AutoSchA: Automatic Hierarchical Music Representations via Multi-Relational Node Isolation: Technical Appendix

Stephen Ni-Hahn[*1], **Rico Zhu**[*1], **Jerry Yin**[2], **Yue Jiang**[1], **Cynthia Rudin**[1], **Simon Mak**[1]

[*]Equal Contribution [1]Duke University [2]Stanford University
{stephen.hahn, rico.zhu}@duke.edu

A Ablation Experiments

A.1 Implementation Details

All models are trained using 4 Nvidia P100 GPUs. For AutoSchA without global embeddings, all GNNs have a hidden dimension of 32; for AutoSchA with both global embedding approaches, the initial node embedding GNN has a hidden dimension of 16, and the scoring GNN has a hidden dimension of 32. We use Adam for our optimizer with a learning rate of 0.001. We also use an exponential learning rate scheduler with $\gamma = 0.1$.

AutoSchA (Base)	AutoSchA (Sequence)	AutoSchA (Grassmann)
676,477	623,421	623,726

Table 1: The number of parameters for the ablation studies.

A.2 Hyperparameter Tuning

We test the effect of modifying the hyperparameters α (the weight for forwards and backwards edges in the directed convolution) and $c_{\min}$ (minimum masking threshold). We measure the average test set accuracy over 10 runs, training the models for 3 epochs each; we always take the accuracy of the last epoch. The results are shown in Figure 1. We see that a combination of a threshold of $c_{\min} = 0.5$ and a directionality weight of $\alpha = 0.75$ is optimal. The performance of $\alpha = 0.5$ is signifiantly stronger than $\alpha = 0$ and $\alpha = 1$, implying that including both forwards and backwards edges is greatly beneficial to the Schenkerian task; a forwards bias in the directionality further improves our performance.

A.3 Node and Edge Features

We also measure the effects of removing node and edge level features. For our each ablated model, we trained 30 models for 3 epochs. Figure 2 shows the maximum test accuracy for each ablated model over the 90 epochs the model copies were trained. Based on these results, we conclude that rhythmic features are most vital to model performance. Surprisingly, including too many pitch features seems to confuse

Figure 1: Hyperparameter ablations of AutoSchA with sequential global embeddings, displaying the effects of $c_{\min}$ and α on mean test set accuracy over 10 runs.

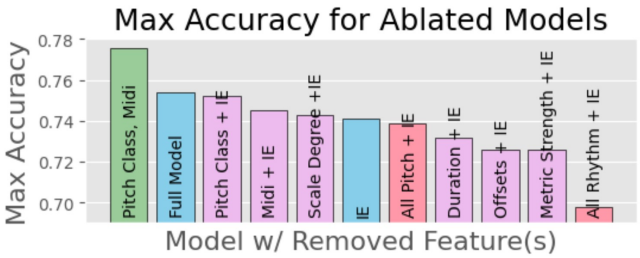


Figure 2: Ablation experiment results. X-axis labels indicate what features were removed from respective models. Blue bars indicate models that include all node features, one including interval edges (IE) and one without. Purple bars indicate models with certain node features removed. Red bars are models with multiple related features removed. The largest, green bar indicates the optimal model.

the model. The maximum accuracy was achieved by removing two of the three pitch features: pitch class and midi. Note that removing all pitch features does seem to negatively affect performance, but to a very small degree.

We hypothesize that larger, more complex musical scores, and scores of different styles may show very different variable importance metrics. For instance, a style that includes many overlapping suspensions would not rely on metric strength as much, for important structural tones would often be offset from strong metrical beats. Certain pitch features, such as pitch class, would be vital when tonicization becomes more abundant.

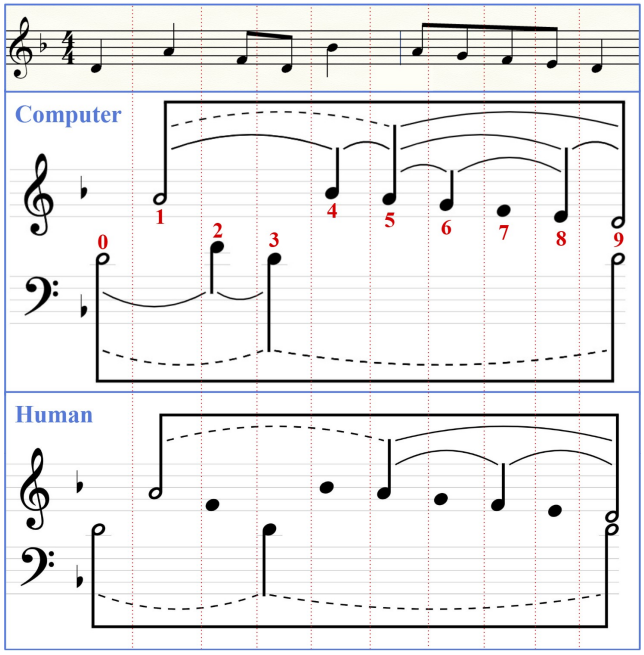


Figure 3: The score of Pachelbel's Primi Toni 1 fugue subject (top), AutoSchA's analysis (middle), and the human analysis (bottom). Note indices are provided in red for convenience.

B Case Study: Pachelbel's Primi Toni

To get a better understanding of the Schenkerian analysis and our model's musical ability, we present a case study of AutoSchA's analysis of Pachelbel's *Primi Toni* 1. Figure 3 shows the score and analyses of the fugue subject.

First, we investigate the voice labeling results. The model agrees with the human analyst on the voice of each note except Note 2. This is impressive considering Notes 8 and 9 are assigned to the treble voice despite being closer to the initial bass note (Note 0) than the initial treble note (Note 1). This suggests the model may learn an understanding of the linear connections in the treble voice.

Regarding the structural analysis itself, there is a considerable difference between the computer and human analysis shown. However, both analyses are perfectly valid, showing different interpretations of the same music. The human analysis sees the excerpt as expanding a single harmony (D minor I), while the computer analysis interprets the excerpt as an entire phrase (I–II–V–I). This is clear when looking at Notes 4-9. The human analysis interprets Note 4 as a foreground upper neighbor to Notes 1 and 5, while the computer recognizes Note 4 as a more structural tone. The human analysis outlines the tonic triad (D minor) by showing Notes 5, 7, and 9 as more structural tones than Notes 6 and 8. On the other hand, the computer analysis outlines a dominant 7th harmony (A dom.) by showing the relative structure of Notes 5 and 8 over Note 7. Further still, Note 6, the 7th of the dominant harmony is given more structure than Note 7, a passing tone connecting the 7th and 5th of the harmony.

C Subspace Merging Technical Details

The subspace merging approach aims to find a unified global topology that can capture the node connections over all edge types (see Method 2 of Figure 4 of the main paper). For simplicity, we focus on the undirected representation of our input graphs, $\hat{\mathbf{A}}_i$ (computed as the logical *or* between $\mathbf{A}_i$ and $\mathbf{A}_i^\intercal$), which we use to generate a fused global graph $\mathbf{A}_{\text{mod}}$. To use this global topological information, we then convolve the initial embeddings over the fused graph, generating node-level features which integrate information over all layers of the graph.

Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with adjacency and degree matrices $\mathbf{A}$ and $\mathbf{D}$ respectively, the *normalized graph Laplacian* is defined to be $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-\frac{1}{2}}$. Given graph Laplacians $\mathbf{L}_1, ..., \mathbf{L}_m$, we can compute an optimal k-dimensional spectral embedding $\mathbf{U}_i \in \mathbb{R}^{n \times k}$ for each graph (**?**). Note that $\mathbf{U}_i$ is the matrix whose columns are the eigenvectors corresponding to the k smallest eigenvalues of $\mathbf{L}_i$. We interpret $\mathbf{U}_1, ..., \mathbf{U}_m$ as a particular representation for $\mathbf{A}_1, ..., \mathbf{A}_m$, defined by a deterministic feature transformation. Because the column space of $\mathbf{U}_i$ spans a k-dimensional subspace of $\mathbb{R}^n$, we can view $\mathbf{U}_i$ as an element of the *Grassmannian* $\mathbf{Gr}(k, n)$, the manifold of all k-dimensional linear subspaces in $\mathbb{R}^n$.

We can measure distance between points $\mathbf{U}_1, \mathbf{U}_2 \in \mathbf{Gr}(k, n)$ using the squared projection distance $d^2(\cdot, \cdot)$ (see **?**):

$$d^2(\mathbf{U}_1, \mathbf{U}_2) = \frac{1}{2}\|\mathbf{U}_1\mathbf{U}_1^\intercal - \mathbf{U}_2\mathbf{U}_2^\intercal\|_F^2$$
$$= \frac{1}{2}(\text{tr}(\mathbf{U}_1\mathbf{U}_1^\intercal) - 2\text{tr}(\mathbf{U}_1\mathbf{U}_1^\intercal\mathbf{U}_2\mathbf{U}_2^\intercal) + \text{tr}(\mathbf{U}_2\mathbf{U}_2^\intercal))$$
$$= k - \text{tr}(\mathbf{U}_1\mathbf{U}_1^\intercal\mathbf{U}_2\mathbf{U}_2^\intercal)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. $d^2(\cdot, \cdot)$ may be interpreted as the square of the *projection metric*, a $\sqrt{2}$-approximation to the true Riemannian geodesic between points on the Grassmann manifold (**?**). With a way to measure distance between transformed adjacency matrices, we can again try to compute an optimal spectral embedding for the aggregate edge representation. This is formulated as a trace minimization problem as shown in **?**:

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times k}} \text{tr}\left(\mathbf{U}^\intercal\left(\sum_{i=1}^{m}(\mathbf{L}_i - \mathbf{U}_i\mathbf{U}_i^\intercal)\right)\mathbf{U}\right), \mathbf{U}^\intercal\mathbf{U} = \mathbf{I}_n,$$

where, via the Rayleigh-Ritz theorem, the solution is given by the corresponding eigenvectors to the k-smallest eigenvalues of the matrix:

$$\mathbf{L}_{\text{mod}} = \sum_{i=1}^{m}(\mathbf{L}_i - \mathbf{U}_i\mathbf{U}_i^\intercal).$$

Here, $\mathbf{L}_{\text{mod}}$ represents a globally fused subspace representation for the graph. With a global characterization for our graph, we now compute an appropriate embedding for each node. We consider a graph convolution on the extracted modified adjacency matrix induced by $\mathbf{L}_{\text{mod}}$:

$$(\mathbf{A}_{\text{mod}})_{g,h} = \begin{cases} 1, & \text{if } g = h, \\ -(\mathbf{L}_{\text{mod}})_{g,h}, & \text{if } g \neq h, \end{cases}$$

where g and h index the matrix row and column respectively.

This additional node representation can be interpreted as adding an alternate view to the input graph, similar to the multi-view framework as established by **?**. We find that this approach gives improvements in predictive accuracy over the baseline model with a comparable number of parameters.

D Human Experiments

D.1 Survey Instrument

Example screenshots of the survey instrument are found in Figures 4, 5, and 6. Figure 4 shows the score overview page, where each analysis may be compared with each other and the score. Figures 5 and 6 show an example page of questions regarding on of the analyses.

D.2 Additional Figures and Analysis

Figures 7, 8, and 9 show the distributions for each major question in the survey experiment over the three different analysts. When it comes to GPA, Figure 7 shows a much broader variance for the flawed model and relatively similar performance between human and computer models.
Musicality (Figure 8) shows a similar trend, with surprisingly lower scores for the human analyst. Finally, Figure 9 shows Turing scores have a heavy weighting around the middle value (5) with a wide variance in all three violin plots, indicating general uncertainty. However, there is a larger bump towards the top of the human and computer models that provides strong evidence for passing the Turing test.

Based on the positive correlation, Figure 10 suggests that survey participants generally gave pieces similar musicality and Turing scores. We noticed a couple participants did not follow this trend, so we asked how they determined their Turing score:

> "I was not attempting to decide whether the computer did a better or worse job than a human. I was actually looking at the types of mistakes made. Basically, I tried to imagine a student making this choice or that choice and/or determine what line of reasoning might lead to the mistake. In other words, I just found certain analyses to have the sort of mistakes that I did not think a human would make. I believe that I probably ranked some of the better analyses as a 5 for the Turing simply because I could not possibly know if it was accurate because it was done by an insightful student or a well-learned machine."

Figure 11 shows the distribution of grades assigned to the human model and computer model for the same musical score. In the majority of cases, the human model outperforms the computer model. However, there are many instances of the reverse. The number of poor grades assigned to the human analyst suggests either a mistake on the analysts part, a mistake on the survey participants part, or a strong difference of opinion between the human analyst and survey participant. We investigate the distribution of grades assigned by individual survey participants in Figure 12. We see that participants 1, 4, and 5 were relatively generous with their grades, while 2 and 3 were more willing to fail analyses.

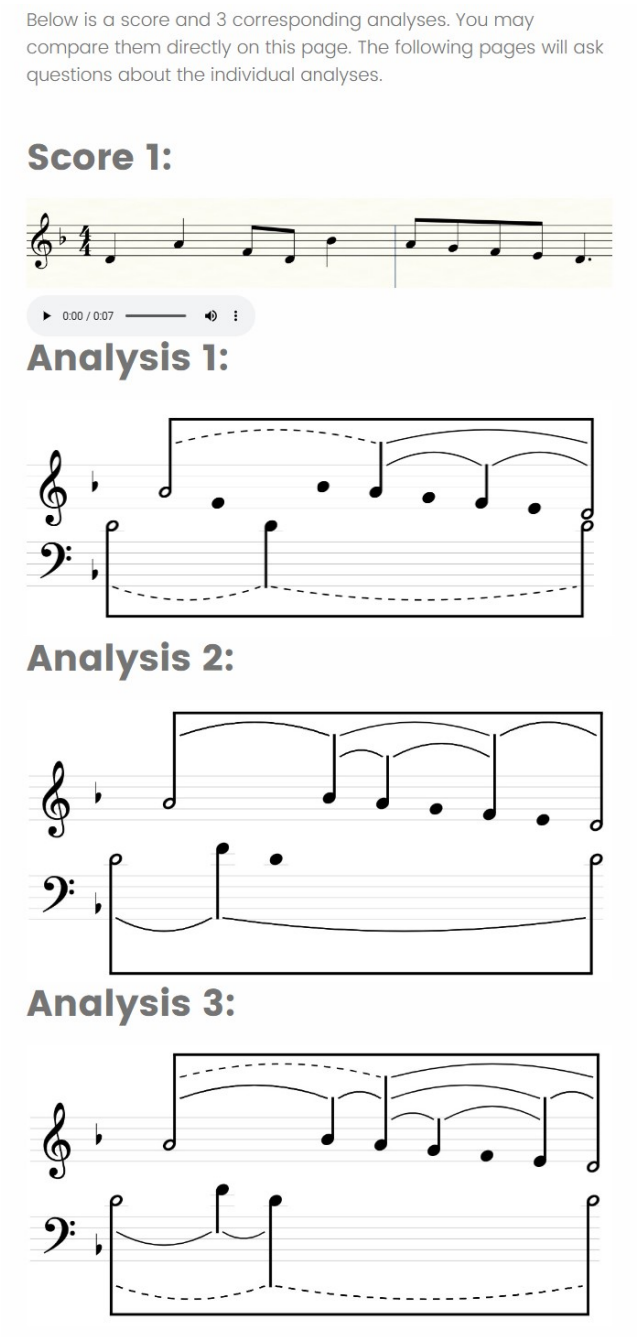


Figure 4: Screenshot of a survey instrument score and analysis overview page.

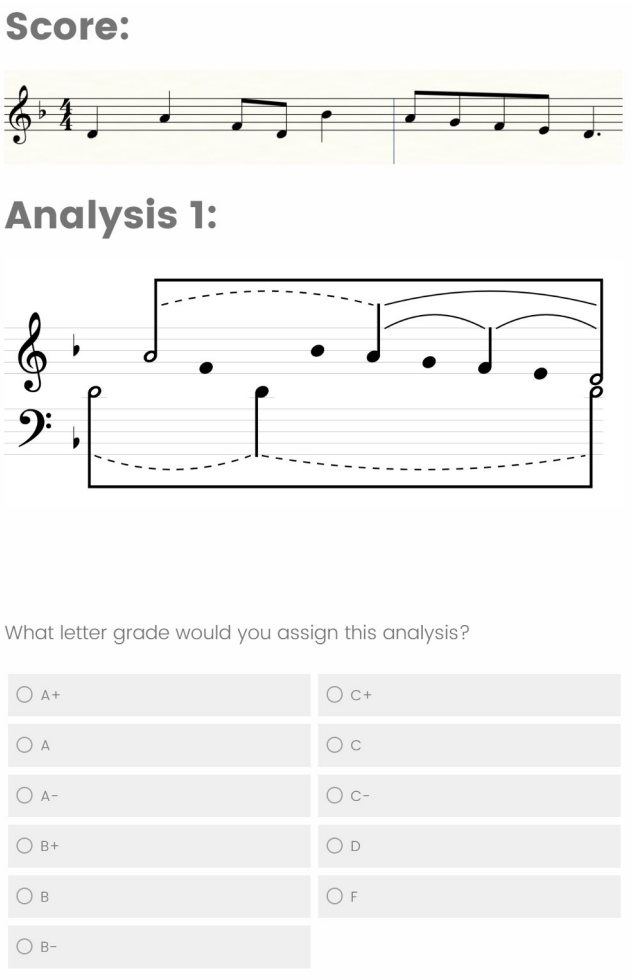


Figure 5: Screenshot of a survey instrument analysis questions page

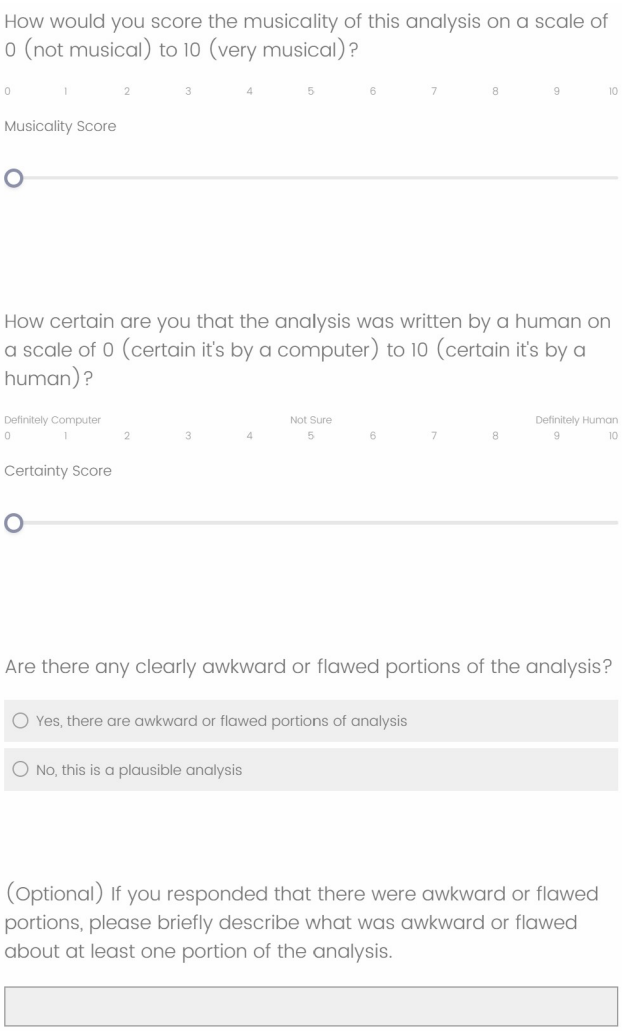


Figure 6: Continuation of analysis questions from Figure 5

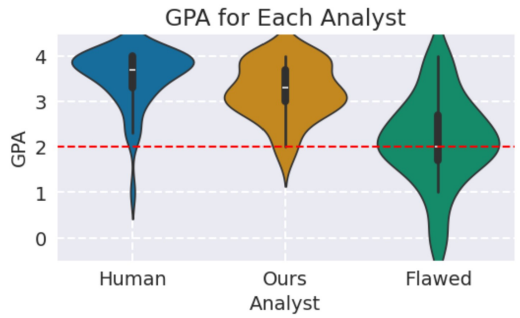


Figure 7: Distribution of GPA for each method of analysis.

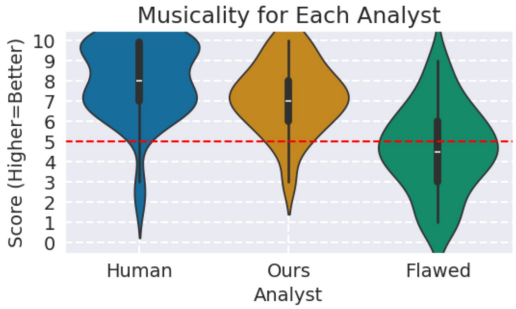


Figure 8: Distribution of musicality for each method of analysis.

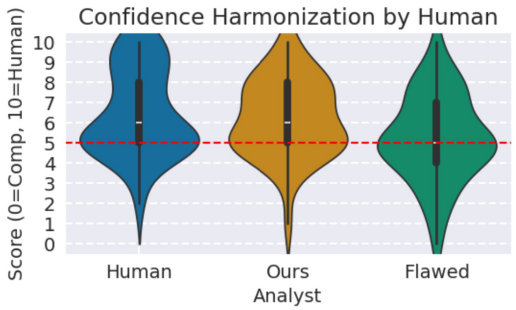


Figure 9: Distribution of Turing test scores for each method of analysis.

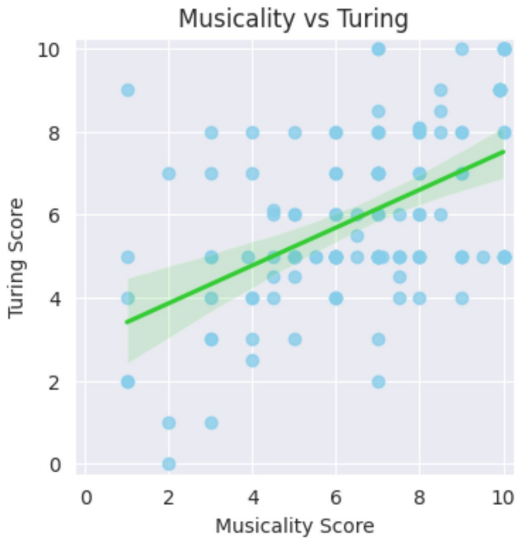


Figure 10: Distribution of musicality and Turing scores with best fit regression line.

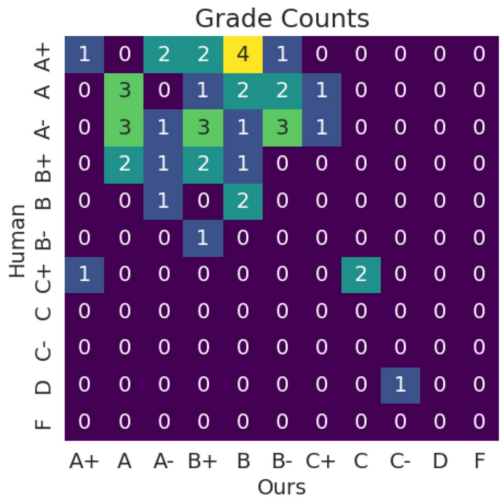


Figure 11: Comparison of assigned grades between the human analyses and AutoSchA analyses.

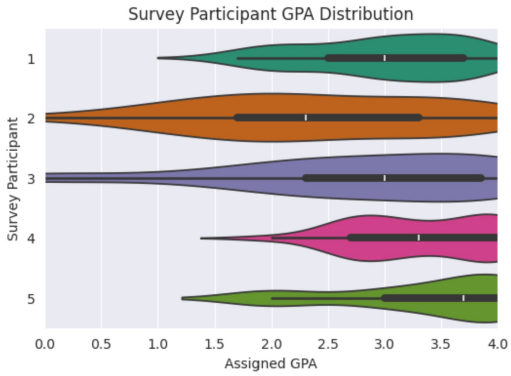
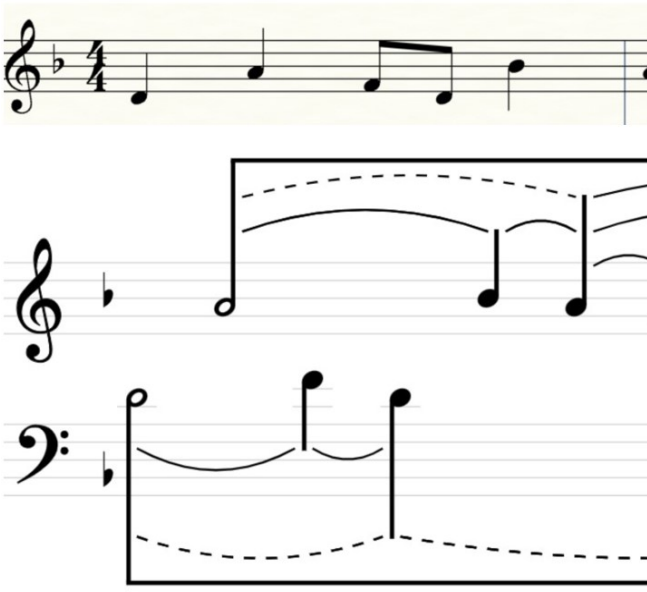


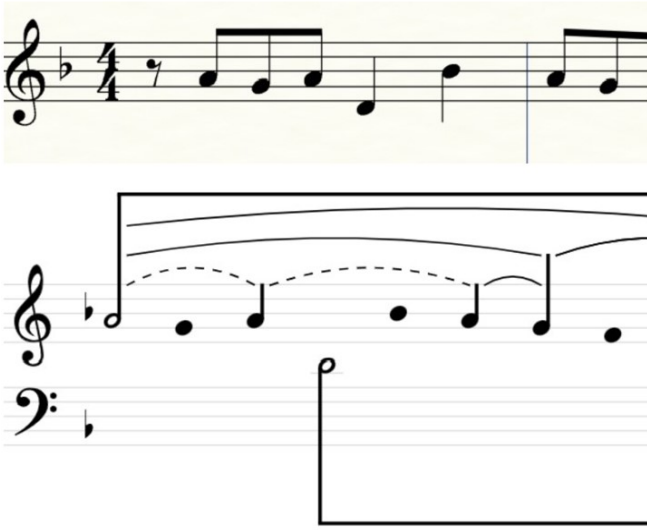
Figure 12: Distribution of GPAs assigned by each survey participant.

E Sample AutoSchA Analyses

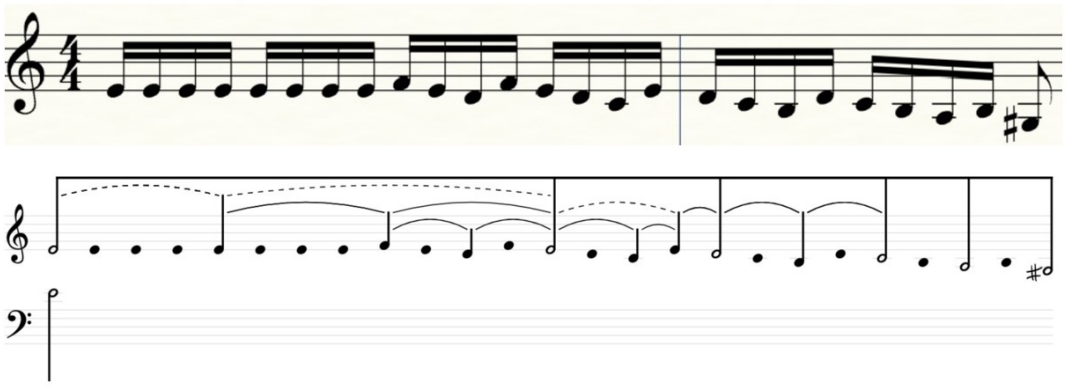
Primi Toni 1



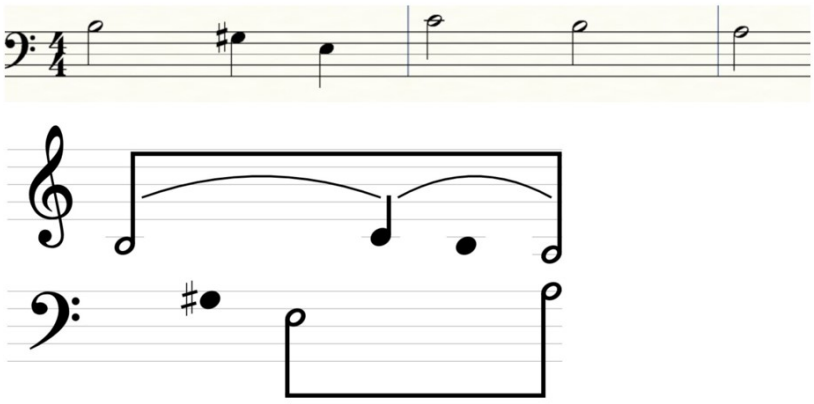
Primi Toni 12



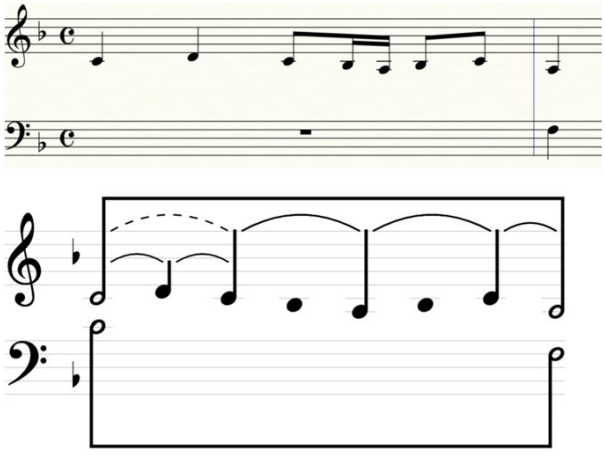
Quarti Toni 4



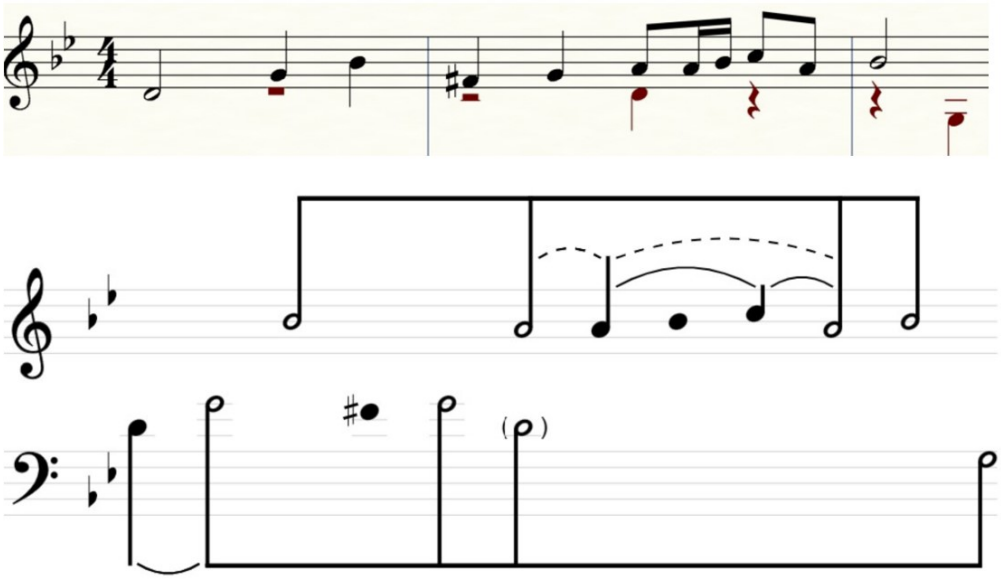
Quarti Toni 5



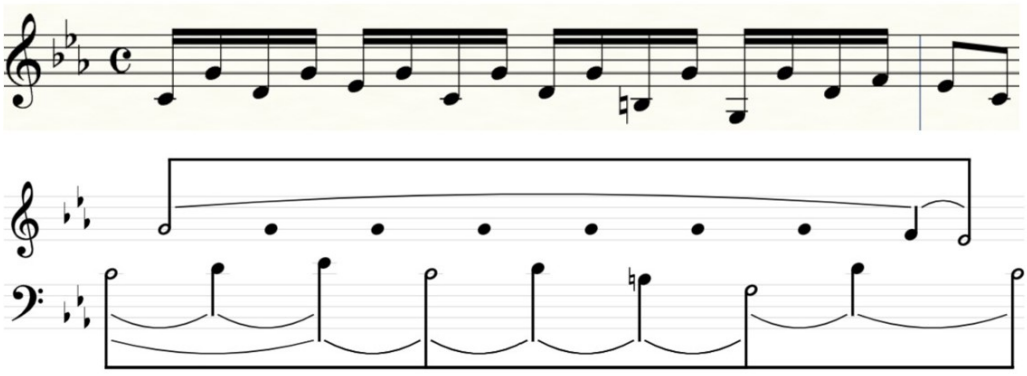
Quinti Toni 7



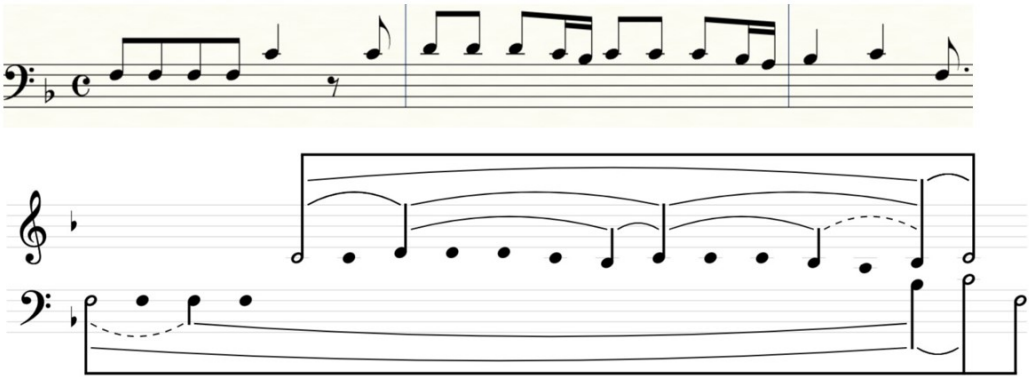
Secundi Toni 6



Septimi Toni 3



Sexti Toni 6



Tertii Toni 4