Where Does Instrumentality Reside in Symbolic Music? A Structural Analysis of a Symbolic-Domain Instrument Classifier

Shun Sawada*
Department of Information Technology and Media Design
Nippon Institute of Technology
Miyashiro, Saitama, Japan

Abstract

Instrument recognition has traditionally relied on timbral cues in audio signals. However, in symbolic music representations such as MIDI, timbre is absent, raising the question of how instrument identity is inferred from note sequences alone. In this work, we investigate how instrumentality is approximated in a learned symbolic-domain instrument classifier by analyzing its behavior under controlled structural perturbations. Rather than focusing on classification accuracy, we examine how perturbations such as pitch normalization, voice extraction, and temporal reversal affect predicted instrument labels and internal representations. Our results show that instrument identity is strongly influenced by structural features such as pitch register and simultaneous note configuration, while temporal ordering also has non-negligible effects. Furthermore, changes in predicted labels are not directly explained by the magnitude of representation shifts, suggesting structured reassignment patterns rather than uniform degradation. These findings suggest that, when timbral information is unavailable, the analyzed classifier relies on structural regularities in symbolic patterns to infer instrument identity. This provides an empirical perspective on how instrumentality can be approximated in learned symbolic representations through structural cues rather than timbral information itself.

1 INTRODUCTION

The rapid expansion of generative AI has transformed how music is created, distributed, and understood. Contemporary systems can produce large volumes of music from symbolic representations such as MIDI, where musical structure is explicitly encoded but timbral information is absent. This shift raises fundamental questions about how musical meaning is represented and inferred when sound itself is no longer directly available.

Among these questions, instrument identity plays a central role. Traditionally, instrument recognition has relied on timbral characteristics in audio signals, which provide rich cues for distinguishing between instruments. Although symbolic representations such as MIDI do not contain explicit timbral information, and instrument labels may only be available as metadata, recent studies have shown that models can infer instrument labels from note sequences to a certain degree (Sawada, 2024, 2025; Ji et al., 2021). This observation suggests that instrumentality may not be solely tied to sound, but could also emerge from structural properties of musical representations.

This raises a fundamental question: where does instrumentality reside in symbolic music? If timbre is absent, what cues allow a model to distinguish between instruments? More importantly, what does this reveal about how instrument identity is internally represented?

In this work, we approach this question not through improving classification accuracy, but through analyzing the behavior of a learned model under controlled structural perturbations. We examine how modifying specific aspects of symbolic input—such as pitch range, voice configuration, and temporal ordering—affects predicted instrument labels and internal representations. By systematically disrupting these structural cues, we aim to identify which aspects of the learned representation in this model contribute to instrument identity.

The contributions of this paper are as follows:

- We apply controlled structural perturbations to analyze how instrument identity is encoded in a learned symbolic-domain instrument classifier.
- We demonstrate that pitch register, simultaneous note configuration, and temporal ordering affect the classifier's behavior in different ways.
- We show that perturbations induce structured reassignment across instrument categories rather than uniform degradation.
- We demonstrate that representation shifts do not directly correspond to classification changes, indicating that the classifier's behavior depends on how perturbations interact with decision boundaries.

2 RELATED WORK

2.1 Instrument Classification in Audio and Symbolic Domains

Musical instrument classification has been studied in both audio and symbolic domains. In the audio domain, prior work has explored a range of approaches, from statistical pattern recognition using hand-crafted acoustic features (Essid et al., 2006) to deep learning models based on spectrograms or audio excerpts (Han et al., 2017). Recent studies have also considered more realistic settings such as hierarchical multi-label instrument classification (Zhong et al., 2023). These approaches typically exploit acoustic characteristics of instrument sounds and have achieved strong performance in audio-based instrument recognition. Compared with audio-based approaches, instrument clas-

*sawada.shun@nit.ac.jp

sification from symbolic representations has received less attention.

Symbolic-domain studies formulate the task as predicting instrument labels from note sequence representations such as MIDI or piano rolls (Ji et al., 2021; Sawada, 2024). Recent work has further investigated audio-informed learning frameworks for this task. For example, knowledge distillation from audio-domain teachers and contrastive Audio–MIDI learning have been used to improve symbolic-domain instrument classification while requiring only symbolic inputs at inference time (Sawada, 2024, 2025). Most of these studies focus on improving classification performance. In contrast, less attention has been paid to analyzing which aspects of symbolic representations contribute to instrument classification decisions.

2.2 Analysis of Learned Representations

In parallel with performance-oriented research, a growing body of work has analyzed the internal representations of neural networks. Probing methods (Alain and Bengio, 2016), network dissection (Bau et al., 2017), and representation similarity analysis (Kornblith et al., 2019) have been used to investigate what information is encoded in learned representations. Another line of research examines model behavior through input perturbation and feature ablation. By modifying or removing parts of the input, these methods analyze how predictions change and identify input components that influence model outputs (Zeiler and Fergus, 2014; Fong and Vedaldi, 2017).

In music research, learned representations have also been studied in symbolic music modeling and classification. However, relatively little work has examined how discriminative models for symbolic-domain instrument classification rely on specific structural cues to make category decisions.

2.3 Position of This Work

This work analyzes how instrument identity is reflected in the learned representations of a symbolic-domain instrument classifier. We apply controlled structural perturbations that modify pitch range, voice configuration, and temporal ordering, and examine their effects on predicted labels and internal representations. Our approach complements prior work by shifting the focus from classification performance to the relationship between symbolic input structure and model behavior.

3 METHOD: STRUCTURAL PERTURBATION ANALYSIS

3.1 Task Setup

We consider the task of musical instrument classification from symbolic representations. Given a segment of a MIDI sequence, the model predicts the instrument label associated with the sequence. Following prior symbolic-domain instrument classification studies (Sawada, 2024, 2025), we use Lakh MIDI-derived note sequences (Raffel, 2016) with the eight-category instrument-label setting derived from Slakh (Manilow et al., 2019).

Each input is a two-measure MIDI segment represented as a binary piano roll. The pitch range is limited to 84 bins, spanning from C1 to B7, and each two-measure segment consists of 192 time steps. Thus, each input is represented as $X \in \{0,1\}^{84\times192}$, where the first dimension corresponds to pitch and the second dimension corresponds to time. The classification target consists of eight instrument categories: Guitar, Piano, Bass, String, Organ, Brass, Pipe, and Reed.

We analyze a supervised symbolic-domain instrument classifier trained directly on MIDI note sequences, following the SIC setting of prior work (Sawada, 2024, 2025). The classifier used in this analysis is based on a ResNet-style convolutional architecture. By focusing on a symbolic-only model, we aim to isolate structural cues available in MIDI representations without introducing acoustic supervision. For representation analysis, we use the output of the encoder part before the final classification layer as the learned representation.

3.2 Structural Perturbations

To investigate which aspects of symbolic representations contribute to instrument identity, we introduce a set of structural perturbations. Let $X \in \{0,1\}^{P\times T}$ denote a binary piano-roll representation, where P is the number of pitch bins and T is the number of time steps. Each perturbation produces a modified piano roll $\tilde{X}$ by selectively altering a specific structural property of X.

Top voice extraction (top_voice): For each time step t, we retain only the highest active pitch:

$$p_t^{\mathrm{top}} = \max\{p \mid X_{p,t} > 0\}. \tag{1}$$

All other active pitches at the same time step are removed. If no pitch is active at time step t, the frame remains silent. This transformation reduces polyphonic note configurations to a single upper voice.

Lowest voice extraction (lowest_voice): For each time step t, we retain only the lowest active pitch:

$$p_t^{\mathrm{low}} = \min\{p \mid X_{p,t} > 0\}. \tag{2}$$

All other active pitches at the same time step are removed, and silent frames remain unchanged. This transformation isolates the lowest voice and removes simultaneous note stacking above it.

Time reversal (time_reversal): The order of time frames is reversed:

$$\tilde{X}_{p,t} = X_{p,T-1-t}. \tag{3}$$

This operation preserves the pitch content and vertical note configuration at each time frame, but reverses the temporal order of the sequence.

Pitch normalization (pitch_center): For pitch normalization, we compute a pitch shift s for each segment and apply it to all active pitch indices so that the mean active pitch is aligned to a fixed target center. Let

$$\mu_X = \mathrm{round}\left(\frac{1}{|\mathcal{A}|}\sum_{(p,t)\in\mathcal{A}} p\right), \tag{4}$$

where $\mathcal{A} = \{(p,t) \mid X_{p,t} > 0\}$ is the set of active note positions. Given a target center p_c, the pitch shift is defined as

$$s = p_c - \mu_X. \tag{5}$$

Thus, the pitch shift s is computed separately for each input segment, rather than being shared across instrument classes or samples.

Each active pitch is shifted by s:

$$\tilde{X}_{p+s,t} = X_{p,t}, \tag{6}$$

when $0 \leq p + s < P$. Notes shifted outside the valid pitch range are removed. In our experiments, we set $p_c = 42$, corresponding to the center of the 84-bin pitch range. This operation reduces absolute pitch-register information by aligning each segment to a common central register.

These perturbations are not intended as realistic musical transformations. Rather, they are controlled interventions designed to isolate specific structural dimensions of the symbolic input, such as absolute pitch range, simultaneous note configuration, and temporal ordering.

3.3 Evaluation Metrics

We evaluate the effect of structural perturbations using both classification-based and representation-based measures.

Classification behavior: We measure standard classification performance using accuracy and F1 score. In addition, we compute the label change rate, defined as the proportion of samples whose predicted label changes after perturbation:

$$R_{\text{change}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(y_i^{\text{orig}} \neq y_i^{\text{pert}}\right), \tag{7}$$

where y_i^{orig} and y_i^{pert} denote the predicted labels before and after perturbation, respectively.

To analyze how instrument categories are redistributed, we also compare confusion matrices before and after perturbation.

Representation analysis: We quantify changes in internal representations using the embedding shift:

$$d_i = \left\| f(x_i^{\text{orig}}) - f(x_i^{\text{pert}}) \right\|_2, \tag{8}$$

where $f(\cdot)$ denotes the encoder.

To evaluate model confidence, we measure the change in maximum predicted probability:

$$\Delta c_i = \max_k p_k^{\text{orig}} - \max_k p_k^{\text{pert}}, \tag{9}$$

where p_k denotes the predicted probability for class k. This metric reflects the change in model confidence regardless of whether the predicted label changes, focusing on the sharpness of the prediction distribution.

3.4 Analysis Framework

Rather than focusing solely on performance degradation, we analyze the relationship between structural perturbations, representation shifts, and classification behavior.

In particular, we examine (i) whether large representation shifts lead to label changes, and more importantly, whether classification behavior can be explained by the magnitude of representation change, (ii) how different perturbations affect the distribution of predicted labels, and (iii) whether instrument categories are degraded or reassigned under structural modifications.

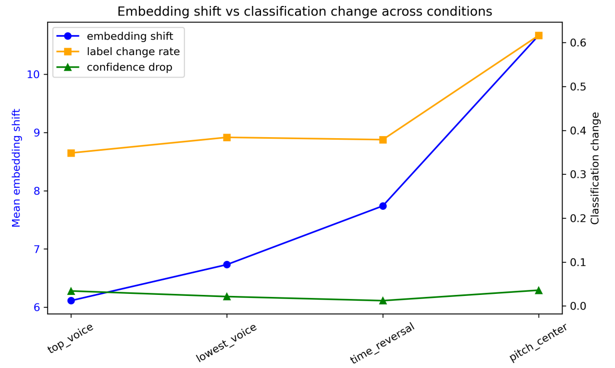


Figure 1: Overall effects of structural perturbations. The plot shows mean embedding shift $\overline{d}$, label change rate R_{change}, and confidence change $\overline{\Delta c}$ across perturbation conditions. Pitch normalization produces the largest label change rate, while representation shift and classification change do not vary in a strictly proportional manner.

This framework allows us to interpret how instrument identity is encoded in learned representations beyond accuracy-based evaluation.

4 RESULTS

4.1 Overall Trends Across Structural Perturbations

We first examine the overall effect of structural perturbations on both representation and classification behavior. Figure 1 shows the mean embedding shift $\overline{d}$ (Eq. 8), label change rate R_{change} (Eq. 7), and the mean confidence change $\overline{\Delta c}$ (Eq. 9) across different perturbation conditions.

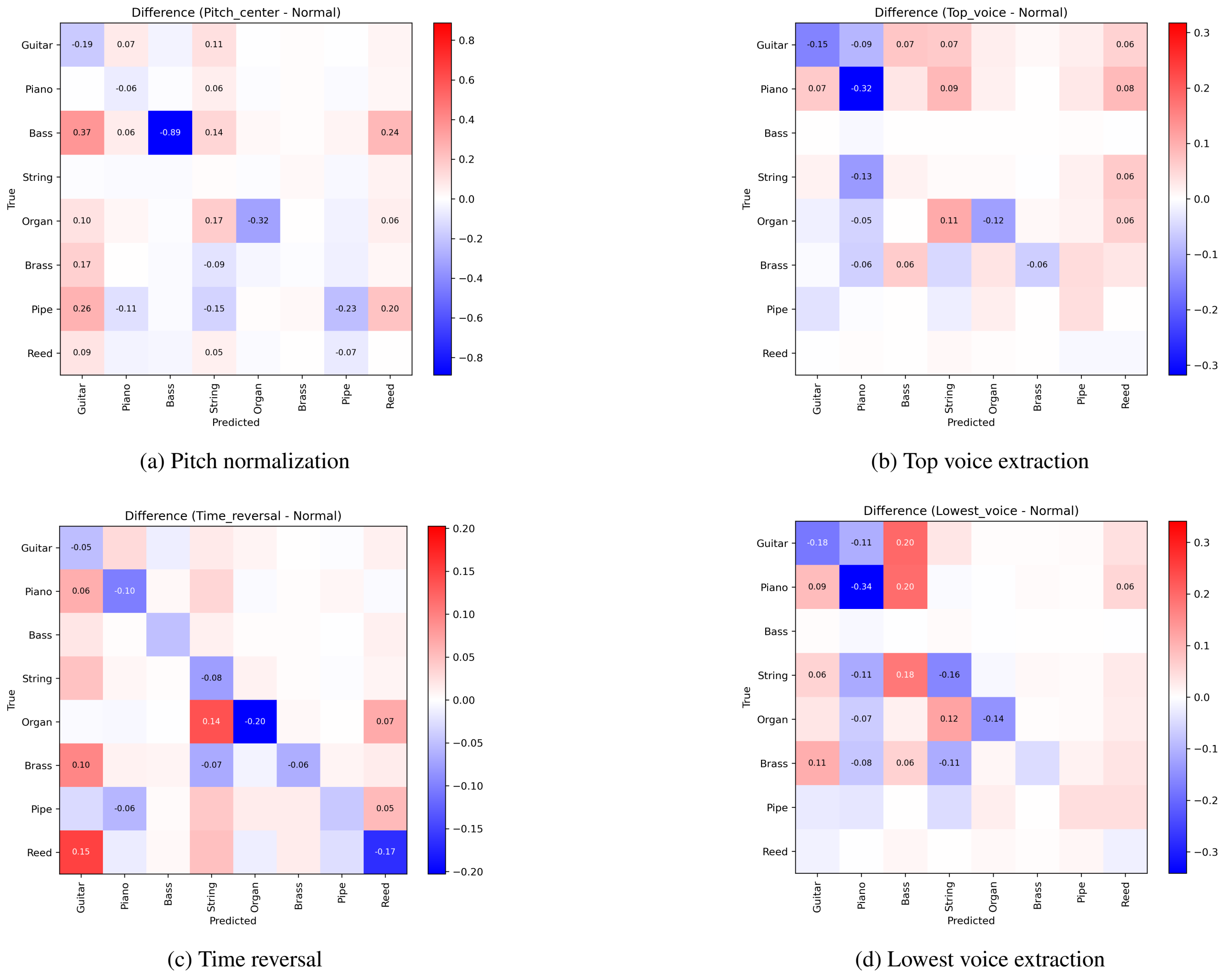
We observe that all perturbations affect the learned representations, although the magnitude of the embedding shift differs across conditions. However, the extent to which these changes affect classification behavior varies significantly across conditions. In particular, pitch normalization leads to a substantially higher label change rate compared to other perturbations, while time reversal has a smaller but non-negligible effect.

Importantly, the magnitude of representation change does not directly correspond to classification change. Although some perturbations produce comparable embedding shifts, their impact on predicted labels differs considerably. This suggests that classification behavior is not solely determined by the amount of representation change, but depends on which structural aspects of the input are modified.

4.2 Sensitivity to Structural Cues

To better understand this discrepancy, we analyze the effect of individual perturbations. Figure 2a compares confusion matrices before and after pitch normalization. We observe substantial changes in the distribution of predicted labels. In particular, certain instrument classes are consistently reassigned to specific alternative categories. This suggests that pitch-related structure plays a central role in determining instrument identity.

In contrast, Figure 2c shows the effect of time reversal. Despite causing noticeable changes in representation, time reversal has a relatively small impact on classification outcomes. The overall distribution of predicted labels remains more stable than under pitch normalization, suggesting that



(a) Pitch normalization

(b) Top voice extraction

(c) Time reversal

(d) Lowest voice extraction

Figure 2: Differences in confusion matrices between structural perturbations and the original input. Each matrix shows the change in prediction distribution after applying a perturbation. Positive values indicate increased predictions after perturbation, while negative values indicate decreased predictions. Pitch normalization produces the largest reassignment, especially for Bass. Voice extraction mainly affects Piano-related predictions, while time reversal produces smaller but non-negligible changes.

temporal direction is a weaker cue than pitch-related structure in this setting. However, the non-negligible representation shift indicates that temporal ordering still contributes to the learned representation.

Voice-related perturbations, namely top voice and lowest voice extraction, also induce structured changes in the predicted label distribution, as shown in Figures 2b and 2d. Compared with pitch normalization, these perturbations reduce polyphonic note configurations to a single voice, thereby removing information about chordal texture and simultaneous note stacking. As a result, instruments whose identity is partly characterized by polyphonic or chordal structures, such as Piano, Guitar, and Organ, tend to be reassigned to alternative categories. For example, both perturbations reduce predictions along the Piano diagonal, while increasing predictions for other classes such as Guitar, Bass, or String depending on the retained voice. This suggests that the model uses simultaneous note configurations as one cue for instrument classification, and that removing such information leads to structured category reassignment rather than uniform degradation.

Together, these results suggest that pitch-related and voice-related structures provide strong cues for instrument classification, but they do not fully account for model behavior. Although time reversal has a smaller effect than pitch normalization, it still changes the learned representations and, in some cases, the predicted labels. This indicates that instrument identity is not determined solely by static properties such as pitch range or polyphony, but also depends on the temporal organization of symbolic patterns.

These patterns indicate that perturbations do not simply degrade classification performance uniformly. Rather, they move inputs toward different regions of the classifier's decision space, resulting in structured reassignment across instrument categories.

4.3 Representation Shift vs. Classification Change

We next examine the relationship between representation changes and classification behavior at the sample level.

Across all perturbation conditions, label-changed and label-unchanged samples largely overlap in the d_i–Δc_i space. This indicates that classification decisions are not governed by representation distance alone. Figure 3 shows pitch normalization as a representative example, because it produces the highest label change rate among the perturbations.

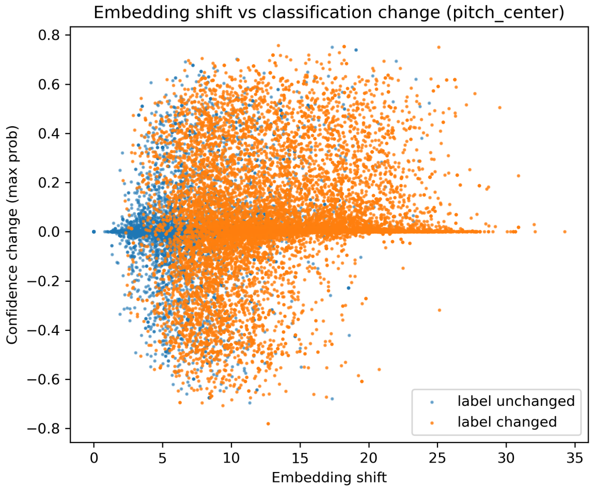


Figure 3: Embedding shift d_i and confidence change Δc_i under pitch normalization, shown as a representative perturbation condition. Samples are colored by whether the predicted label changes after perturbation. The overlap between label-changed and label-unchanged samples indicates that representation shift alone does not determine classification changes.

In this condition, label-changed and label-unchanged samples substantially overlap in the d_i–Δc_i space. Some samples exhibit large d_i values without changes in predicted labels, while others show label changes despite relatively small d_i values. This suggests that classification changes depend not only on the magnitude of representation shift, but also on the direction of the shift relative to the classifier's decision boundaries.

4.4 Summary of Findings

Overall, the results show that structural perturbations affect internal representations and classification behavior in different ways. Pitch- and voice-related cues strongly influence instrument classification, while temporal ordering also contributes to the learned representation. Moreover, label changes are not explained by representation shift alone, but reflect structured reassignment across instrument categories.

5 DISCUSSION

5.1 Instrumentality Beyond Timbre

Our results suggest that instrument identity in symbolic music can be approximated from structural properties of note sequences. This interpretation is consistent with affordance-based perspectives on musical instruments, which emphasize that instruments are not defined solely by their acoustic outputs, but also by the actions, constraints, and musical patterns they support (Mooney, 2011; Tullberg, 2022).

In particular, pitch register and simultaneous note configuration provide strong cues for classification, while temporal ordering also contributes to the learned representation. These structural cues can be interpreted as proxies for instrument identity: they are not explicit representations of timbre, but symbolic features that correlate with how instruments are typically used in musical data. For example, Bass is strongly affected by pitch normalization, suggesting that absolute register serves as an important cue for this category. Similarly, voice extraction reduces Piano-related predictions, indicating that simultaneous note configurations and chordal textures contribute to the model's interpretation of Piano-like patterns.

This does not imply that instrument identity in musical practice is independent of timbre. Rather, our results suggest that, when acoustic information is unavailable, the analyzed model relies on structural regularities as alternative signals for distinguishing instrument categories. From this perspective, the analyzed model infers instrument identity through structural patterns that are statistically associated with instrument labels in the training data, rather than through timbral information itself.

5.2 Reassignment Rather than Degradation

The difference matrices reveal that perturbations do not lead to uniform degradation. Instead, predicted labels shift toward specific alternative categories in a structured manner. This suggests that perturbations move inputs toward different regions of the classifier's decision space, rather than simply reducing confidence or increasing random errors.

For example, pitch normalization changes how Bass-like patterns are interpreted by removing absolute register cues, while voice extraction changes how Piano-like patterns are interpreted by removing simultaneous note configurations. These observations support the view that classification behavior reflects coherent category reassignment under structural modification.

5.3 Limitations and the Role of Musical Context

Our analysis is based on local symbolic segments and does not explicitly incorporate broader musical context. In particular, we do not account for factors such as long-term structure, ensemble interaction, or the functional role of instruments within a composition.

In musical practice, instrument identity is often shaped by such contextual factors. The same pitch pattern may be interpreted differently depending on its position in a piece, its relationship to other instruments, or its functional role (e.g., melody, accompaniment, or bass line). Therefore, the structural cues identified in this work should be understood as local approximations rather than complete determinants of instrumentality.

Another limitation concerns the model architecture. The classifier used in this study is based on a CNN-style architecture that processes piano rolls as two-dimensional patterns. Such models may emphasize local pitch-time configurations more strongly than global temporal direction. Therefore, the relatively weaker effect of time reversal should not be interpreted as evidence that temporal ordering is unimportant for instrument identity in general.

Future work could extend this analysis by incorporating longer temporal contexts or multi-track representations, allowing for a more direct investigation of how contextual and relational information contributes to instrument identity.

5.4 Implications for AI Music Research

These findings suggest several implications for AI-based music understanding and generation.

First, they indicate that a symbolic-domain classifier can rely on structural regularities present in the training data when inferring instrument identity. In this sense, the instrument identity captured by the model is shaped not only by the nominal instrument labels, but also by how instruments are represented and used in symbolic data.

Second, the distinction between representation shift and classification change suggests that changes in internal representations are not always reflected directly in predicted labels. This highlights the importance of analyzing both representation-level and decision-level behavior when interpreting music AI models.

Finally, our results suggest that model-based notions of instrumentality should be understood in relation to the data, representation, and architecture through which they are learned. Such analysis may contribute to the development of more interpretable and musically grounded AI systems.

6 CONCLUSION

In this paper, we investigated how instrument identity is approximated in a learned symbolic-domain instrument classifier when timbral information is unavailable. Rather than focusing on classification performance, we analyzed the behavior of the classifier under controlled structural perturbations.

Our results show that classification behavior is strongly influenced by structural cues such as pitch register and simultaneous note configuration, while temporal ordering also has a weaker but non-negligible effect. Furthermore, we found that changes in predicted labels are not directly explained by the magnitude of representation shifts. Instead, instrument categories are often reassigned in a structured manner, indicating that classification decisions depend on which structural aspects of the input are modified.

These findings suggest that, in the absence of timbral information, the analyzed classifier relies on structural regularities in symbolic patterns to infer instrument identity. From this perspective, instrumentality in the analyzed learned symbolic representation can be understood as a property shaped by data distributions and representational constraints.

At the same time, our analysis is limited to local symbolic representations and does not explicitly incorporate broader musical context such as long-term structure or ensemble interaction. Future work may extend this approach by integrating contextual and relational information, providing a more comprehensive understanding of how instrument identity is represented in music.

Overall, this work offers an empirical perspective on how instrumentality is approximated in a learned symbolic-domain instrument classifier, and highlights the importance of structural analysis for interpreting learned representations in AI music systems.

REFERENCES

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549.

Essid, S., Richard, G., and David, B. (2006). Musical instrument recognition by pairwise classification strategies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1401–1412.

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437.

Han, Y., Kim, J., and Lee, K. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):208–221.

Ji, K., Yang, D., and Tsai, T. (2021). Instrument classification of solo sheet music images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 546–550. IEEE.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529.

Manilow, E., Wichern, G., Seetharaman, P., and Le Roux, J. (2019). Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 45–49.

Mooney, J. (2011). Frameworks and affordances: Understanding the tools of music-making. *Journal of Music, Technology & Education*, 3(2–3):141–154.

Raffel, C. (2016). *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Columbia University.

Sawada, S. (2024). Symbolic-domain musical instrument classification using knowledge distillation from audio-teacher to symbolic-student. In *Proceedings of the 32nd European Signal Processing Conference (EUSIPCO)*, pages 191–195. IEEE.

Sawada, S. (2025). Contrastive audio-midi learning for symbolic-domain musical instrument classification. In *Proceedings of the 33rd European Signal Processing Conference (EUSIPCO)*, pages 426–430. IEEE.

Tullberg, M. (2022). Affordances of musical instruments: Conceptual consideration. *Frontiers in Psychology*, 13:974820.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833. Springer.

Zhong, Z., Hirano, M., Shimada, K., Tateishi, K., Takahashi, S., and Mitsufuji, Y. (2023). An attention-based approach to hierarchical multi-label music instrument classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.