

Supplementary Material for “Where does Haydn end and Mozart begin? Composer classification of string quartets”

All plots and analyses here can be found in the publicly accessible Github repository at <https://github.com/wongswk/haydn-mozart>. Implementation of heatmaps are through the R package *pheatmap* (Kolde, 2012).

This supplement contains the following material, as referenced in the main text.

- S1. Exploring Similarity among Segment Features (referenced in Section 3.2)
- S2. Investigating Dependence among Movements by Set, Quartet, and Date of Composition (referenced in Section 5.2)
- S3. Comparing LOO on the Full Feature Set versus the Reduced Feature Set (referenced in Section 5.2)

S1. Exploring Similarity among Segment Features

As noted in Section 3.2, we have performed exploratory analyses to determine appropriate candidates for segment lengths of m notes for the segment features. We compute all segment features for segment lengths $m = 8, 9, 10, \dots, 17, 18$. For various subsets of these segment features, we then compute correlation matrices and visualize them via heatmaps. If X is an $n \times p$ data matrix consisting of n movements and p standardized (centered and scaled) features, then the correlation matrix is computed as $X'X$; for $i, j = 1, 2, \dots, p$, the entry $X'X_{ij}$ is the (Pearson sample) correlation between features i and j over the n movements. From the heatmaps, we can discern that in most cases, features with segment lengths of distance 1 apart (e.g., segment lengths 8 and 9) have extremely high correlations of 0.9 or more. Meanwhile, when the segment lengths differ by 2, the correlations are weaker and in some cases lower than 0.5. Therefore, different information is captured for some features with segment lengths of distance 2 or more apart, while mostly the same information is reflected by segment lengths of distance 1 apart. Meanwhile, the number of proposed features would be increased by about 62% (from 1182 to 1914) by including the segment features with segment lengths of distance 1 apart in the feature set. In particular, a dimensionality approaching 2000 features, relative to a sample size of 107 or 285, would potentially require more specialized feature selection or dimension reduction strategies. Thus, as a balance between maximizing statistical information and minimizing computational and modeling complexity, we find it reasonable to include only segment length $m = 8, 10, 12, 14, 16, 18$ as candidates for segment features in our feature set and use feature selection (Section 4) to more rigorously identify the optimal segment lengths.

As an example, Figure S1 shows the correlation matrix for the recapitulation location features for duration in the HM285 dataset. The yellow diagonal represents correlations of 1, since each feature is perfectly correlated with itself. The four blocks forming around the diagonal correspond to each part of the string quartet, since, e.g., Viola features are correlated among themselves. Within each block, it is observed that correlations between pairs of segment features decrease as the distance between segment lengths increases. In particular, pairs of recapitulation segment features with segment lengths of distance 1 apart (e.g., $m = 8$ and $m = 9$) tend to have very strong correlations of 0.9 or higher; pairs of segment features with segment lengths of distance 2 apart (e.g., $m = 8$ vs. $m = 10$) have slightly weaker correlations around 0.8 or lower. Heatmaps for the remaining subsets of features can be found in the Github repository.

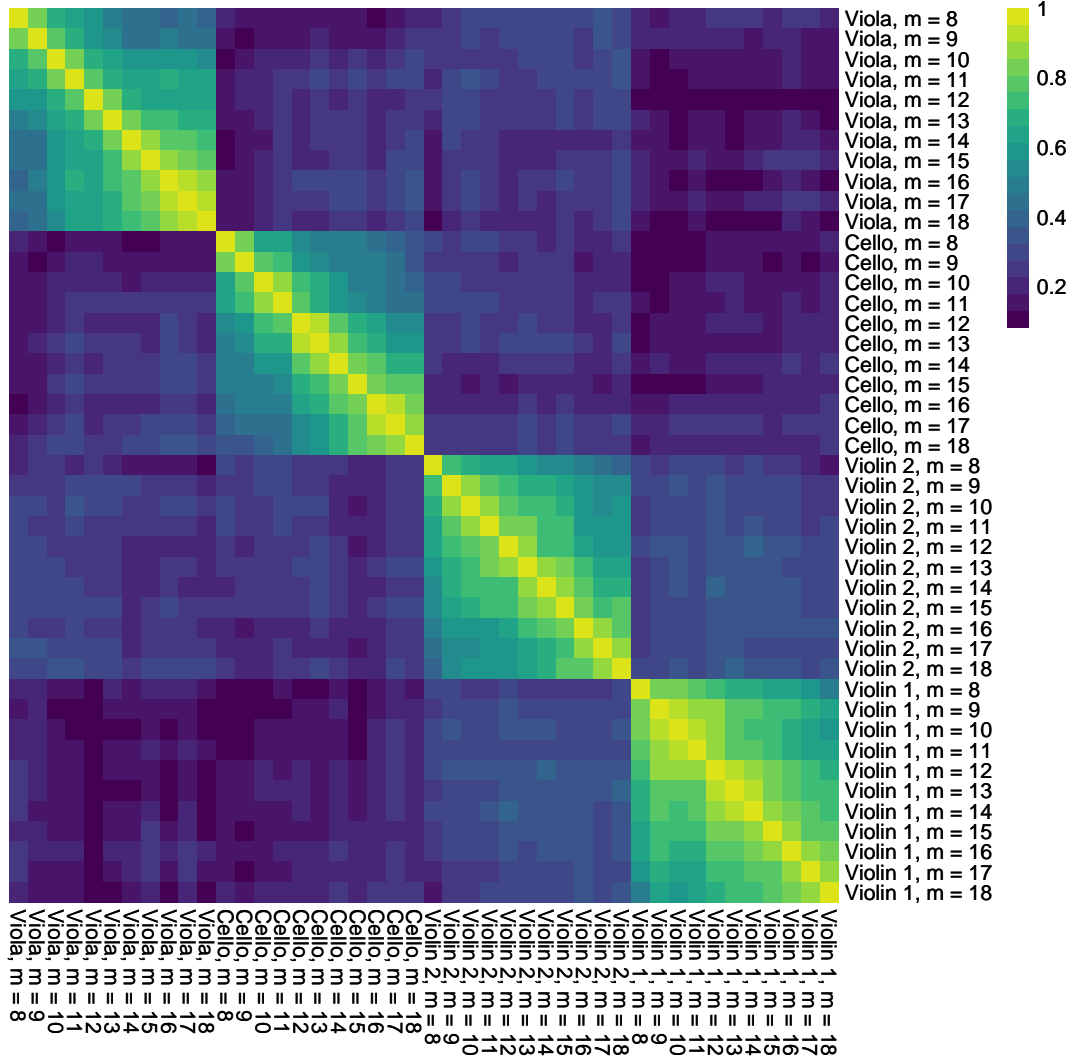


Figure S1.: The heatmap shows the $p \times p$ correlation matrix $X'X$ for the recapitulation location duration features (for segment lengths of $m = 8, 9, 10, \dots, 17, 18$ notes) in the HM285 dataset.

S2. Investigating Dependence among Movements by Set, Quartet, and Date of Composition

As mentioned in Section 5.2, there are natural groupings among Haydn and Mozart string quartet movements that could potentially violate the assumption of independence required for validity of LOO CV. In addition to the potential within-quartet dependency, these composers often published quartets as sets, especially when commissioned, e.g., Haydn’s ‘Prussian’ quartets dedicated to King Frederick William II of Prussia or Mozart’s No. 20 quartets (K. 499) published by Franz Anton Hoffmeister. More broadly, there could be temporal structure, since pieces are written over the course of a composer’s life.

We explore potential dependencies among movements by set, quartet, and date of composition through descriptive statistics: we visualize matrices of features and their correlations through heatmaps. Let $X = \begin{pmatrix} X_M \\ X_H \end{pmatrix}$ be the $n \times p$ data matrix consisting of p standardized features (centered and scaled) for n movements, ordered such that the first rows form a block X_M of Mozart movements and the final rows form a block X_H of Haydn movements. Within each composer block, we order each set of quartets by date of composition, then by quartet number, and finally by movement number. For example, within the ‘Milanese’ set of the Mozart block, the K. 155 quartets (movements 1, 2, then 3) would be followed by the K. 156 quartets. Define also the Gram matrix as XX' , which has XX'_{ij} as the (Pearson sample) correlation between movement i and movement j over the p standardized features, for all $i, j = 1, \dots, n$. We analyse the data matrix X , the Gram matrix XX' , and their variants, discussed as follows.

The heatmap corresponding to the data matrix X is displayed in Figure S2. Hierarchical clustering is then applied to X to help identify potential patterns among movements and features. We opt for the standard choices of complete-linkage and Euclidean distance for clustering, resulting in the heatmap in Figure S3. If dependencies within sets, quartets, composer category, or date were present, the following patterns would be observed. In Figure S2, block patterns over the heatmap rows would form. In Figure S3, clusters corresponding to sets, quartets, composer, or date would form, which would be indicated through the column annotations. However, no such patterns emerge. We note that clustering does result in the formation of some other strong clusters (e.g., the bright yellow blocks at the bottom of the heatmap), but they appear unrelated to sets, quartet, date, or composer. This is an interesting result, because it suggests movements that could be musically similar, despite belonging to different groups. The same analysis is repeated for various subsets of movements and features, e.g., the ‘recapitulation features’ and the ‘recapitulation features for the Mozart block X_M ’; none of the resulting heatmaps (which can be found in the Github repository) show patterns by set, quartet, or date.

The Gram matrix is visualized through the heatmap in Figure S4. Note that the colors associated with quartets have been randomized here (rather than following the order from Figures S2 and S3), for improved visualization. Similarly to the correlation matrix shown in Figure S1, the yellow diagonal corresponds to correlations of 1 between each movement and itself. Dependencies within sets, quartets, date, or composer would result in block-diagonal patterns on this heatmap; e.g., if the ‘Milanese’ quartets had correlated features, then a square block of yellow colors would be visible along the diagonal. While we do observe a few blocks of yellow colors (e.g., in the upper left corner of the heatmap), the corresponding correlation values are quite weak at 0.2 to 0.4; additionally, their rectangular, non-square shapes indicate that they span multiple

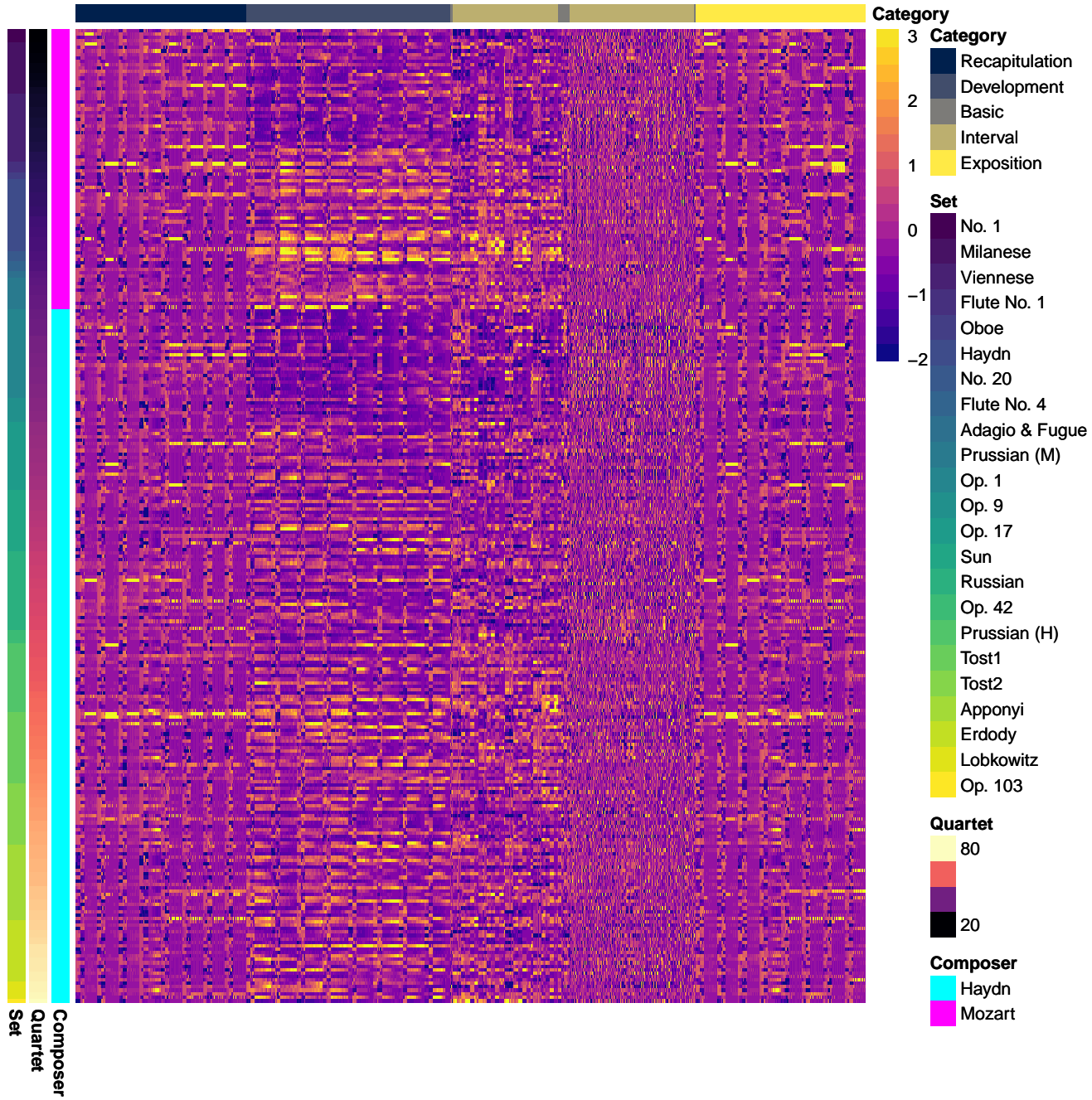


Figure S2.: The heatmap represents the column-standardized, ordered $n \times p$ feature matrix X on the HM285 dataset. The rows and columns correspond to movements and features, respectively. The left annotations mark set, quartet, and composer groups of each movement, while the top annotation denotes the category of each feature.

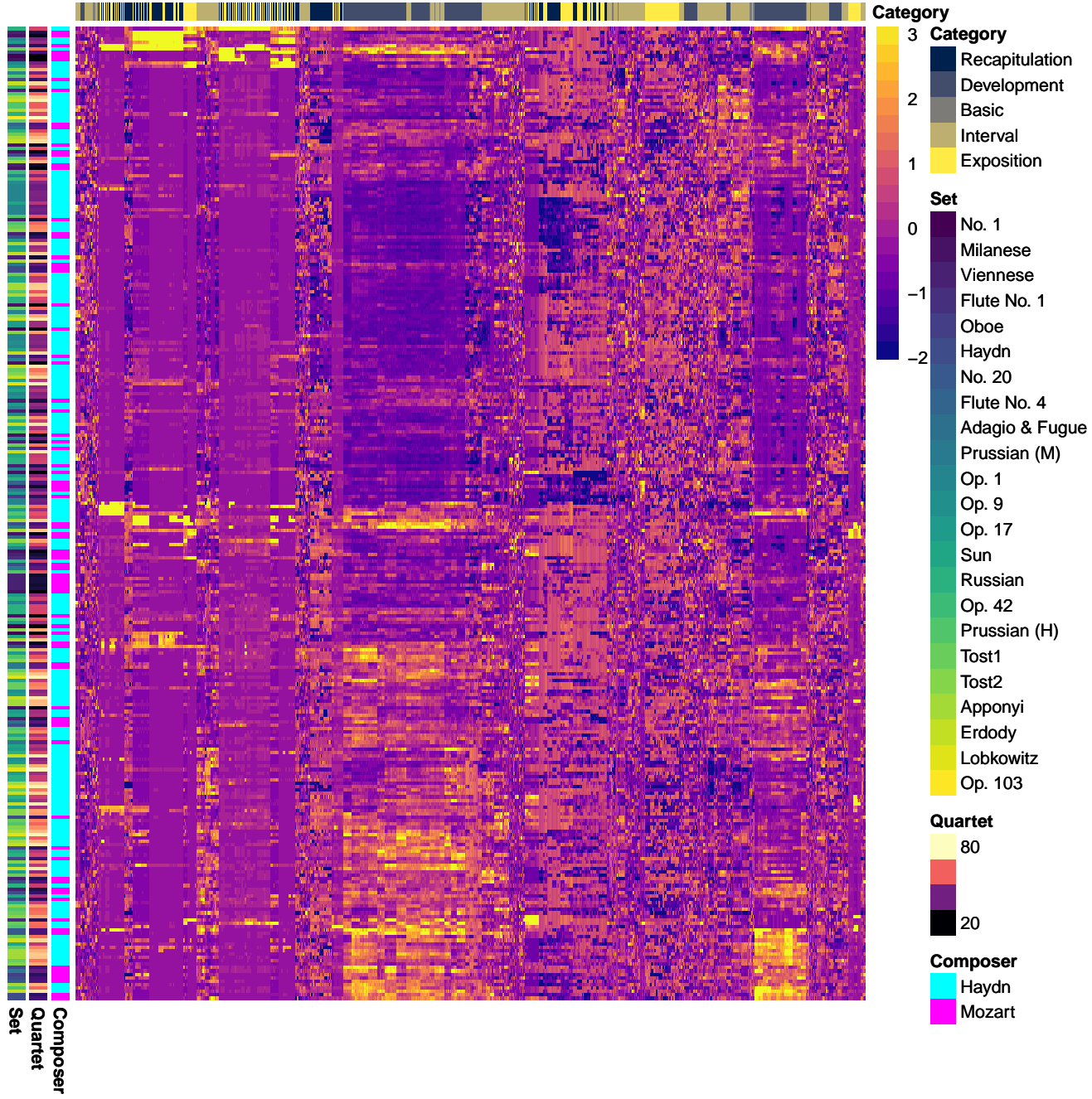


Figure S3.: The heatmap represents the column-standardized $n \times p$ feature matrix X on the HM285 dataset from Figure S2, with rows and columns reordered via hierarchical clustering. The rows and columns correspond to movements and features, respectively. The left annotations mark set, quartet, and composer groups of each movement, while the top annotation denotes the category of each feature.

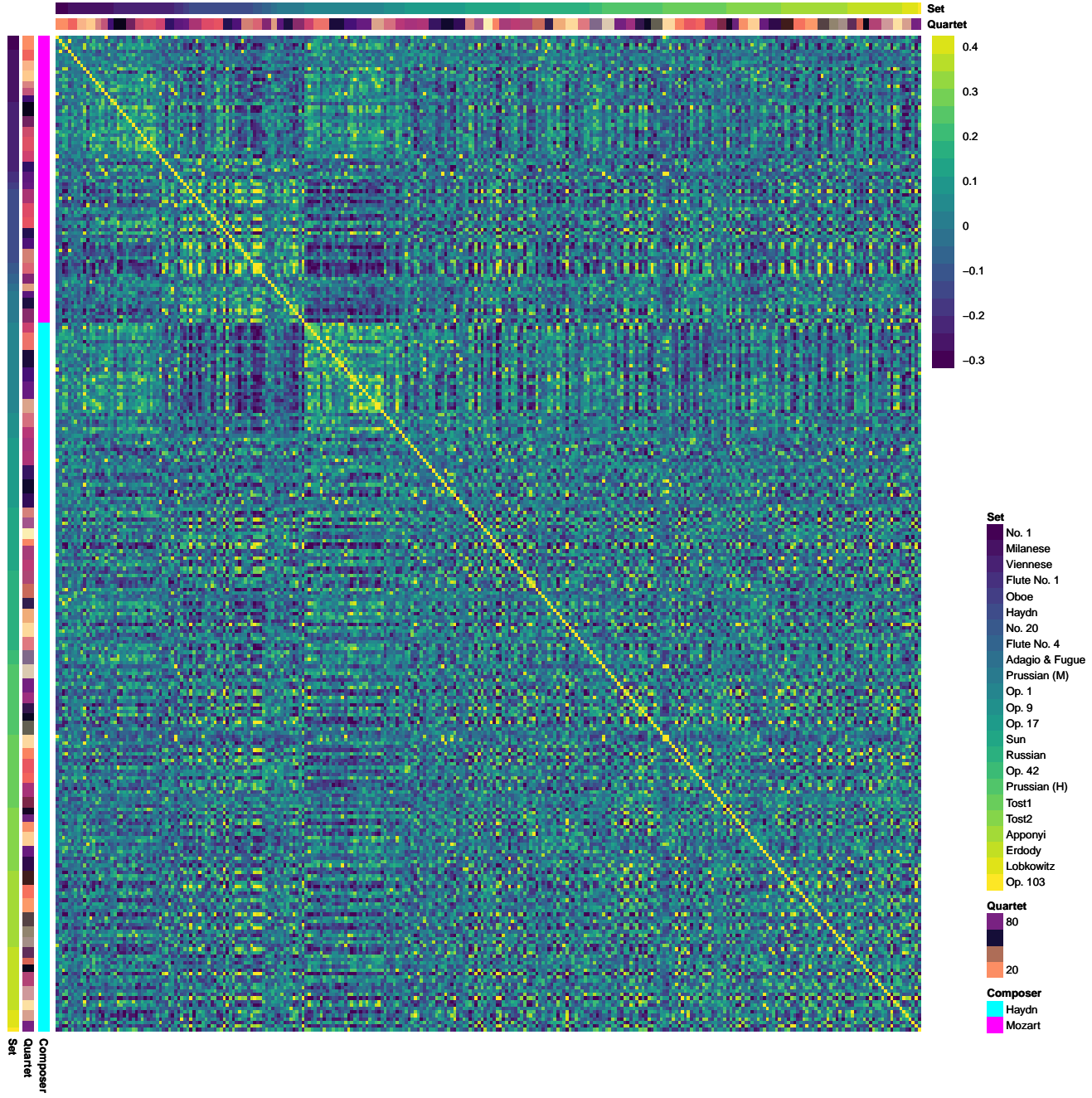


Figure S4.: The heatmap depicts the $n \times n$ Gram matrix XX' of the column-standardized, ordered feature matrix X . For all $i, j = 1, 2, \dots, n$, row i and column j of XX' corresponds to the correlation between rows i and j of X . The left annotations mark set, quartet, and composer groups of each movement, and the top annotations repeat the set and quartet groups for improved readability.

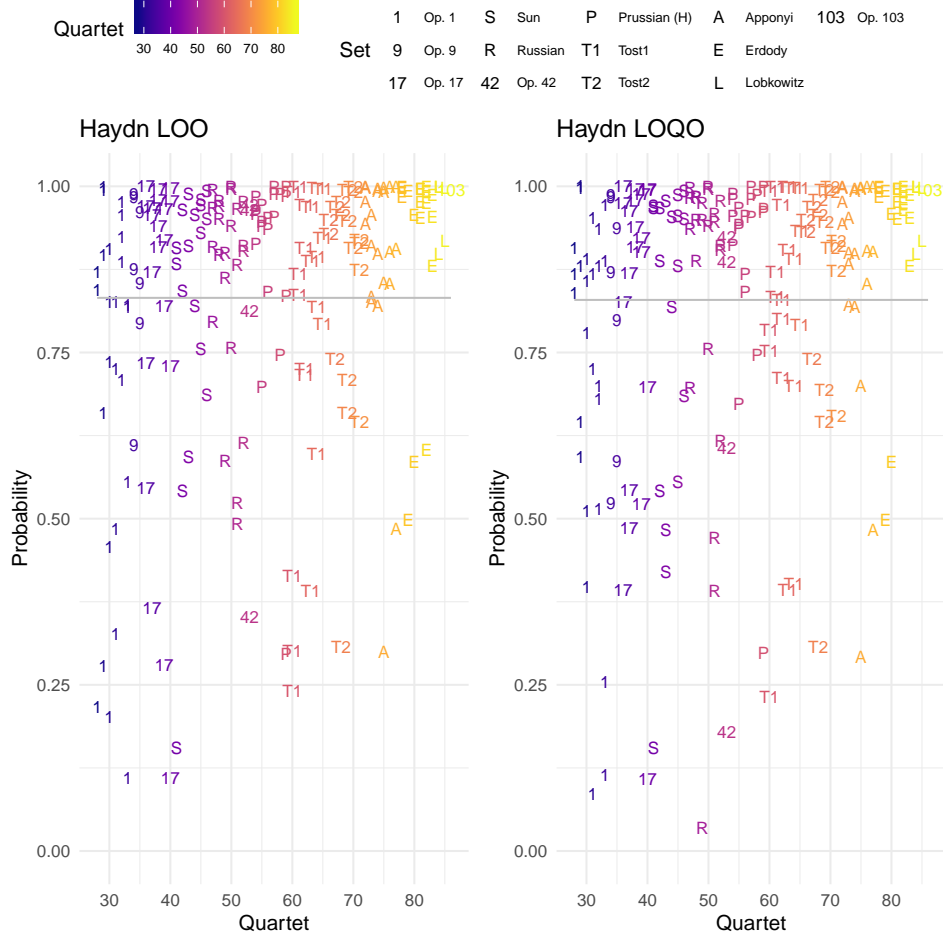


Figure S5.: For Haydn movements in the HM285 dataset, we plot the estimated probabilities by quartet and set, which result from the LOO CV scheme (left) and the LOQO CV scheme (right). In each plot, the gray line marks the mean of the estimated probabilities over the movements. For improved visualization in these plots, quartets have been ordered by approximate date of composition then converted to a numerical scale.

quartets and sets. Thus, any patterns in the Gram matrix heatmap are mild and seem more related to random chance than quartet, set, composer, or date. As above, we form heatmaps for other subsets of features and movements (included in the Github repository), but they display no notable patterns.

Next, we analyse the estimated probabilities from LOO and LOQO to identify potential patterns. Figures S5 (Haydn) and S6 (Mozart) show estimated probabilities from LOO (left) and LOQO (right) by quartet and set. For each composer, the plots look very similar for LOO and LOQO, indicating that LOQO has not substantially affected the results, despite reducing the size of available training data in each CV fold. This stability across the CV schemes suggests that it was appropriate for movements within a quartet to be treated as approximately independent for LOO. Additionally, there are no discernible patterns by set or date of composition in these plots, implying that movements within a set or similar time period are also approximately independent.

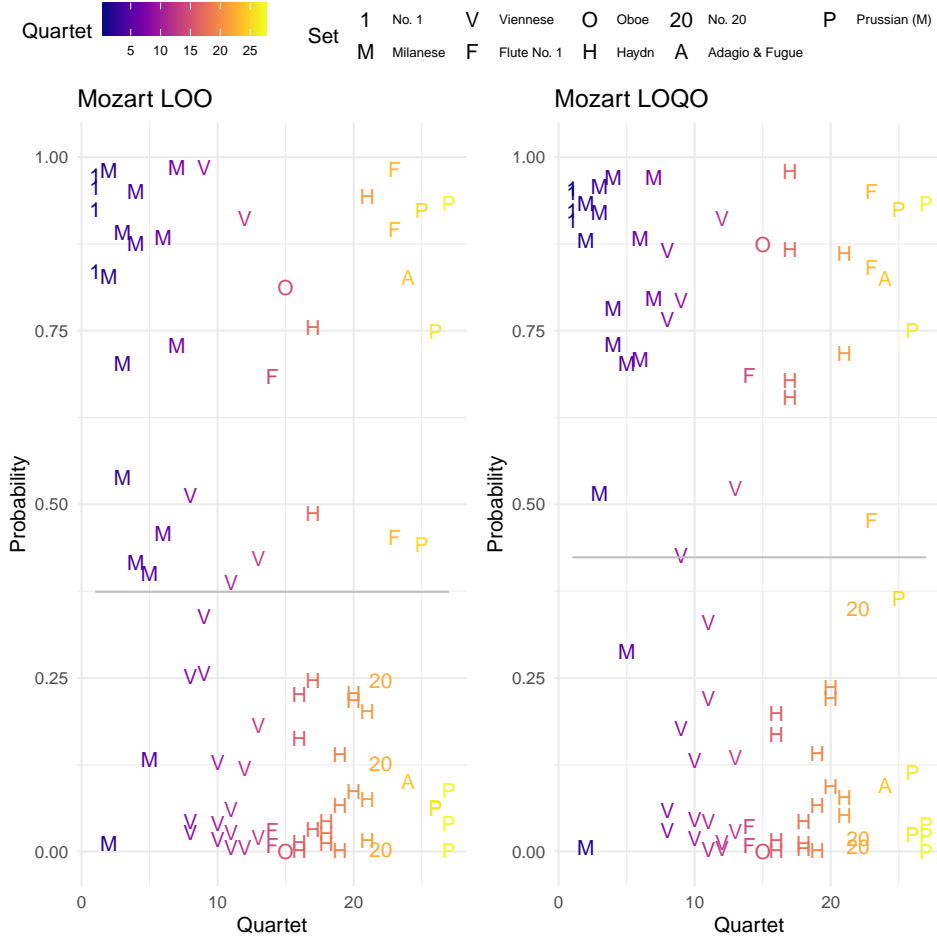


Figure S6.: For Mozart movements in the HM285 dataset, we plot the estimated probabilities by quartet and set, which result from the LOO CV scheme (left) and the LOQO CV scheme (right). In each plot, the gray line marks the mean of the estimated probabilities over the movements. For improved visualization in these plots, quartets have been ordered by approximate date of composition then converted to a numerical scale.

| Group | Percentage |
|---|------------|
| LOO probability < LOQO probability | 26.67% |
| LOO probability = LOQO probability | 44.21% |
| LOO probability > LOQO probability | 29.12% |
| LOO class < LOQO class | 5.61% |
| LOO class = LOQO class | 91.93% |
| LOO class > LOQO class | 2.46% |
| LOO probability < LOO-reduced probability | 45.96% |
| LOO probability = LOO-reduced probability | 2.11% |
| LOO probability > LOO-reduced probability | 51.93% |
| LOO class < LOO-reduced class | 10.88% |
| LOO class = LOO-reduced class | 81.75% |
| LOO class > LOO-reduced class | 7.37% |

Table S1.: The estimated composer probabilities and corresponding classes from the Bayesian logistic regression model (with class label 0 for Mozart and class label 1 for Haydn) are compared among LOO, LOQO, and LOO on the reduced feature set. Estimated probabilities for each movement have been rounded to the nearest hundredth when determining if they are equal here.

Numeric summaries further demonstrate that the estimated probabilities and classes are similar for LOO and LOQO. As displayed in Table S1 (top rows), close to half of the movements in HM285 (44.21%) have the same estimated probabilities from LOO and LOQO CV, up to rounding. Any differences in estimated probabilities between the CV schemes infrequently translate to different predicted classes. Only about 9% of movements in the dataset have different predictions for LOO versus LOQO, mostly corresponding to correct classifications in LOO and incorrect classifications in LOQO. While Mozart movements are more likely than Haydn movements to have different estimated probabilities between the two CV schemes (72% for Mozart, 50% for Haydn), Mozart is almost as likely as Haydn to have classifications that agree across CV (90% for Mozart, 92% for Haydn). Overall, these analyses confirm the stability and validity of our CV schemes.

S3. Comparing LOO on the Full Feature Set versus the Reduced Feature Set

As highlighted in Section 5.2, we compare the testing results of LOO on the full feature set versus the reduced feature set. Such stability comparisons can offer information about the contribution of our novel sonata-inspired features, which are excluded in the reduced feature set. We start by analysing the estimated probabilities and classes from each testing fold (left-out movement). As displayed in Table S1 (bottom rows), only about 2.11% of the movements in HM285 have the same estimated probabilities (up

to rounding) for LOO on the full set versus the reduced set. Recall from above that LOQO gave the same estimated probability as LOO for about half the movements. About 18% of movements in HM285 have different class predictions for LOO on the full versus reduced sets; in contrast, LOQO gave different class predictions than LOO for about 9% of the movements. Additionally, we form plots that compare estimated probabilities on the full versus reduced feature sets, similarly as in Figures S5 and S6 and included in the Github repository. Altogether, these graphical and numeric results show that probability and class stability is maintained across CV schemes (from LOO to LOQO), but instability occurs across feature sets (from the full feature set to the reduced feature set).

Next, we consider the stability with respect to selected features of LOO on the full feature set versus the reduced feature set. Of the 56 movements misclassified in LOO on the reduced subset, 34 of them were also misclassified in LOO on the full feature set, while 22 of them were classified correctly when given access to the full feature set. Examining the models for the latter 22 movements/folds can suggest potentially important features. For LOO on the full feature set, all 22 of the models contain at least one development feature (most commonly, the ‘standard deviation count’ feature labeled (E) in Figure 3, represented in 20 models), and 20 of the models additionally contain a recapitulation feature (the ‘maximum fraction of overlap’ feature labeled (G) in the figure). For LOO on the reduced feature set, each of the 22 models contain the basic ‘standard deviation of duration’ feature (feature (A) in Figure 3), while 19 of the models include a ‘mean proportion of minor third intervals for Viola’ feature (similar to feature (D) in the figure). These models imply that some of the basic and interval features can contribute to high composer classification accuracies, but greater success in discriminating Mozart requires the sonata-inspired features.

Interestingly, there are 16 movements misclassified from LOO on the full feature set but classified correctly from LOO on the reduced subset. One possibility is that these movements have structures outside sonata and related forms. However, there are no clear patterns among the movements, other than Haydn predominance (15 movements by Haydn and only 1 by Mozart). This result reinforces that, without the sonata-inspired features, a classifier tends to guess the majority class, Haydn.

References

Kolde, R. (2012). Pheatmap: pretty heatmaps. *R package version, 1*(2).