From Composition to Compositionality: Discovering Reusable Structure in Polyphonic Music Embeddings

Zhijin Guo [1 2] **Richard Freedman** [3] **Martha Lewis** [4]

Abstract

Composer, *composition*, *compositional*, *compositionality*: four words associated with the same Latin root *componere*, "to put together." We take this overlap literally and study Renaissance polyphony as a naturalistic testbed for compositional representation learning: composers build works by recombining reusable melodic units under constraints of composer, genre, and voice. We model polyphonic music as a multi-relational graph over interval n-grams, with sequential, vertical, and thematic relations capturing how musical units are combined melodically, contrapuntally, and motivically. Embeddings trained only from these relations, without attribute supervision, recover interpretable additive factors corresponding to externally documented stylistic categories. Held-out (composer, genre, voice) combinations are recoverable bidirectionally, and the real attribute–embedding correspondence outperforms 100 permutation baselines across reconstruction, cosine, and retrieval metrics. Finally, projecting out the genre subspace selectively reduces genre information while preserving much of the relational structure, suggesting that relational musical context can induce partially separable compositional representations.

1. Introduction

Composer, *composition*, *compositional*, *compositionality*: four words associated with the Latin root *componere*, "to put together" (Online Etymology Dictionary). The shared root is more than a linguistic curiosity. It names both a human generative act and a central goal of representation learning. A composer creates a composition by selecting, transforming, and recombining reusable musical materials under structural constraints (Freedman, 2020). A compositional representation, in the machine-learning sense, should likewise encode reusable factors that can be recombined to support systematic generalization (Lake & Baroni, 2018). This paper takes the overlap literally: we study whether embeddings learned from composed music recover the compositional structure by which the music was made.

This question is central to compositional learning. Modern AI systems can achieve strong in-distribution performance while still failing on cases that require recombining known components in unfamiliar ways (Lake et al., 2017; Andreas, 2019; Dziri et al., 2023). Such failures matter for robustness and safety, because models that rely on shallow correlations may behave unpredictably under distribution shift. They also matter for interpretability: if a model generalizes compositionally, its internal representation should expose reusable factors and combination rules rather than opaque memorised associations. The challenge is that, in many domains, it is hard to know whether the relevant components are genuinely present in the data or merely imposed by the benchmark.

Renaissance polyphony provides an unusually clear test case. In sixteenth-century imitative music, composers built new works by recombining short melodic units (*soggetti*) drawn from motets, madrigals, chansons, and chant materials, under constraints shaped by composer, genre, and voice (Freedman & Fiala, 2024). These factors are not labels invented for an Machine Learning evaluation; they are part of the documented compositional practice of the repertoire. The domain therefore lets us ask a precise interpretability question: can an unsupervised model recover, from relational musical context alone, the reusable stylistic factors that musicologists identify in the source process?

Polyphony is especially suitable because it is intrinsically relational. A melodic unit is shaped by what follows it in the same voice, what sounds simultaneously in other voices, and how similar material recurs elsewhere in the corpus. We encode these three dimensions as sequential, vertical, and thematic relations in a multi-relational graph over interval

[1]Department of Engineering Science, University of Oxford, Oxford, UK [2]Department of Computer Science, University of Bristol, Bristol, UK [3]Department of Music, Haverford College, Haverford, PA, USA [4]Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands. Correspondence to: Zhijin Guo <zhijin.guo@eng.ox.ac.uk>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

n-grams. These relations provide a natural inductive bias: they expose the different ways in which musical units are put together, without giving the model composer, genre, or voice labels during training.

We then test whether the learned embedding space is compositional: whether it contains interpretable reusable factors, whether those factors combine additively, and whether they support generalization to held-out combinations rather than memorisation of observed co-occurrences. Specifically, we ask whether embeddings trained only from relational structure recover externally documented factors of composer, genre, and voice, and whether held-out (composer, genre, voice) combinations can be reconstructed bidirectionally. In this way, composed music becomes a naturalistic testbed for studying the internal representation of compositionality.

Contributions. Our contributions are fourfold:

- **Compositional signals in a natural domain.** We study Renaissance polyphony as a setting where composers explicitly recombine melodic units under documented constraints of composer, genre, and voice. This gives us a naturalistic testbed in which compositionality is both the source process that generated the data and the representational property we aim to recover.
- **Relational evidence for composition.** We use sequential, vertical, and thematic relations as the only training signal. These relations capture how musical units are combined across melodic continuation, contrapuntal coordination, and motivic recurrence, allowing us to test whether compositional factors emerge from musical context alone.
- **Interpretable factor structure.** We show that embeddings learned without attribute supervision linearly align with externally documented factors of composer, genre, and voice. Moreover, these factors admit an additive decomposition that outperforms 100 permutation baselines across reconstruction and retrieval metrics.
- **Generalization and separability.** We show that held-out factor combinations are recoverable bidirectionally, from embeddings to attributes and from attributes to embeddings. We also find that genre information can be partially removed while preserving much of the learned relational structure, suggesting partially separable compositional signals.

2. Relational Context for Polyphonic Music

Language models are typically trained on text as a linear sequence, which tokens occur near which others, despite the underlying syntactic structure. Polyphonic music, by contrast, provides a richer substrate with multiple explicit

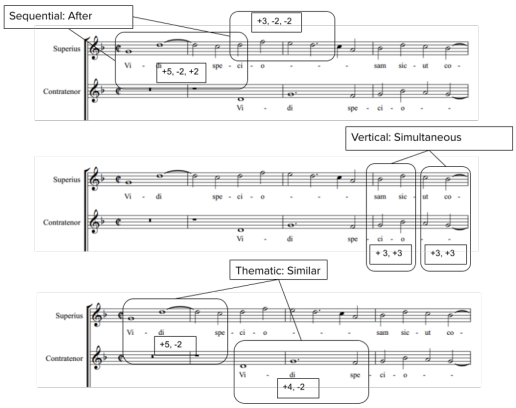


Figure 1. Three context types in polyphony. *Sequential*: patterns follow within a voice. *Thematic*: similar patterns recur elsewhere. *Vertical*: patterns sound simultaneously across voices.

dimensions along which units relate. A composer puts reusable units together along several axes at once, and our modelling choice is to expose these axes as distinct relational contexts. Thus, if the learned geometry shows compositional structure, it is recovering the same kind of "putting together" that generated the score.

Tokens: interval n-grams. We represent melodies as intervals between successive pitches rather than absolute pitches, so that transpositions such as C–D–F–E and G–A–C–B map to the same pattern, $(+1, +2, -1)$. We use length-5 diatonic interval n-grams from the CRIM corpus (Freedman & Fiala, 2024), yielding 16,703 unique n-grams across approximately 250,000 occurrences in about 250 sixteenth-century polyphonic compositions. Processing details are in Sections B and C.

Three context types. We build a multi-relational graph over n-grams using three relations (Figure 1). The *sequential* relation connects patterns that follow one another within a voice. The *vertical* relation connects patterns that sound simultaneously across voices, capturing contrapuntal constraint. The *thematic* relation connects melodically similar patterns elsewhere in the corpus, capturing motivic recurrence. Together these relations encode how musical units are combined melodically, contrapuntally, and thematically.

Learning without attribute supervision. We learn an embedding $\mathbf{e}_i \in \mathbb{R}^d$ for each n-gram from this graph using a contrastive knowledge-graph embedding objective (Yang et al., 2015; Bordes et al., 2013); details are in Section D. Composer, genre, and voice labels are never used in training. The only signal is which n-grams co-occur under which relation, so any stylistic structure recovered by the embeddings must be induced from relational context alone. Link-prediction Hits@1 is 0.77 against a random baseline

of 0.33, confirming that the relational structure is encoded.

3. Evaluating Linearity and Compositional Generalization

We adapt the two-step protocol of Guo et al. (2025) for quantifying compositionality in classical and state-of-the-art embeddings to the music domain. The protocol takes embeddings learned without attribute supervision and asks two things of them: first, whether known categorical attributes are linearly recoverable from the embedding geometry (a representation question); second, whether held-out attribute combinations can be reconstructed by composing factor vectors (a generalization question). Applying it to relational music embeddings tests whether the additive compositionality canonically observed in word embeddings emerges from a richer, multi-relational context.

Attributes. Each n-gram i is associated with a (composer, genre, voice) attribute combination via Boolean indicators, providing 131 distinct attribute groups across the corpus. These categories are defined by musicological practice and external to training; no attribute information was used to learn the embeddings.

Step 1: linearity. We seek a factor matrix $\hat{\mathbf{U}}$ such that the average n-gram embedding within each attribute group is approximately reconstructed by a linear combination of factor vectors corresponding to that group's composer, genre, and voice. We measure linear-system loss (ℓ_2), cosine similarity between reconstructed and true group-mean embeddings, and identity retrieval@10 over the 131 candidate groups. Significance is established against a permutation null: 100 random shuffles of the attribute–n-gram correspondence, under which no relationship between attributes and embeddings exists by construction.

Step 2: generalization. The same factor matrix $\hat{\mathbf{U}}$ is then used to recover held-out attribute combinations bidirectionally, from embedding to attributes and from attributes to embedding, testing whether factor vectors learned aggregately across the corpus combine correctly to handle configurations the linear system was not directly fit to reconstruct.

4. Results

4.1. Attribute Factors Are Linearly Reflected in the Embedding

We first ask whether externally documented stylistic factors are linearly reflected in the learned embedding space. If the relational objective has recovered composer, genre, and voice structure, then attribute indicators should align with directions in the embedding geometry, despite never being used during training.

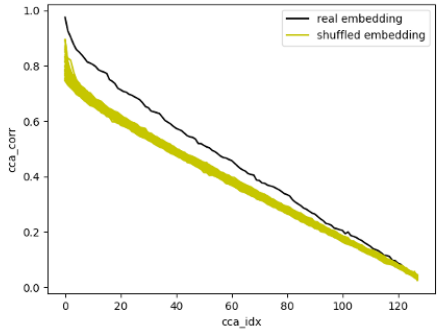


Figure 2. Canonical correlation between attribute indicators and n-gram embeddings, sorted by component. The real embedding (black) yields higher canonical correlations across all components than any of 100 random shuffles of the attribute–embedding correspondence (yellow band).

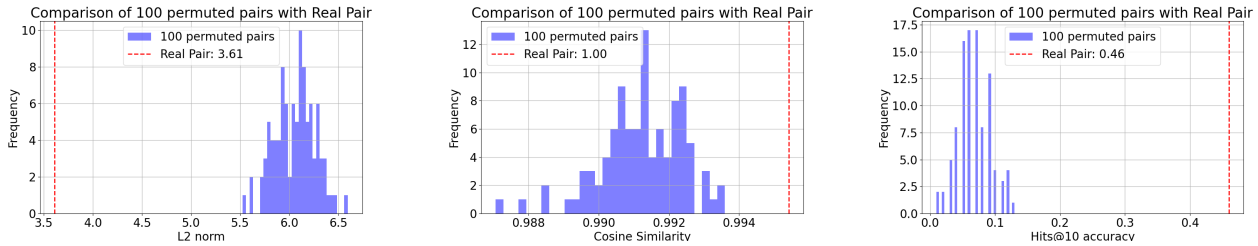
We test this with CCA between attribute indicators and n-gram embeddings. Figure 2 shows that the real attribute–embedding pairing provides higher canonical correlations than 100 random shuffles, indicating that composer, genre, and voice are specifically and linearly present in the embedding space.

4.2. Embeddings Admit an Additive Factor Decomposition

Linearity alone shows that stylistic information is present, but not that it composes. We therefore test whether group-mean embeddings can be reconstructed by adding factor vectors for composer, genre, and voice: $\mathbf{e}_i \approx \mathbf{e}^{\text{composer}}_{c(i)} + \mathbf{e}^{\text{genre}}_{g(i)} + \mathbf{e}^{\text{voice}}_{v(i)} + \mathbf{r}_i$, where $\mathbf{r}_i$ is a residual capturing group-specific variation not explained by composer, genre, or voice. We evaluate reconstruction using linear-system loss (ℓ_2), cosine similarity, and identity retrieval@10 over 131 attribute groups, comparing against 100 random permutations of the attribute–embedding correspondence.

Figure 3 shows that the real embedding lies outside the permutation distribution on all metrics. The linear-system loss is lower than the most favourable random shuffle (3.61 vs. 5.5), cosine similarity exceeds the best permuted value, and retrieval@10 reaches 0.46 compared with a maximum permuted value of 0.13. Thus, the embedding does not merely contain stylistic information; it organizes composer, genre, and voice as additive factors.

This result is important because the factors are external to training. The relational objective never observes composer, genre, or voice labels, yet the learned geometry places musicologically defined categories at locations that combine linearly. In this sense, the embedding recovers a composi-



(a) Linear-decomposition loss (ℓ_2). *(b)* Cosine similarity. *(c)* Identity retrieval@10.

Figure 3. Distribution of reconstruction statistics under 100 random permutations of the attribute–embedding correspondence, with the real embedding marked. The real result lies outside the permutation distribution on all three metrics.

tional structure from relational context alone.

4.3. Additive Factors Generalize to Held-Out Combinations

Additive structure is useful only if it supports recombination. We therefore test whether factor vectors learned from observed groups can recover held-out (composer, genre, voice) combinations. The corpus is sparse: many composers appear with only some genres, and not all voices appear in all composer–genre pairs. Recovering a held-out group therefore requires composing factors learned from other configurations.

We evaluate this bidirectionally. In the *embedding-to-attribute* direction, a held-out group-mean embedding is used to identify the correct composer, genre, and voice combination among 131 candidates. In the *attribute-to-embedding* direction, a held-out attribute combination is used to synthesize an embedding near the correct empirical group mean. Retrieval@10 reaches 0.46, compared with a permutation-null maximum of 0.13. The bidirectional test distinguishes compositional structure from probing. Recovering attributes from an embedding could mean only that the embedding contains label information. Synthesizing an embedding from attributes is stronger: it requires the geometry to support construction by vector addition. The success of both directions indicates that the learned factors can be recombined, rather than merely decoded.

4.4. Partial Genre Separability

Finally, we ask whether additive factors are functionally separable. We remove the genre subspace from the embedding using nullspace projection at threshold $\theta = 0.2$ and measure both genre classification and structural link prediction.

Table 1 shows that genre AUCs move toward the chance baseline of 0.50 after projection, while link prediction declines only moderately and remains well above its random baseline of 0.33. Genre information is therefore partially separable: removing it selectively reduces genre predictability without destroying the relational structure needed for link prediction.

Table 1. Genre classification AUC before and after projecting genre out of the embedding. Hits@1 reports link-prediction accuracy on held-out edges.

Genre / Metric	Before	After
Agnus Dei	0.59	0.56
Chanson	0.72	0.59
Credo	0.57	0.55
Gloria	0.53	0.52
Kyrie	0.54	0.52
Madrigal	0.57	0.53
Motet	0.55	0.53
Plainchant	0.57	0.51
Sanctus	0.57	0.54
Hits@1 (link pred.)	**0.77**	**0.67**

5. Conclusion and Future Work

We studied Renaissance polyphony as a naturalistic testbed where compositionality is both the source process that generated the data and the representational property being tested. From only sequential, vertical, and thematic relations between interval n-grams, embeddings recover externally documented factors of composer, genre, and voice without attribute supervision.

The results show three compositional signals: stylistic attributes are linearly reflected in the embedding space, group embeddings decompose additively into composer, genre, and voice factors, and held-out factor combinations are recoverable bidirectionally. The genre-removal experiment further suggests partial separability, since genre information can be reduced while much of the relational structure remains. Future work should ablate the three relation types and test stronger forms of transfer, such as holding out entire composers or genres. Additionally, scaling this framework to larger corpora represents a natural next step. This could involve incorporating related 16th-century repertoires—such as those curated by The Josquin

Research Project, The 1520s Project, or The Tasso in Music Project—or expanding chronologically to encompass 17th- and 18th-century polyphonic styles. Finally, while the current model successfully captures stylistic signatures using purely sequential and simultaneous relations, it notably omits explicit duration data. Given the results, the topological and chronological order of interval patterns alone carries immense stylistic information; however, explicitly integrating note and pattern durations remains a critical avenue for future investigation.

Beyond these directions, structural network-analysis methods could complement our geometric study. Spectral and partition-based analyses of signed networks (Guo et al., 2026) could be applied by casting contrapuntal relations (e.g., consonance vs. dissonance) as signed edges, probing how voices and stylistic groups cohere or oppose within the learned space.

Acknowledgements

This work was supported by the University of Bristol's Jean Golding Institute Seed Corn Funding. The authors would like to warmly thank Nello Cristianini for providing invaluable guidance and discussing the foundational ideas of this paper, as well as Chenhao Xue for insightful discussions that helped shape this work.

Impact Statement

This paper aims to advance the understanding of compositional representation learning. We do not foresee that the musicological setting introduces consequences requiring dedicated discussion beyond those standard in the field.

References

Andreas, J. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, 2013.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI Conference on Artificial Intelligence*, 2018.

Cuthbert, M. S. and Ariza, C. music21: A toolkit for computer-aided musicology and symbolic music data. In *International Society for Music Information Retrieval Conference*, 2010.

Dziri, N., Lu, X., Sclar, M., et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, 2023.

Ethayarajh, K., Duvenaud, D., and Hirst, G. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Freedman, R. Citations: The renaissance imitation mass. *Le Studium Multidisciplinary Journal*, 2020. URL https://hal.science/hal-03219135.

Freedman, R. and Fiala, D. Citations: The Renaissance Imitation Mass Project (CRIM). https://crimproject.org/, 2024. Project directors. Accessed 2026-05-07.

Guo, Z., Xue, C., Xu, Z., Bo, H., Ye, Y., Pierrehumbert, J. B., and Lewis, M. Quantifying compositionality of classic and state-of-the-art embeddings. *arXiv preprint arXiv:2509.19332*, 2025.

Guo, Z., Zhang, L., Bonnet, T., Pierrehumbert, J. B., and Dong, X. Linking extreme discourse to structural polarization in signed interaction networks. *arXiv preprint arXiv:2605.12814*, 2026.

Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2018.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. A review of relational machine learning for knowledge graphs. In *Proceedings of the IEEE*, volume 104, pp. 11–33, 2016.

Online Etymology Dictionary. Compose: Etymology, origin and meaning. https://www.etymonline.com/word/compose. Accessed 2026-05-07.

Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations*, 2015.

A. Related Work

Compositional structure in distributional embeddings. The observation that word embeddings support vector arithmetic over latent factors (Mikolov et al., 2013) has been variously interpreted as a curiosity of co-occurrence statistics (Ethayarajh et al., 2019), evidence of geometric structure that approximates linguistic relations, or a benchmark for compositional generalization in its own right. Ethayarajh et al. (2019) provide a theoretical account showing that linear analogy structure follows from co-occurrence shifted-PMI when the relevant factors are statistically independent in the corpus, a condition that holds approximately for many semantic relations in text but is hard to verify externally. Guo et al. (2025) formalise this as a two-step generalised evaluation: first measure linearity between known entity attributes and embeddings via canonical correlation analysis, then evaluate additive generalisation by reconstructing embeddings for unseen attribute combinations under reconstruction-loss, cosine, and retrieval metrics. We adopt this protocol and apply it to a relational embedding of polyphonic music. Our setting extends the line of work along two axes. First, we operate over a richer relational substrate: three distinct relation types rather than one, generated by partially independent compositional rules in the source domain. Second, the latent factors being recovered are independently documented stylistic categories from musicology rather than implicit semantic axes inferred from the embedding behaviour, which provides external validation of what the geometry has learned.

Compositional generalization. A separate literature, exemplified by SCAN and the broader systematic-generalization agenda (Lake et al., 2017; Andreas, 2019), asks whether models recombine seen primitives to handle unseen combinations. Most work in this area uses downstream task performance as the test signal: a model is trained on commands like "jump twice" and "walk thrice" and tested on "jump thrice." Our test is geometric rather than task-based: we ask whether held-out factor combinations are recoverable from the embedding space itself, in both directions (from embeddings to attributes and from attributes to embeddings). The two formulations are complementary: task-based tests reveal whether a model can act compositionally, while geometric tests reveal whether its representations are themselves composed in a way that supports the action. Andreas (2019) introduces the closely related notion of representational compositionality, which we test directly on a domain with documented compositional generative structure.

Disentanglement. Unsupervised disentanglement is provably impossible without inductive biases (Locatello et al., 2019). Standard approaches supply such bias through architectural or objective choices, or through grouping signals (Bouchacourt et al., 2018). Our setting supplies an implicit inductive bias through the structure of the relational data itself: the three relation types are governed by partially independent generative rules in the source domain. The question of whether this kind of relational diversity is sufficient to induce disentangled representations is, to our knowledge, underexplored, and the present work provides one positive data point. The factors we recover (composer, genre, voice) are partially separable in the sense of Section 4.4: removing one degrades classification on it without destroying structural prediction. This is a weaker form of disentanglement than the strict factor-orthogonality posited in some of the disentanglement literature, but it is a form that emerges naturally from relational training without targeted objectives.

Knowledge-graph embeddings. We use DistMult (Yang et al., 2015) as a representative of the knowledge-graph embedding family (Bordes et al., 2013; Nickel et al., 2016). Our framing departs from the standard application of these methods (link prediction over factual relations) by treating the embedding itself as the object of study and link prediction as a sanity check on representation quality. The choice of DistMult is conservative: it is the simplest bilinear scorer that supports multiple relation types, and any compositional structure we observe is therefore not a function of expressive bias in the scoring function.

Symbolic music modelling. Symbolic music has a long history of representation-learning approaches (Cuthbert & Ariza, 2010), but most prior work treats the score as a se-

quence and applies sequential models accordingly. The present approach, by contrast, models music as an explicitly relational object with three distinct edge types. This is closer in spirit to musicological practice (which treats counterpoint, motivic working, and harmonic constraint as separate analytic dimensions) than to standard audio or symbolic music ML.

B. Musicological Background

The data considered in this paper come from musical works of nearly 500 years ago: a corpus of polyphonic vocal music in which four, five, or more independent lines are combined in a contrapuntal weave, making sense both individually and as a whole. This is contrapuntal music, tightly bound by a set of stylistic rules; later centuries called it the "strict" style, and it was very carefully composed. Thousands of such pieces were composed and printed in the sixteenth century. We focus on a particular sacred genre: a cyclic treatment of the five Ordinary (fixed-text) movements of the Catholic Mass: Kyrie, Gloria, Credo, Sanctus, and Agnus Dei. Many such pieces were based on still other polyphonic works, variously sacred (motets) or secular (madrigals, French chansons), and were filled with complex allusions, citations, and transformations.

In Latin such pieces were called *missa ad imitationem condita*, an interesting turn of phrase that is worth considering for its original meanings and modern resonances. *Ad imitationem* means simply "in imitation." But *condita* has special meanings. It is the past participle of *condere*, and probably had a number of nuanced connotations for sixteenth-century readers. According to the French humanist Robert Estienne's famous trilingual dictionary of Latin, French, and English (issued in London in 1552), the infinitive could mean "to compose, to make, to build, or to edify."[1] But it might also mean "to enclose," or even "to hide." Such associations encourage us to think of these pieces as things not only *built upon* their models, but structures in which borrowed patterns were at times *hidden* and *enclosed*. They were filled, in short, with forms and meanings that can be both obvious and covert, which is one reason such pieces are particularly relevant to the theme of compositionality we consider here: learned representation spaces are themselves both built up from training signal and filled with latent meanings recoverable only through analysis.

C. Symbolic Score Processing

The data are derived not from audio sources but from symbolic scores. The original sixteenth-century sources are first transcribed into modern notation, with all parts on a single page. From these we produce logical XML encodings according to the open-source MEI schema.[2] Using the Python libraries Music21 (Cuthbert & Ariza, 2010)[3] and CRIM Intervals,[4] we then derive recurring patterns of notes in various dimensions. These come in the form of *intervals*: vectors representing not the notes themselves but the distances between them, measured against a standard diatonic scale of pitches (as on a keyboard). Sequential patterns are like standard n-grams, expressed as patterns such as $(-3, +2, -3)$ followed by $(+2, -3, +4)$. Vertical patterns are those formed between tones occurring simultaneously in different voice parts. Thematic patterns are those similar to one another occurring elsewhere in the corpus. Figure 1 in the main text illustrates all three on a sample passage.

The three relations together form what Freedman & Fiala (2024) call a "double helix" threading the corpus: sequential threading each voice as a chain of motivic units, vertical binding voices to one another under contrapuntal constraints, and thematic linking recurrent material across the entire corpus regardless of voice or temporal position.

D. Embedding Training Details

We learn embeddings $\mathbf{e}_h, \mathbf{e}_t \in \mathbb{R}^d$ for each n-gram and a relation matrix M_r for each relation $r \in \{\text{seq}, \text{vert}, \text{them}\}$ via the DistMult scoring function $S(h, r, t) = \mathbf{e}_h^\top M_r \mathbf{e}_t$ (Yang et al., 2015). We optimise the contrastive negative-sampling loss

$$\mathcal{L} = -\sum_{f \in F} \log \frac{e^{S(f)}}{e^{S(f)} + \sum_{f' \in F'} e^{S(f')}}, \tag{1}$$

where F are observed (head, relation, tail) triples and F' are corrupted negatives. Training uses learning rate 0.1 for 50 epochs. The choice of DistMult is conservative: it is the simplest bilinear scorer that supports multiple relation types, and any compositional structure observed is therefore not a function of expressive bias in the scoring function.

E. Permutation-Test Summary Table

F. Concrete Example of Bidirectional Recovery

To make the bidirectional recovery concrete, consider the group (Palestrina, Kyrie, Tenor). The associated factor vectors are $\mathbf{e}^{\text{composer}}_{\text{Palestrina}}$, $\mathbf{e}^{\text{genre}}_{\text{Kyrie}}$, and $\mathbf{e}^{\text{voice}}_{\text{Tenor}}$, each fit jointly with its respective category over the entire corpus. The Palestrina vector is shaped by every Palestrina n-gram across all his Mass movements (Kyrie, Gloria, etc.) and across all voices. The Kyrie vector is shaped by every composer's Kyrie move-

[1] From the *Dictionariolum puerorum, tribus linguis Latina, Anglica & Gallica* (Paris 1553), p. 72.

[2] https://music-encoding.org/

[3] https://github.com/cuthbertlab/music21

[4] https://github.com/HCDigitalScholarship/intervals

Table 2. Linear decomposition of n-gram embeddings into stylistic factors. Lower loss, higher cosine similarity, and higher retrieval indicate stronger additive structure. Permutation results report the range over 100 random shuffles of the attribute–embedding correspondence. The real embedding lies entirely outside the permutation distribution on all three statistics.

METRIC	REAL	PERMUTED (RANGE)
LINEAR LOSS (↓)	**3.61**	5.5 TO 6.6
COSINE SIMILARITY (↑)	**1.00**	0.988 TO 0.994
RETRIEVAL@10 (↑)	**0.46**	0.01 TO 0.13

ments. The Tenor vector is shaped by every Tenor line in the corpus, regardless of composer or genre. None of these factor vectors is fit specifically on (Palestrina, Kyrie, Tenor) co-occurrences. Yet their additive combination lands close enough to the empirical (Palestrina, Kyrie, Tenor) group mean that retrieval@10 succeeds, and does so symmetrically in the inverse direction, where the empirical group mean is presented to the linear system and the correct attribute combination is recovered. This is not a probe; it is a generalization test in the strict sense.

G. What Additive Structure Means Formally

Section 4.2 is approximate, and the residual $\mathbf{r}_i$ is non-negligible: it captures motivic content shared across composers, genres, and voices. What we claim is that the projection of $\mathbf{e}_i$ onto the stylistic-factor subspace can be expressed as the sum of three factor vectors with low residual, where "low" is operationalised by the permutation tests of Table 2. This is weaker than perfect linear factorization but stronger than mere correlation: the addition operation, applied to factor vectors learned independently of one another, recovers stylistic-group means at rates inconsistent with the null hypothesis of no relationship. Ethayarajh et al. (2019) provide a theoretical condition under which this kind of structure emerges in word embeddings, namely that the latent factors are statistically independent in the corpus. The condition is approximately satisfied here: composer, genre, and voice are not perfectly independent in CRIM (some composers favour certain genres), but they are sufficiently independent that the relational learning objective can place them at near-orthogonal directions in the embedding space.

H. Comparison to Vector Arithmetic in Word Embeddings

The closest analog in the literature is the analogy task "a is to b as c is to ?" in word embeddings, evaluated over held-out analogy quadruples (Mikolov et al., 2013; Ethayarajh et al., 2019). Our setup differs in two respects. First, we operate over factor combinations rather than word quadruples, which makes the compositional claim more general: any combination of factors, not just analogies of fixed form. Second, our domain has documented compositional structure: the additive geometry recovers categories that musicologists already use, which gives the result a level of external validation that the word-vector setting cannot easily provide.