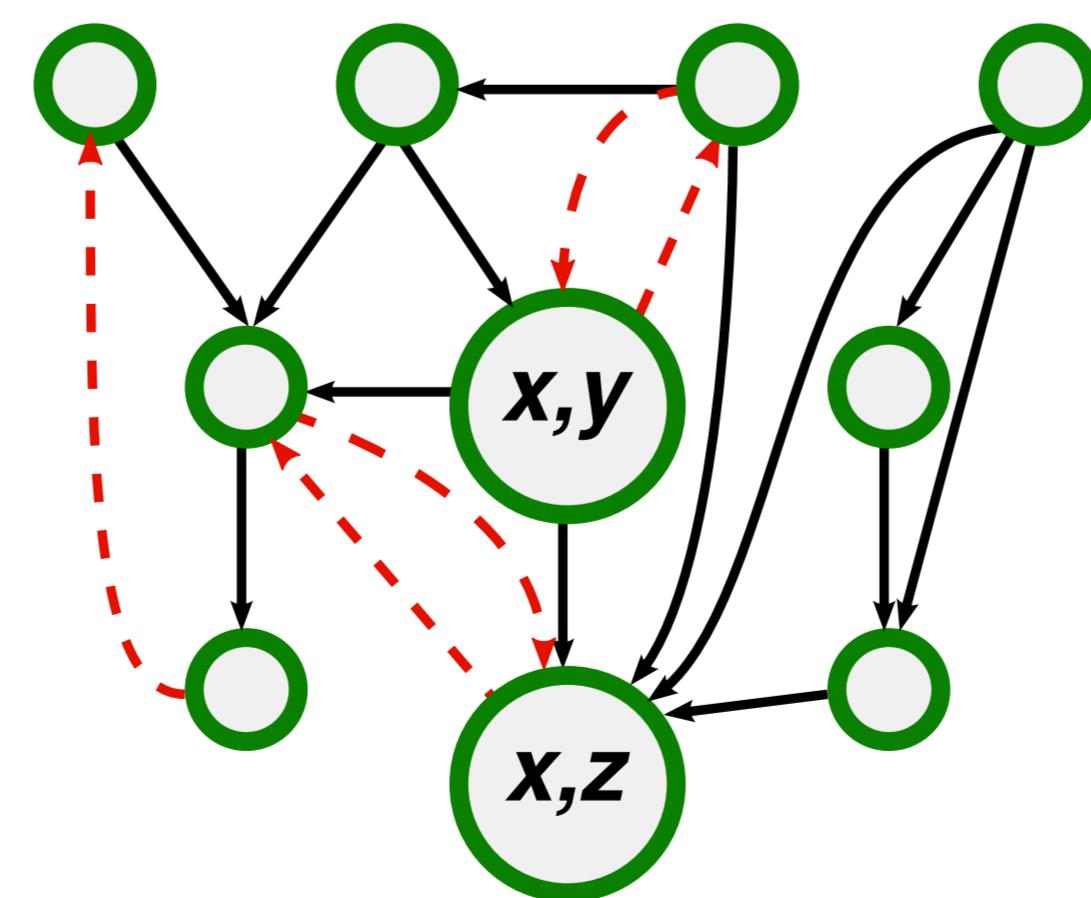
Example



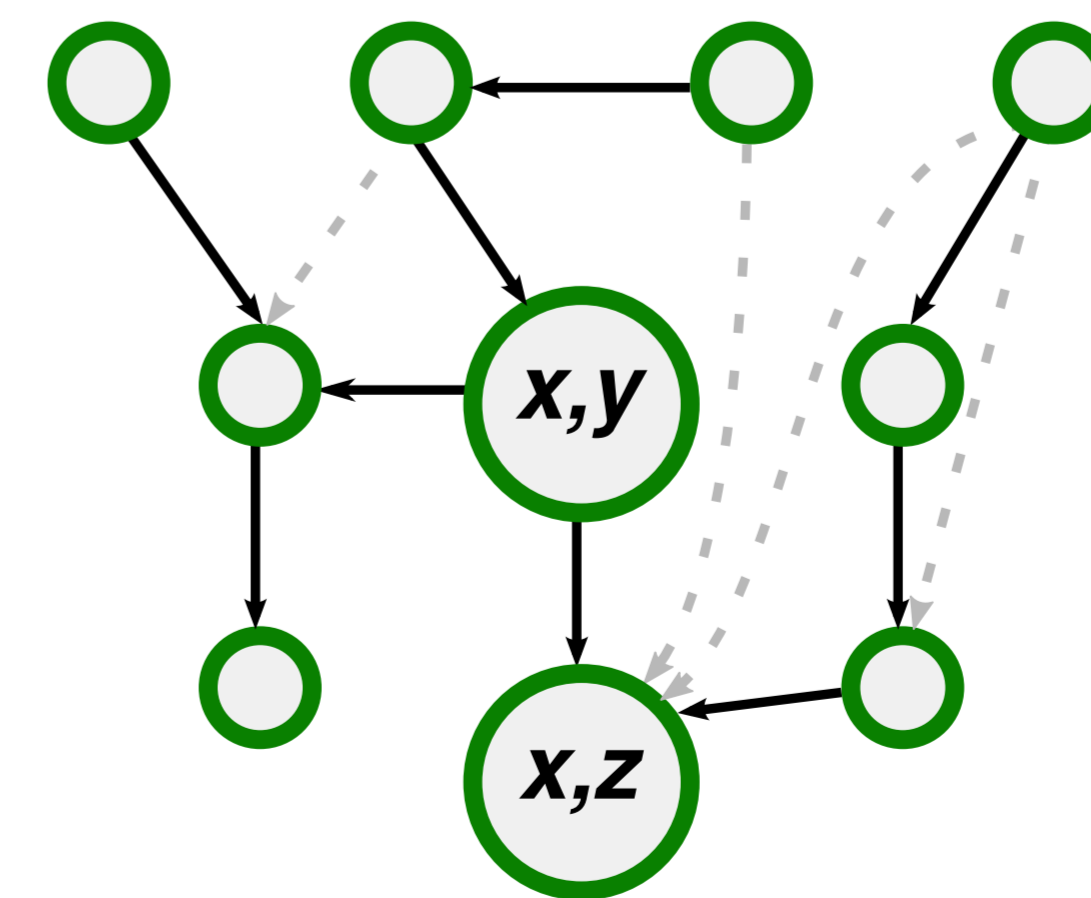
1. Humans supply similarity measurements



2. Similarity measurements are combined to form a **directed graph** over pairs



3. Inconsistencies are pruned to reveal a **directed acyclic graph (DAG)**



4. The DAG is simplified to a minimal equivalent edge set. These edges form the constraints of the optimization.

More similar

Less similar

References

- [1] Brian McFee and Gert Lanckriet. Partial order embedding with multiple kernels. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.
- [2] D. Ellis, B. Whitman, A. Berenzweig and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR2002)*, pp.170-177, 2002.
- [3] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):467-476, February 2008.