

Leakage-Controlled Multimodal Prediction of Piano Performance Expressivity Using Sequential and Engineered MIDI Representations

Geping Jin

Shanghai Conservatory of Music, Shanghai, China

Dian Jin

jindian1233212026@163.com

École Normale Supérieure de Musique de Paris

Research Article

Keywords: Expressive piano performance, symbolic music processing, MIDI, multimodal learning, continuous regression, group-disjoint validation, explainable artificial intelligence

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11022159/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Leakage-Controlled Multimodal Prediction of Piano Performance Expressivity Using Sequential and Engineered MIDI Representations

¹ Geping Jin, ^{2*} Dian Jin

1. Shanghai Conservatory of Music, Shanghai, China
2. École Normale Supérieure de Musique de Paris, Paris, France

Corresponding Author: jindian1233212026@163.com

Abstract

Predicting expressivity in piano performance is difficult, because the perceptual judgement is the result of a combination of temporal, dynamic, articulatory, pedalling and tonal features. We propose a Residual Cross-Gated Dual-Branch MIDI Attention Fusion Network (RCG-DMAF-Net) to continuous perceptual rating prediction. The PercePiano dataset was remediated into 1202 performance segments, grouped into 99 composition-segment groups. Variable-length note-event sequences containing eight attributes and 51 engineered performance descriptors were used for each segment. The two higher level perceptual dimensions (Music Making Pure Expressive and Interpretation Doubtful Convincing) were modelled. For leakage of excerpts, the performance of the same segment of the composition stayed in one partition. The 244 observations from 20 groups without overlap were set aside as a locked evaluation set, not used for tuning and final model selection and 958 observations were used for four-fold group-disjoint validation and hyperparameter selection. The R^2 value for pure expressiveness was 0.4581, and 0.4765 for convincingness, respectively, while the RMSE value for the same was 0.0870 and 0.0908. The corresponding R^2 values for ExtraTrees were 0.4979 and 0.4930, but there was no significant difference between the two models at the 95% confidence level using the paired group-bootstrap intervals. This controlled ablation demonstrated that multimodal fusion greatly boosted sequence-only learning, while the differences between the fusion variants were relatively negligible. The main association SHAP made with the predictions were pedalling, velocity variation, onset structure, tempo and polyphony. The principal contribution is a leakage-aware evaluation of complementary sequential and engineered MIDI representations, together with an assessment of whether target-specific gated residual fusion provides an advantage over simpler fusion strategies.

Keywords: Expressive piano performance; symbolic music processing; MIDI; multimodal learning; continuous regression; group-disjoint validation; explainable artificial intelligence

1.Introduction

Expressive piano performance is shaped by coordinated variations in timing, tempo, articulation, dynamics, pedalling, note duration and phrasing. These differences enable each individual musician to create a different performance of the same music score. Such differences in perceptual attributes are difficult to predict with the computer since expressivity is not a directly measurable physical parameter nor can it be described simply with a low-level perceptual attribute. It grows out of the temporal relations between the notes that are played and larger statistical properties of an interpretation. Representations which capture local performance sequences as well as global musical behaviour are therefore required for reliable modelling. The symbolic representation has been used in computational music performance research as well as the audio representation. Audio-based methods will contain timbral and recording related features, while symbolic MIDI representations will allow direct access to the note onset, duration, pitch, velocity, inter-onset interval and pedal behaviour[1]. The value of MIDI is that it allows many aspects related to the performance to be separated from the recording situation and it offers the possibility of performing temporal analysis at the level of notes. Performance modelling and transcription studies have been enabled by large symbolic datasets like MAESTRO and PercePiano was motivated by the need to assess subtle perceptual aspects of piano repertoire. PercePiano offers annotations on several different perception levels, including timings, articulations, mood, musical character and interpretation, based on the expert's perception. The two modelling strategies mentioned here are particularly relevant to symbolic performance evaluation. Bidirectional long short-term memory networks and Transformers are sequential neural networks that model the relationships between the sequence of note events[2].

They can thus perceive expressive qualities of music that develop through changes in tempo, dynamic, length and pedalling. Reliable training with moderate size datasets is, however, challenging in the case of sequence models, and stable global features of a performance might be underutilized. Engineered descriptors, on the other hand, provide brief descriptions of aspects like tempo distribution, velocity variation, pitch range, polyphony, onset density or use of pedals. Tree ensembles are good at capturing non-linear relationships between these descriptors efficiently and

can provide relatively direct feature-level explanation, though aggregation can lose the temporal information that can differentiate expressive trajectories. One possible approach to integrating these complementary sources of information is known as multimodal or multi-representation learning. The local note-event evolution can be encoded in a sequential branch, and the global performance statistics in a tabular branch. However, just putting information in a row does not guarantee that the resulting series will be useful. Unrestricted fusion can add noise or overfitting, if a strong tabular predictor exists and explains a lot of the target variation. A more conservative approach is to have the engineered-feature pathway as the main prediction and the sequence pathway as the adaptively controlled correction, which is the residual formulation. Target-specific gating can also influence the degree to which sequential information influences each perceptual dimension. Another great concern of piano-performance modelling is the design of evaluation. Observations can share composition, score segment, melodic structure and harmonic content because more than one performer can play the same musical excerpt[3].

A split into a training and evaluation set may thus arise at a random observation level, where similar performances end up in both sets. The scores obtained might be somewhat due to recognition of the common excerpts, and not necessarily generalization to unknown music. To ensure all performances of a given composition segment are in a single partition it is thus necessary to include Group-disjoint validation. It is also important to keep selection of the model apart from the test; otherwise, a number of tests of test performance can result in over-optimistic estimates. Although previous studies have examined symbolic note-event sequences and engineered MIDI descriptors, comparatively little attention has been given to their complementary value under composition-aware evaluation. In particular, it remains necessary to determine whether multimodal fusion improves generalization to previously unseen musical excerpts when compared with strong tree-ensemble baselines and whether any improvement results from the multimodal representation itself or from a particular fusion mechanism. Controlled ablation is therefore required to determine whether any observed improvement arises from the sequential representation, engineered descriptors, multimodal combination, residual formulation, or gating mechanism. Statistical uncertainty must also be reported because small numerical differences between models may not represent reproducible performance differences. Accordingly, this study formulates piano-performance expressivity assessment as continuous multi-output regression rather than converting the perceptual ratings into artificial classes. Two higher-order PercePiano

dimensions - Music Making Pure Expressive and Interpretation Doubtful Convincing are considered as the main outcomes. There are 1,202 performance segments each represented by an eight attribute MIDI note-event sequence and 51 engineered performance descriptors[4].

The proposed RCG-DMAF-Net is a Residual Cross-Gated Dual-Branch MIDI Attention Fusion Network, which is composed of a bidirectional LSTM attention Encoder (LA) and a tabular neural pathway. The tabular branch generates the primary estimate, while a target-specific gate controls the magnitude of a sequence-informed residual correction. This study makes four contributions. First, it develops a dual-representation framework combining temporally ordered MIDI note events with 51 interpretable global performance descriptors. Second, it evaluates a residual cross-gated fusion architecture in which the sequence representation provides a target-specific correction to a tabular prediction. Third, it applies composition-segment-disjoint development and evaluation partitions to prevent performances of the same musical excerpt from appearing in both training and evaluation data. Fourth, it compares the proposed architecture with strong linear, tree-ensemble, recurrent, Transformer, single-task fusion, and multi-task fusion baselines using controlled ablation, paired group-bootstrap analysis, and SHAP-based interpretation. The study is not designed to imply that a neural fusion model should be better than any other traditional model. Rather, it examines the possibility of using a combination of sequential and engineered MIDI representation to verify and interpret perceptual prediction within a strict generalization protocol. This distinction is relevant for moderate sized music sets where tree ensembles can be a strong contender and margins of performance can be small, and thus care must be taken when interpreting them[5].

The overall flow of the study is illustrated in **Figure 1**, ranging from dataset reconstruction to generation of MIDI sequences and engineered features, group-disjoint partitioning, development-only tuning, evaluation of frozen model on the locked blind set, ablation analysis and model interpretation. In Section 2, the data set, its representations, proposed architecture and evaluation protocol are described. Section 3 presents the methodology. The comparative, ablation, statistical and explainability results are displayed in Section 4. The future research and limitations are discussed in Section 5 and the study concluded in Section 6.



Figure 1. Leakage-controlled workflow for multimodal piano performance expressivity prediction

2. Literature Review

2.1 Computational Assessment of Expressive Piano Performance

Expressive music performance research investigates the process used by musicians to create an interpretive performance of a notated score using timing, dynamics, articulation, pedalling and phrasing. The early computational studies concentrated primarily on quantifying the difference between events performed and notated or on modelling rules of performance-rendering. Later studies gradually came to the conclusion that technical accuracy is not satisfactory for describing the listeners' impressions of the performance. Expressiveness, tension, musical character and interpretative conviction are higher level concepts that need perceptual annotations and models to represent interacting performance cues. Expressive music performance modelling has been reviewed and the importance of musical expression being hierarchical and context dependent has been stated. Note-level deviations can take significance when they interact with the surrounding notes, phrases and structural elements. So, no individual note marking or single overall score is a good representative of an interpretation. Perceptual evaluation is also an inherently variable

process due to the fact that there can be several musically valid interpretations to a given score. Estimating reconstruction error from one performance might thus, when listeners hear it in terms of wider interpretative features, only be part of an assessment of reconstruction quality or similarity[6].

There are now many more data that can be used for data-driven analysis, with aligned symbolic-performance datasets. At scale, MAESTRO offers aligned audio and MIDI recordings, while also enabling research in the field of piano transcription and representation learning. But its metadata didn't focus heavily on subtle perceptual judgments. PercePiano tackles this drawback by offering ratings of piano performances, based on expert judgments, on 19 perceptual aspects, including the timing, articulation, pedalling, mood, musical character and interpretation. The data has 1,202 annotated performance segments and allows the continuous prediction of perceptual attributes, not just the detection of technical errors or artificially created categories. The present study mainly deals with the following, which are considered as comparatively high-level judgments of musical communication and interpretation: Music Making Pure Expressive and Interpretation Doubtful Convincing. They are harder to forecast than other measurable and perceptible attributes like tempo or note duration, which are represented by a single value. Their prediction, therefore, offers a good test of the complementarity of local MIDI sequences and global performance descriptors.

2.2 Sequential Modelling of Symbolic Music

Symbolic music can be described as a sequence of events, consisting of pitch, onset, duration, velocity, tempo and pedal. This representation retains the temporal organization without some of the sources of variation that would arise from the recording environment, microphone placement, instrument acoustics. Sequence models are able to process note events in the order they occur during a performance, and learn some contextual relationships which are not kept by mere aggregation. Music generation, rendering music performances, music transcription and music sequence classification have been among the wide range of applications involving recurrent neural networks and long short-term memory networks. When a whole performance segment is available, bidirectional LSTMs are relevant as they incorporate contextual information from both past and future note event. For expressive-performance assessment, patterns of acceleration, onset delay, dynamic changes in articulation, and coordinated changes in articulation can, in theory, be learned

in the recurrent models. However, their performance is dependent on the size of the dataset, the construction of the sequences, the choice of regularization and how variable length sequences are handled[7].

Recurrent encoding can be supplemented by attention mechanisms, so as to assign different importance to events. Temporal attention creates a weighted representation of the sequence of the multiple positions rather than a compressed representation of a entire performance in the final recurrent state. This can be helpful if perceptually important events are clustered at phrase boundaries, accented notes, or changes of pedalling/dynamics. However, attention weights cannot be taken for granted to be a causal explanation, and sequence models can still fail to do well if there are relatively few independent musical excerpts available. There are other ways of modelling long range musical dependency, namely through transformer architectures. Event-specific tokenization and self-attention have been adopted to learn structural relationships from large corpora in symbolic-music systems[8]. When a large amount of pre-training data is available, transformer-based models have demonstrated their ability to understand music well. However, they can have a larger variance and computational cost for smaller supervised data sets. Therefore, it is informative to compare the performance of the recurrent, attention based and Transformer models, but differences should be understood in context of dataset size and representation[9].

2.3 Engineered MIDI Descriptors and Ensemble Learning

Engineered performance descriptors are summaries of musically interpretable properties of the MIDI data. Descriptors of the timing can be such as tempo statistics, inter-onset intervals, onset density variability and onset duration variability. Velocity mean, velocity range, velocity standard deviation, velocity quantiles and velocity CoV can be used to represent dynamic behaviour. Tonal and textural properties are characterized by pitch range, pitch-class entropy, polyphony and note density; properties of sustain-pedal use are characterized by pedal down fraction and by the count of pedal events. These are compact global representations that lose the fine temporal order, but are suitable for moderate-sized datasets. Linear regression is an easily interpretable reference model, but cannot represent complex interactions unless these interactions are explicitly modeled. With relatively simple preprocessing, Random Forest and ExtraTrees models are able to capture nonlinear relationships and feature interactions. XGBoost incorporates sequential gradient-based tree construction and regularization, and serves as a solid reference of structured music descriptors.

Tree ensembles are important comparators for neural models because when the number of independent groups is small, and tabular representation already summarizes informative domain knowledge, deep learning is not necessarily superior. They can also be analysed using feature-importance methods. While they indicate relationships between features and model outputs, they are not causal musical mechanisms, but rather local and global measurements of this relationship. Feature interpretation in expressive-performance research is particularly useful as it allows them to link the model behaviour with identifiable elements of timing, velocity, pedalling, pitch and texture. Limitations of global descriptors are that the same summary statistics can be generated from different temporal trajectories. Two performances may be of the same average tempo and the same velocity variation, but be different in the locations of the points of emphasis and/or acceleration and/or pedalling. The role of engineered features and note-event sequences should thus be seen as complementary rather than alternative[10].

2.4 Multimodal Fusion for Performance Prediction

Multimodal learning refers to learning representations that represent various aspects of a single observation. A sequence encoder may be used to model temporal evolution of performed notes in piano-performance assessment, a tabular encoder may be used to represent global statistical properties. Usually, early fusion joins the representations prior to the prediction. This is a very simple method, but it requires that both modalities are always represented in the same way and it may enable a weak representation to be affected by a strong one even if the latter is noisy. Adaptive fusion: learned weights/gates to control contribution of each representation[11]. Concatenation, element-wise multiplication, and absolute differences between latent representations are all suitable ways of expressing cross-modal interactions. In addition, a residual design offers one more constraint: The second branch of the network learns only a correction to the prediction made by the first branch. This structure is suitable where there is a good baseline provided by an engineer and it is believed that information to be added sequentially will contribute selectively. Multi-output perceptual modelling is also relevant for the target-specific gates since the timing, mood and interpretation may vary to different extents depending on the sequence information. Previous work on expressive-performance systems has used recurrent networks, score-aligned representations and perceptual features, and more recent systems based on symbolic music have considered the use of pre-trained contextual representations. But, less work has been done to investigate the

viability of a sequence-derived correction in addition to a well-designed engineered-feature predictor with a common leakage-controlled protocol. Furthermore, comparing the sophisticated fusion models with classical ensembles is still essential as the generalization ability does not necessarily increase with the complexity of the architecture.

2.5 Validation, Leakage and Interpretability

For certain composition sections there may be more than one performer; validation procedures are important then. Common pitch sequences, harmonic progressions, phrase structures and durations are found in performances of a shared excerpt. A model can use the identity of the excerpt—rather than learning the general relationship between the behaviour of a performance and a perception—when the excerpt is randomly distributed between train and test sets. Consequently, optimistic estimates may be obtained using observation-level random splitting. A validation that is more challenging is called piece-disjoint or composition-segment-disjoint validation, which requires that related performances stay in one partition. In the current context, the grouping of the composition segments would be more appropriate; the main dependency structure being formed by the same local excerpt but played by different pianists. When developing, use group-disjoint cross-validation, and also keep a locked set of groups that have not been seen before when selecting hyperparameters. When the performance of two models is close, uncertainty estimation is also of importance. Spending just one number of R^2 or a single number of rmse can make it seem that there is an assurance to the data that isn't warranted. Group bootstrap procedures maintain the dependency structure by resampling whole groups of composition segmentation and can yield confidence intervals for individual metrics and for model differences. Controlled ablation is also necessary to determine if improvement occurs due to the sequence branch, tabular branch, fusion mechanism or residual correction/gating.

2.6 Research Gap

Perceptual annotation and symbolic sequence learning and engineered performance descriptors have proven their worth in the literature, but four gaps can be identified. First, many studies report the results of either sequential or aggregated representations but not their complementary role. Second, in some cases, there is not a proper tree ensemble set to be compared with high-performance neural architectures. Third, too, splitting the observations randomly may not capture

common musical excerpts. Fourth, standard architectural claims are often constructed on point estimates and without group-aware uncertainty analysis and controlled ablation. RCG-DMAF-Net is proposed as the solution of the above issues in this study, which is called Residual Cross-Gated Dual-Branch MIDI Attention Fusion Network. The model is designed to combine an attention-weighted bidirectional LSTM representation of a note-event sequence with a neural representation of 51 engineered note descriptors. The primary prediction is made through a tabular pathway, while a target-specific gate regulates a sequence-informed residual correction. The linear model, tree-ensemble model, recurrent model, Transformer model and alternative fusion model are evaluated against the architecture by composition-segment-disjoint development folds, and a locked blind set. Two-step analysis of numerical performance and statistically supported results, and of interpretable musical associations, is achieved by pairing with groups of bootstraps and by component ablation and SHAP analysis. **Table 1** shows the comparison of the modelling approaches that have been evaluated in this study.

Table 1. Comparison of the modelling approaches evaluated in this study

Approach	Input representation	Principal strength	Principal limitation	Role in this study
Ridge regression	Fifty-one engineered MIDI descriptors	Transparent linear reference with low computational cost	Cannot capture complex nonlinear relationships without explicit transformations	Linear baseline
Random Forest	Engineered MIDI descriptors	Robust nonlinear ensemble with implicit feature interactions	Aggregation removes temporal ordering	Tree-ensemble baseline
ExtraTrees	Engineered MIDI descriptors	Strong variance reduction through highly randomized trees	Does not directly model note-event sequences	Leading structured-data baseline
XGBoost	Engineered MIDI descriptors	Regularized boosting with nonlinear interaction modelling	Sensitive to tuning and limited to fixed-length inputs	Gradient-boosting baseline

Approach	Input representation	Principal strength	Principal limitation	Role in this study
Bidirectional LSTM	Eight-attribute MIDI note-event sequences	Models ordered contextual dependencies in both temporal directions	May underperform with limited independent musical groups	Recurrent sequence baseline
Transformer	MIDI note-event sequences	Represents long-range dependencies using self-attention	Data and computational requirements can be high	Attention-based sequence baseline
DMAF-Net	MIDI sequences and engineered descriptors	Learns adaptive single-task multimodal fusion	Fusion can introduce noise when one representation dominates	Single-task fusion comparator
MT-DMAF-Net	MIDI sequences and engineered descriptors with 19 outputs	Shares information across perceptual dimensions	Susceptible to negative transfer among heterogeneous targets	Multi-task fusion comparator
RCG-DMAF-Net	MIDI sequences and engineered descriptors	Uses tabular primary prediction with target-specific gated residual correction	Additional complexity does not guarantee superiority over strong tree ensembles	Proposed model

3. Methodology

3.1 Study Design and Data Source

In this study, two complementary symbolic representations of piano performance – temporally ordered MIDI note events and engineered MIDI descriptors of fixed length – were used to investigate the continuous prediction of perceptual attributes of piano performance. The experiments were conducted with the openly available dataset PercePiano developed for expressive piano performance evaluation by experts. The source repository was obtained from the official PercePiano GitHub repository and committed at e672299b8661f63802cd5cd73bb3c76218a13029 in order to facilitate the ability to reconstruct the

dataset from an immutable version. A total of 1,202 annotated performance segments from 22 musical works were reconstructed. Each observation represents one pianist’s performance of a particular segment of a musical composition. The same composition-segment identifier was given to performances of the same composition and segment number. This procedure created 99 composition-segment groups which enabled related performances to be clustered when data was partitioned. Twenty-two performer identifiers were added to the metadata and were not used as predictors. There were no missed MIDI files, no missing target values and no observations for which no MIDI file was found or extracted from[12].

All 1,202 observations were successfully matched to their respective file. PercePiano offers normalized perceptual ratings for 19 bipolar attributes of timing, articulation, pedalling, timbre, dynamics, musical character, mood and interpretation. All 19 targets were included as part of the exploratory multi-task modelling of the present study, but two high-level dimensions were selected as primary outcomes: Music Making Pure Expressive and Interpretation Doubtful Convincing. These dimensions were chosen prior to RCG-DMAF-Net hyperparameter optimization as they represent general perceptions which are unlikely to be explicitly defined by a single directly observable MIDI property. The Dataset composition and primary prediction targets is shown in **Table 2**. The original continuous ratings were preserved. They did not get converted to artificial classes. For Music Making Pure Expressive, the range was from 0.1786 to 0.9048 with a mean of 0.5846 and for Interpretation Doubtful Convincing the range was from 0.1714 to 0.9643 with a mean of 0.5603. The target ratings were normalized to the interval used in the released PercePiano dataset. Higher values for Music Making Pure–Expressive indicate greater perceived expressiveness, whereas higher values for Interpretation Doubtful–Convincing indicate greater perceived interpretative conviction[13].

Table 2. Dataset composition and primary prediction targets

Component	Value
Annotated performance segments	1,202
Musical works	22
Composition-segment groups	99

Component	Value
Performer identifiers	22
Available perceptual targets	19
Primary targets	2
Engineered MIDI descriptors	51
Note-event attributes	8
Maximum padded sequence length	512 events
Minimum observed sequence length	51 events
Median observed sequence length	106 events
Maximum observed sequence length	413 events
Truncated sequences	0
Missing MIDI files	0
Feature-extraction failures	0

3.2 Prediction Targets

The primary task was formulated as two-output continuous regression. For observation i , the target vector was defined as

$$\{y\}_i = \left[y_{\{i, \{expressive\}\}}, y_{\{i, \{convincing\}\}} \right]^{\{T\}}$$

The values $y_{i, expressive}$ and $y_{i, convincing}$ represent the normalized Music Making Pure Expressive and Interpretation Doubtful Convincing scores, respectively. Two continuous outputs were generated from the models specifically created for the primary analysis. MT-DMAF-Net was also trained on all 19 PercePiano dimensions to investigate if multi-task learning was beneficial to the two main outcomes. The continuous ratings were not lost due to the construction of classes based on thresholds. It also enabled the prediction error to be used to represent the size of the

discrepancy between actual and perceived perceptual scores, instead of assuming that all errors were equal class mismatches.

3.3 Input Representation

3.3.1 MIDI Note-Event Sequences

All the MIDI files were read with PrettyMIDI. No drums and all other instrument notes were grabbed and sorted by onset time, pitch, and offset time. Eight attributes were used for each note event: 1. MIDI pitch / 127; 2. MIDI velocity / 127 (Normalized); 3. number of notes. 4. natural log of 1 plus time since last onset; 5. Pitch interval of the preceding note / 24; 6. Position in time of note onsets in relation to the duration of the segment; 7. Mix in binary sustain-pedal state at note onset; and 8. binary chord-onset indicator. The Sustain-pedal state was obtained from the midi cc number 64. The notes that started at a point before 30 ms were considered more or less simultaneous for the construction of the chord onset indicator. Each sequence was padded to a maximum length of 512 events, and its true length and binary padding mask were retained. No sequence was truncated because the observed sequence lengths ranged from 51 to 413 events. Sequence features were standardized separately within each training fold. The means and standard deviations were calculated using only valid training tokens and were subsequently applied to the corresponding validation or evaluation observations. Padded positions were reset to zero after standardization, preventing information from validation or evaluation groups from influencing preprocessing[14].

3.3.2 Engineered MIDI Descriptors

A second representation was created to abstract each performance with 51 fixed-length MIDI descriptors. Each of these variables was directly derived from the associated MIDI file, and grouped into six musically interpretable sets. Engineered MIDI feature groups are shown in **Table 3**.

Table 3. Engineered MIDI feature groups

Feature group Number		Included descriptors
General temporal structure	5	Segment duration, note count, unique-onset count, note density and onset density

Feature group Number		Included descriptors
Pitch and polyphony	12	Pitch-class entropy, maximum and mean polyphony, chord-onset ratio, and the mean, standard deviation, median, minimum, maximum, lower quartile, upper quartile and range of pitch
Pedalling and tempo	7	Pedal-event count, pedal-down fraction, tempo-change count, and the mean, standard deviation, minimum and maximum of tempo
Velocity	9	Mean, standard deviation, median, minimum, maximum, lower quartile, upper quartile, range and coefficient of variation
Note duration	9	Mean, standard deviation, median, minimum, maximum, lower quartile, upper quartile, range and coefficient of variation
Inter-onset interval	9	Mean, standard deviation, median, minimum, maximum, lower quartile, upper quartile, range and coefficient of variation
Total	51	All engineered MIDI descriptors

For neural fusion models, each tabular variable was normalized using parameters of the corresponding training partition. Median imputation and standardisation were used in ridge regression. No missing or infinite values were found in the extracted feature matrix. The engineered descriptors were provided to the tree-based models on the same numerical scale as they were.

3.4 Leakage-Controlled Data Partitioning

A number of performers performed the same piece of the composition. The observations have a score content, melodic structure, harmonic progression and approximate phrase organization; thus these observations cannot be considered as completely independent of each other. In all of the model development and evaluation, `composition_segment_group` was employed as a grouping variable to prevent the performance of the same excerpt from being included in different partitions. The first method was based on the approximate target stratification performed using a deterministic five-way `StratifiedGroupKFold` procedure with quintiles of Music Making Pure Expressive. The first method was based on approximate target stratification by applying a deterministic five-way `StratifiedGroupKFold` procedure using quintiles of the Music Making Pure Expressive. Fold 0 was

designated as locked holdout set and the other four partitions were the development set. For this procedure random seed 2026 was used.

The development set had 958 observations from 79 composition-segment groups, the locked set had 244 observations from 20 groups. There was no overlap in the groups. The observations of the development set were then split into 4 folds using StratifiedGroupKFold with a random seed set to 2027, ensuring that the folds were disjoint sets based on the observation's composition. Fold sizes were 249, 224, 253 and 232 observations. All of the composition-segment groups were present in a single fold of development. The locked performance identifiers were stored with a SHA-256 fingerprint to prevent any unexpected change in membership. The locked partition was not used for hyperparameter optimization or stop epochs or stopping ensemble configuration of RCG-DMAF-Net and was fixed before the hyperparameter optimization. It was opened for one final comparison of the frozen RCG-DMAF-Net ensemble and classical baselines (fixed). The final results were not used to add tuning or model selection[15].

3.5 Comparative Models

3.5.1 Classical Regression Baselines

Five regression baselines with constant length were tested. DummyMean predicted the mean training target and gave an uninformative comparison. For ridge regression, $\alpha = 1.0$ was used for the l2 regularization. Random Forest and ExtraTrees were both based on 500 trees, square-root feature sampling and a minimum leaf size of 2. XGBoost had 500 estimators, 0.03 learning rate, four maximum tree depth, three minimum child weight, 0.8 row subsampling, 0.8 column subsampling, 0.05 L1 regularization and 1.0 L2 regularization. These models were added as good baseline models for medium size structured data sets[16].

3.5.2 Bidirectional LSTM

The sequence-only recurrent baseline made a 32-dimensional representation by applying a linear layer with a layer normalization and rectified linear activation to the eight input attributes. Packed variable length sequence was encoded in a single bidirectional LSTM layer with 64 hidden units on each direction. The last forward and backward hidden states were connected and fed into a 30% dropout, 64-unit hidden layer, rectified linear activation, 20% dropout and a scalar output regression head. Two main targets were separated by training two BiLSTM models.

3.5.3 Transformer

The Transformer baseline was used to map each note event to a 64 dimensional latent space. The positions were embedded, and the following two Transformer encoder layers contained four attention heads and a feed-forward dimension of 128. The dropout was to be 0.20. The padded events were excluded from padding masks and the mean pooling for these events was masked to create the segment level representation fed into the regression head. Two targets were predominant and separate models were trained for each.

3.5.4 DMAF-Net

The Dual-Branch MIDI Attention Fusion Network is a fusion of sequence and tabular representations. It was followed by a bidirectional LSTM with 48 hidden units in each direction and temporal attention, as well as its sequence branch predicting the 8 note attributes to 32 dimensions. The attention-weighted context was projected down to 64 dimensions. The tabular branch performed 128 and 64 unit layers of tabular data with layer normalization, GELU activation and dropout to transform 51 engineered descriptors. A learned feature-wise sigmoid gate was used to combine the two 64-dimensional representations.

$$h_f = g \odot h_s + (1 - g) \odot h_t$$

where h_s and h_t denote the sequence and tabular representations, g is the learned fusion gate, and $\odot$ denotes element-wise multiplication. Separate single-output DMAF-Net models were trained for the two primary targets.

3.5.5 Multi-Task DMAF-Net

MT-DMAF-Net adopted the same sequence and tabular encoders as DMAF-Net, but gave the fused representation to all 19 perceptual dimensions. There was a shared 64 unit layer followed by a 19 output regression head. The goal was the average Smooth L1 loss over the perceptual targets. This model examined the hypothesis that there is shared supervision from lower- and higher-level perceptual attributes which would enhance the two main outcomes[17].

3.6 Proposed RCG-DMAF-Net

The proposed Residual Cross-Gated Dual-Branch MIDI Attention Fusion Network was devised for scenarios that have a good initial prediction from engineered descriptors, and a second prediction that is more detailed, but only relevant to a subset of the items. The sequence branch

first expanded each note event of the sequence to 32 dimensions, and then fed this expanded sequence into a one-layer bidirectional LSTM to process. The attention score for note event j was computed as

$$e_{i,j} = w_a^T \tanh(W_a h_{i,j} + b_a)$$

where $e_{i,j}$ is the attention score for note event j in performance i , and $h_{i,j}$ is its BiLSTM hidden representation and the normalized attention weight was

$$a_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{L_i} \exp(e_{i,k})}$$

where L_i is the valid sequence length of performance i . The summation therefore excludes padded events. The sequence context was calculated as

$$h_{s,i} = \sum_{j=1}^{L_i} a_{i,j} h_{i,j}$$

where $a_{i,j}$ is the normalized attention weight and $h_{i,j}$ is the BiLSTM hidden representation for event j and projected to the fusion dimension. The tabular encoder transformed the 51 standardized descriptors into a latent representation h_t . A dedicated tabular prediction head generated the primary two-target estimate

$$\hat{y}_{t,i} = f_t(h_{t,i})$$

where $h_{t,i}$ is the tabular representation of performance i , and $f_t(\cdot)$ is the tabular prediction head. Cross-modal interaction was represented using the sequence vector, tabular vector, element-wise product and absolute difference

$$h_{c,i} = f_c([h_{s,i}, h_{t,i}, h_{s,i} \odot h_{t,i}, |h_{s,i} - h_{t,i}|])$$

where the square brackets represent vector concatenation and $\odot$ denotes element-wise multiplication. The interaction representation produced a two-dimensional residual correction

$$r_{i,k} = f_{(r,k)}(h_{c,i})$$

where $r_{i,k}$ is the residual correction for performance i and perceptual target k , and $f_{(r,k)}(\cdot)$ is the target-specific residual head. A separate gate was learned for each target

$$g_{i,k} = \text{sigmoid}(f_{(g,k)}([h_{s,i}, h_{t,i}]))$$

where $g_{i,k} \in (0,1)$ controls the contribution of the sequence-derived residual correction for performance i and target k . The final prediction was

$$\hat{y}_{i,k} = \hat{y}_{t,i,k} + \lambda(g_{i,k} \cdot r_{i,k})$$

where λ is a tuned residual-scale coefficient. This formulation allows the tabular pathway to retain its primary prediction while sequence information modifies each target only to the extent permitted by its learned gate.

The model selected had 48 LSTM hidden units per direction, 96 dimensional fusion space, and 64 units tabular hidden layer. The entire architecture had 96,839 trainable parameters. Figure 2 should show the architecture of the two inputs to the sequence and tabular encoders, cross-interaction block, target-specific gate, residual correction and final outputs.

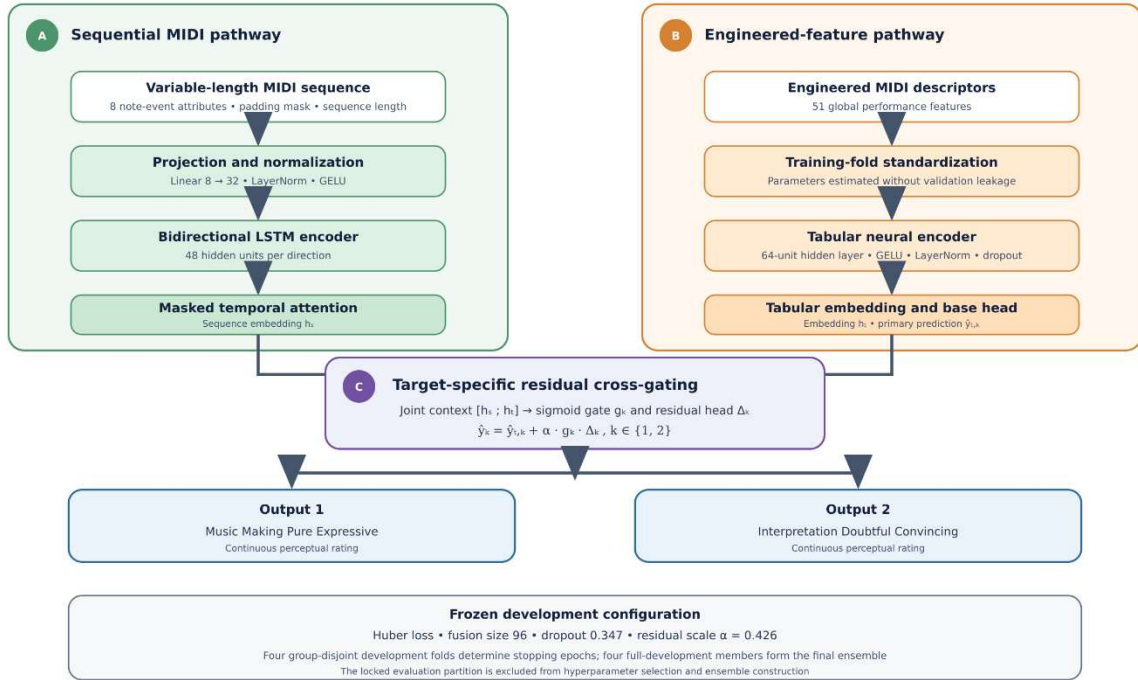


Figure 2. Architecture of the proposed Residual Cross-Gated Dual-Branch MIDI Attention Fusion Network

3.7 Hyperparameter Optimization and Training

The hyper-parameters of RCG-DMAF-Net were optimized via Optuna with Tree-structured Parzen estimator and Median pruning. The search goal was the average RMSE over the two main targets and four development folds. This is 958 development observations that were used only for 12 trials. The trials were given a maximum of 70 epochs, patience was set to 10 epochs and batch

size to 32. The parameters in the hyperparameter configuration are illustrated in the following **Table 4**.

Table 4. RCG-DMAF-Net hyperparameter search space and selected configuration

Hyperparameter	Search space	Selected value
LSTM hidden units per direction	32, 48, 64, 96	48
Fusion dimension	32, 64, 96	96
Tabular hidden units	64, 128, 192	64
Dropout	Continuous 0.10–0.40	0.34724
Learning rate	Log-uniform 1×10^{-4} – 2×10^{-3}	1.107×10^{-3}
Weight decay	Log-uniform 1×10^{-6} – 5×10^{-3}	6.136×10^{-6}
Residual scale λ	Continuous 0.10–1.00	0.42569
Loss function	Mean squared error or Smooth L1	Smooth L1
Optuna trials	Fixed	12
Batch size	Fixed	32
Maximum tuning epochs	Fixed	70
Early-stopping patience	Fixed	10

The optimization of the neural network was done using AdamW. A ReduceLROnPlateau scheduler reduced the learning rate by half after five epochs if no improvement was witnessed on the validation. Maximum norms of gradients 1.0 were applied. For the final development models, the maximum number of epochs was 100 and the patience was 15 epochs for early stopping. The epochs they chose were 38, 37, 8 and 20 for folds 0–3, respectively. Hyperparameters and epochs per fold were frozen and four members of the ensemble were trained on all the development observations. The time they have reached coincided with the four values they had been developed with. The four member predictions were averaged to receive locked-set evaluation. The neural

predictions were limited to the interval of valid normalized-rating scores (between 0 and 1). The configurations of the comparison models is shown in **Table 5**.

Table 5. Fixed configurations of the comparison models

Model	Main configuration	Loss or objective	Maximum epochs
Ridge	$\alpha = 1.0$ with standardized descriptors	Squared error with L2 penalty	Not applicable
Random Forest	500 trees, max_features = sqrt, minimum leaf = 2	Ensemble regression	Not applicable
ExtraTrees	500 trees, max_features = sqrt, minimum leaf = 2	Ensemble regression	Not applicable
XGBoost	500 estimators, depth = 4, learning rate = 0.03, subsample = 0.8, column subsample = 0.8	Squared-error regression	Not applicable
BiLSTM	32-dimensional projection and 64 hidden units per direction	Smooth L1	100
Transformer	d_model = 64, four heads, two layers and feed-forward size 128	Smooth L1	100
DMAF-Net	48 hidden units per direction and fusion size 64	Smooth L1	120
MT-DMAF-Net	48 hidden units per direction, fusion size 64 and 19 outputs	Mean Smooth L1	150
RCG-DMAF-Net	Selected Optuna configuration in Table 4	Smooth L1	100

3.8 Controlled Ablation Analysis

A development-only ablation study was performed using five configurations, but the same optimization settings were used. Tabular only (with the engineered-feature encoder and primary prediction head); 2. Use sequence only and embedded with projected BiLSTM-attention representation; 3. Simple fusion (concatenated sequence, tabular); 4. Ungated residual fusion, tabular prediction and an unrestricted sequence-informed residual; and Full RCG-DMAF-Net,

with gating of the residual correction using targets. The training of each configuration was repeated over the four development folds, using a maximum of 80 epochs, training patience of 12 and batch size 32[18].

No locked observations were used. The number of active parameters was obtained for each forward path based on the number of modules involved in each path and not the number of modules in the common ablation wrapper. Counts obtained were 13,058 for tabular only, 49,219 for sequence only, 74,851 for simple fusion, 96,453 for ungated residual fusion and 96,839 for full RCG-DMAF-Net.

3.9 Explainability Analysis

ExtraTrees was chosen as the baseline structured data model for which to perform SHAP analysis due to its best performance and ability to provide efficient explanations of the tree. All 958 development observations were used to train separate ExtraTrees models for the two main targets. Seed 31415 was used to calculate SHAP values for a representative set of 500 development observations. The mean absolute SHAP values were used to highlight the importance of features across the world and the beeswarm plots helped visualize the contribution of features and their distribution. It is important to emphasize that SHAP results were not considered as evidence of causal mechanisms in music but as model-specific predictive associations. Explanation generation did not take place with locked-set observations.

3.10 Evaluation Metrics

The coefficient of determination, mean absolute error, root mean squared error, Pearson correlation and Spearman rank correlation were used to assess the performance of regression.

For n observations with observed values y_i , predictions $\hat{y}_i$ and observed mean $\bar{y}$, the coefficient of determination was calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Mean absolute error was

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

and root mean squared error was

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Pearson correlation was used to measure linear association between observed and predicted ratings while Spearman correlation was used to measure rank-order agreement. R^2 was used as the primary measure of goodness of fit; the primary measure of error for optimization and paired comparisons was RMSE.

3.11 Statistical Uncertainty and Model Comparison

Uncertainty was estimated as the resampling of whole composition-segment groups instead of individual observations. Confidence intervals for the model metrics and paired RMSE differences were obtained from 2,000 group- bootstrap replications for the first group-disjoint benchmark. It has 20 composition-segment groups and 5,000 pairs of bootstrap samples were drawn for the final locked-set comparison. In each of the two models, the same groups were sampled which ensured paired predictions and within-group dependence.

The RMSE difference and corresponding 95% confidence interval (CI) were reported for each comparison. A difference was deemed to be statistically significant if the whole interval extended above or below the value of zero. Also, the proportion of the bootstrap repetitions favouring RCG-DMAF-Net was computed as the probability that the proposed model resulted in lower RMSE. Separate 5000-repetition paired group bootstrap was used to assess the control ablation predictions across the 79 development groups. This analysis separated numerical differences among variants of the fusion from differences supported by the available grouped data.

3.12 Reproducibility

All the preprocessing, model development and evaluation were performed in Python with the help of NumPy, pandas, scikit-learn, XGBoost, PyTorch, Optuna, PrettyMIDI, SciPy and SHAP. Random seeds were predetermined for data partitioning, model development, tuning and bootstrap analysis. Fold assignments, the locked-set fingerprint, extracted features, padded sequence tensors, out-of-fold predictions, blind predictions, training histories, hyperparameter trials, model

checkpoints, statistical comparisons and figure-generation code were retained. There were 8 final RCG-DMAF-Net checkpoints, half of them were fold-specific development models and half of them were full-development ensemble members. The entire package is also reproducible and includes both SHA-256 checksums and a modelling-status manifest and the notebook itself also contains the executable notebook. The materials allow for the checking of reported data partitions, selected parameters, numerical results.

4. Results and Discussion

4.1 Preliminary Group-Disjoint Model Benchmark

The classical, sequential and multimodal model families were first benchmarked with four-fold, composition-segment-disjoint cross-validation (CV) of the 1202 observations. The results of this analysis were compared with representation strategies and provided motivation for the further development of RCG-DMAF-Net. All predictions were made outside the fold and performances corresponding to the same segment of the composition were restricted to one-fold. The result for both primary targets is best achieved with the tree ensembles, as presented in Table 6. Random forest had the best R^2 score of 0.4873 and the lowest RMSE of 0.0870 at Interpretation Doubtful Convincing. ExtraTrees and XGBoost were next with R^2 scores of 0.4752 and 0.4632. The best results were obtained by XGBoost with an R^2 of 0.4678, followed by Random Forest with an R^2 of 0.4615 and ExtraTrees with an R^2 of 0.4555 for Music Making Pure Expressive[15].

Table 6. Preliminary group-disjoint out-of-fold performance

Target	Model	R^2	95% bootstrap interval for R^2	MAE	RMSE	Pearson r	Spearman ρ
Interpretation Doubtful Convincing	Random Forest	0.4873	0.4347–0.5338	0.0676	0.0870	0.7006	0.6526
Interpretation Doubtful Convincing	ExtraTrees	0.4752	0.4246–0.5187	0.0688	0.0880	0.6941	0.6457

Target	Model	R ²	95% bootstrap interval for R ²	MAE	RMSE	Pearson r	Spearman ρ
Interpretation Doubtful Convincing	XGBoost	0.4632	0.3970–0.5173	0.0694	0.0890	0.6829	0.6268
Interpretation Doubtful Convincing	DMAF-Net	0.4245	0.3672–0.4726	0.0722	0.0921	0.6520	0.5963
Interpretation Doubtful Convincing	MT-DMAF- Net	0.3990	0.2869–0.4804	0.0730	0.0941	0.6358	0.6010
Interpretation Doubtful Convincing	Ridge	0.3581	0.2936–0.4139	0.0762	0.0973	0.6131	0.5684
Interpretation Doubtful Convincing	Transformer	0.3093	0.2165–0.3807	0.0786	0.1009	0.5576	0.5194
Interpretation Doubtful Convincing	BiLSTM	0.2777	0.1898–0.3441	0.0804	0.1032	0.5453	0.5263
Music Making Pure Expressive	XGBoost	0.4678	0.3878–0.5312	0.0676	0.0862	0.6853	0.6606
Music Making Pure Expressive	Random Forest	0.4615	0.3917–0.5170	0.0686	0.0867	0.6848	0.6623
Music Making Pure Expressive	ExtraTrees	0.4555	0.3880–0.5099	0.0690	0.0872	0.6836	0.6647
Music Making Pure Expressive	DMAF-Net	0.4138	0.3208–0.4860	0.0715	0.0905	0.6439	0.6187
Music Making Pure Expressive	MT-DMAF- Net	0.3974	0.2858–0.4834	0.0721	0.0918	0.6359	0.6230

Target	Model	R ²	95% bootstrap interval for R ²	MAE	RMSE	Pearson r	Spearman ρ
Music Making Pure Expressive	Ridge	0.2791	0.1238–0.4001	0.0780	0.1004	0.5640	0.5580
Music Making Pure Expressive	Transformer	0.2603	0.1710–0.3327	0.0808	0.1017	0.5171	0.5025
Music Making Pure Expressive	BiLSTM	0.1815	0.0974–0.2471	0.0847	0.1069	0.4507	0.4365

The sequence-only models (BiLSTM and Transformer) achieved less accuracy than the structured baselines. The Transformer was found to be better than BiLSTM, but didn't outperform Ridge regression for either of the targets. This finding suggests that at the current dataset scale temporal ordering was not enough to utilise the information available efficiently. The difference between sequence models and tree ensembles was significantly reduced by DMAF-Net. The RMSE was reduced from 0.1032 to 0.0921 for convincingness and 0.1069 to 0.0905 for pure expressiveness compared to BiLSTM. MT-DMAF-Net was slightly inferior than its single task counterpart on the two major targets.

Multi-task learning achieved good performances for some of the lower-level PercePiano dimensions, but did not benefit the two broad high-level outcomes. This pattern indicates low-level negative transfer, due to dissociations between perceptual dimensions. These observations were validated with paired group-bootstrap comparisons. For both targets, the leading tree model outperforms BiLSTM and Transformer in terms of RMSE. The difference between the leading tree model and DMAF-Net was significantly decreased, but not statistically separable in the preliminary tree benchmark. The results provided the motivation to stick to the residual formulation adopted in RCG-DMAF-Net, where a robust tabular prediction was kept and sequential data was limited to a gated correction[19].

4.2 RCG-DMAF-Net Hyperparameter Selection and Development Performance

All 12 configurations were tested with just the 958 development observations in the Optuna search. The best trial yielded a validation RMSE of 0.08928, with a mean of four-fold for the two major

targets. The model selected had 48 BiLSTM hidden units, 96-dimensional fusion space, 64 tabular hidden units, dropout rate of 0.3472, learning rate of 1.107×10^{-3} , weight decay of 6.136×10^{-6} , and residual scale of 0.4257, in addition to Smooth L1 loss. This configuration was then re-trained in the four development folds. The optimal validation epochs were in the range from 8 to 38 epochs (as seen in Table 7), which shows that it is possible to use the early stopping technique depending on the fold. There was little variation between folds with mean validation RMSE of 0.09023.

Table 7. Development-fold performance of the selected RCG-DMAF-Net configuration

Development fold	Validation observations	Best epoch	Best validation RMSE
0	249	38	0.08994
1	224	37	0.08890
2	253	8	0.09149
3	232	20	0.09058
Mean	239.5	25.75	0.09023

Four final ensembles trained on full development partition with number of epochs for each fold. By taking the average of their predictions they still retained the frozen configuration but at the same time reduced the dependency on a specific neural initialization[20]. The Coefficient of determination of all the models presented on the locked evaluation set as presented in **Figure 3**.

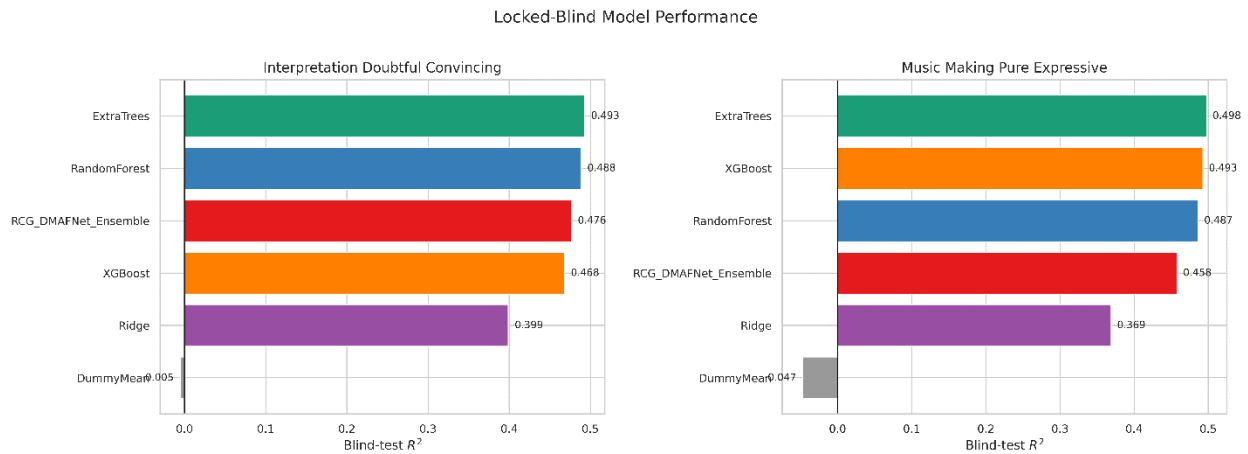


Figure 3. Coefficient of determination for all models on the locked evaluation set

4.3 Locked-Set Model Comparison

The final comparison for the 244 locked observations is given in Table 8. ExtraTrees had the best R^2 value (0.4930) and the lowest RMSE (0.0894) for Interpretation Doubtful Convincing. Random Forest followed closely with $R^2 = 0.4882$ and $RMSE = 0.0898$. RCG-DMAF-Net ranked third by R^2 and RMSE, achieving $R^2 = 0.4765$ and $RMSE = 0.0908$. It still achieved the best MAE of 0.0690, just a little worse than ExtraTrees and Random Forest. RCG-DMAF-Net also outperforms XGBoost, Ridge and DummyMean on this target. Again, ExtraTrees had the highest R^2 of 0.4979 and the lowest RMSE of 0.0837, in Music Making Pure Expressive. The R^2 values for XGBoost and Random Forest were 0.4928 and 0.4866, respectively. RCG-DMAF-Net obtained $R^2 = 0.4581$, $MAE = 0.0698$ and $RMSE = 0.0870$. It did not outperform the top tree ensembles, but significantly outpaced the Ridge regression and the mean-prediction baseline.

Table 8. Final performance on the locked composition-segment-disjoint set

Target	Model	R^2	MAE	RMSE	Pearson r	Spearman ρ
Interpretation Doubtful Convincing	ExtraTrees	0.4930	0.0691	0.0894	0.7125	0.6802
Interpretation Doubtful Convincing	Random Forest	0.4882	0.0691	0.0898	0.7074	0.6707
Interpretation Doubtful Convincing	RCG-DMAF-Net Ensemble	0.4765	0.0690	0.0908	0.6957	0.6594
Interpretation Doubtful Convincing	XGBoost	0.4681	0.0702	0.0915	0.6864	0.6479
Interpretation Doubtful Convincing	Ridge	0.3989	0.0742	0.0973	0.6512	0.5930
Interpretation Doubtful Convincing	DummyMean	-0.0049	0.1005	0.1258	Not defined	Not defined
Music Making Pure Expressive	ExtraTrees	0.4979	0.0666	0.0837	0.7427	0.7188
Music Making Pure Expressive	XGBoost	0.4928	0.0662	0.0841	0.7122	0.6761

Target	Model	R ²	MAE	RMSE	Pearson r	Spearman ρ
Music Making Pure Expressive	Random Forest	0.4866	0.0669	0.0847	0.7263	0.7023
Music Making Pure Expressive	RCG-DMAF-Net Ensemble	0.4581	0.0698	0.0870	0.6807	0.6357
Music Making Pure Expressive	Ridge	0.3692	0.0751	0.0938	0.6143	0.5636
Music Making Pure Expressive	DummyMean	-0.0474	0.0996	0.1209	Not defined	Not defined

The R² values of the DummyMean models were negative, indicating that all substantive models learned out-of-group information. The model ranking supports the preliminary finding that engineered global MIDI descriptors are highly informative at the available dataset size. Despite this, RCG-DMAF-Net was still competitive with the absolute RMSE difference from ExtraTrees of 0.0020 for convincingness and 0.0033 for pure expressiveness. Complementary overview of the correlations. ExtraTrees achieved Pearson correlations of 0.7125 and 0.7427 for convincingness and pure expressiveness, respectively, while RCG-DMAF-Net achieved 0.6957 and 0.6807. All substantive models showed positive Spearman correlations, suggesting meaningful rank ordering of predictions also[21].

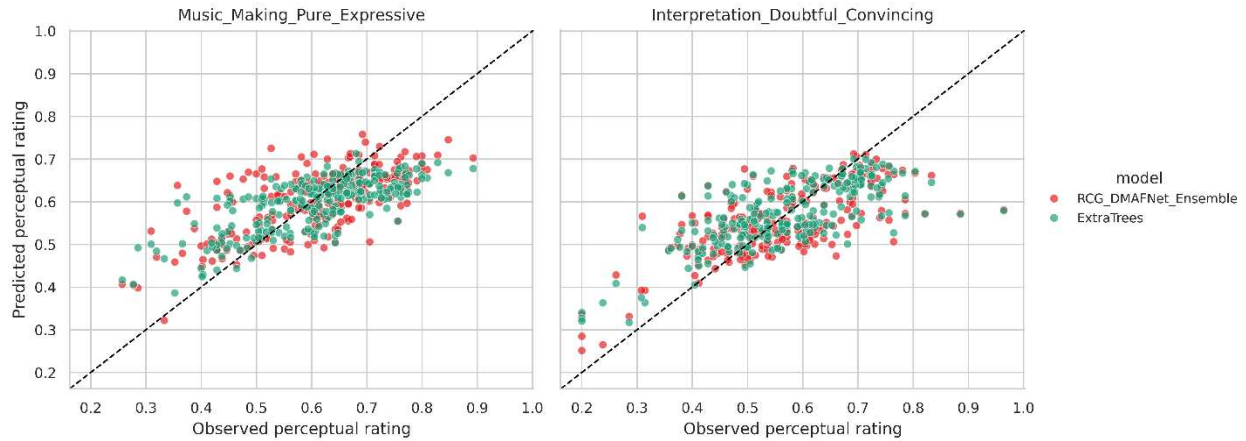


Figure 4. Observed and predicted perceptual ratings for RCG-DMAF-Net and the leading tree ensemble on the locked set. The dashed diagonal indicates perfect agreement.

It is clear that the prediction trend was similar to the observed rating trend, although the predicted ratings were contracted towards the centre of the target range as can be seen in **Figure 4**. This regression to the mean effect was most clearly observed at the extremes of perceptual ratings, and is similar to the intermediate R^2 values. This is probably due to the relatively small sample of independent score-segment groups, and to the subjectivity of the high-level perceptual annotations. The Residual distributions of the evaluated models on the locked set is shown in **Figure 5**.

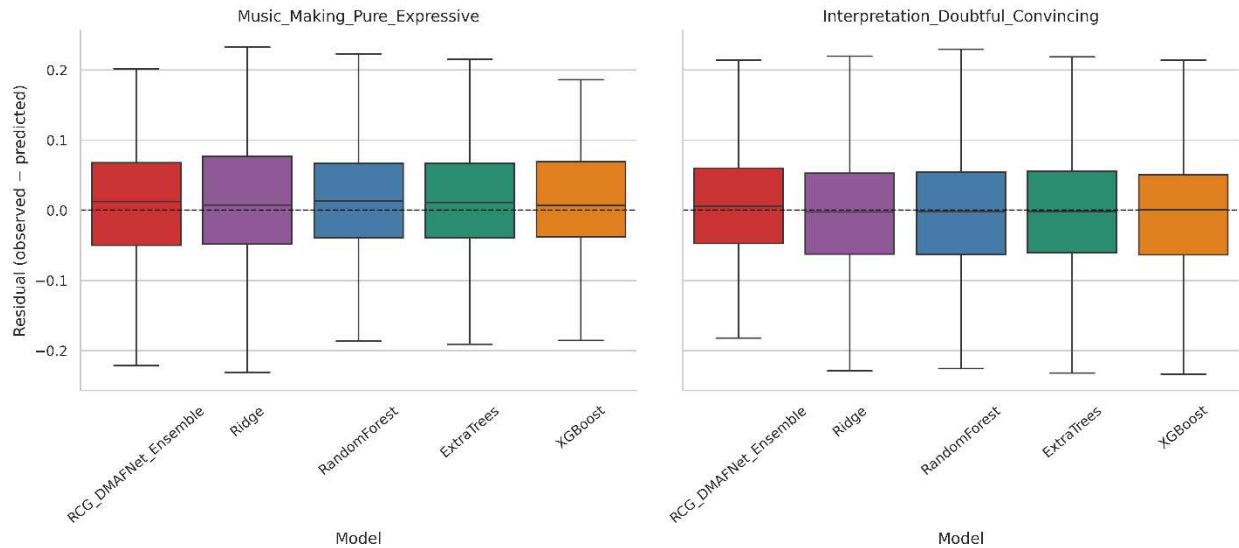


Figure 5. Residual distributions of the evaluated models on the locked set

There was no obvious gross systematic bias with residual distributions generally centred around zero. The errors were fairly small for the top models, but the Ridge and DummyMean models had clearly larger errors. Pearson and Spearman correlations were undefined for DummyMean because it produced a constant prediction and therefore had zero predicted variance. The RCG-DMAF-Net ensemble's development and locked-set RMSE can be seen in **Figure 6**.

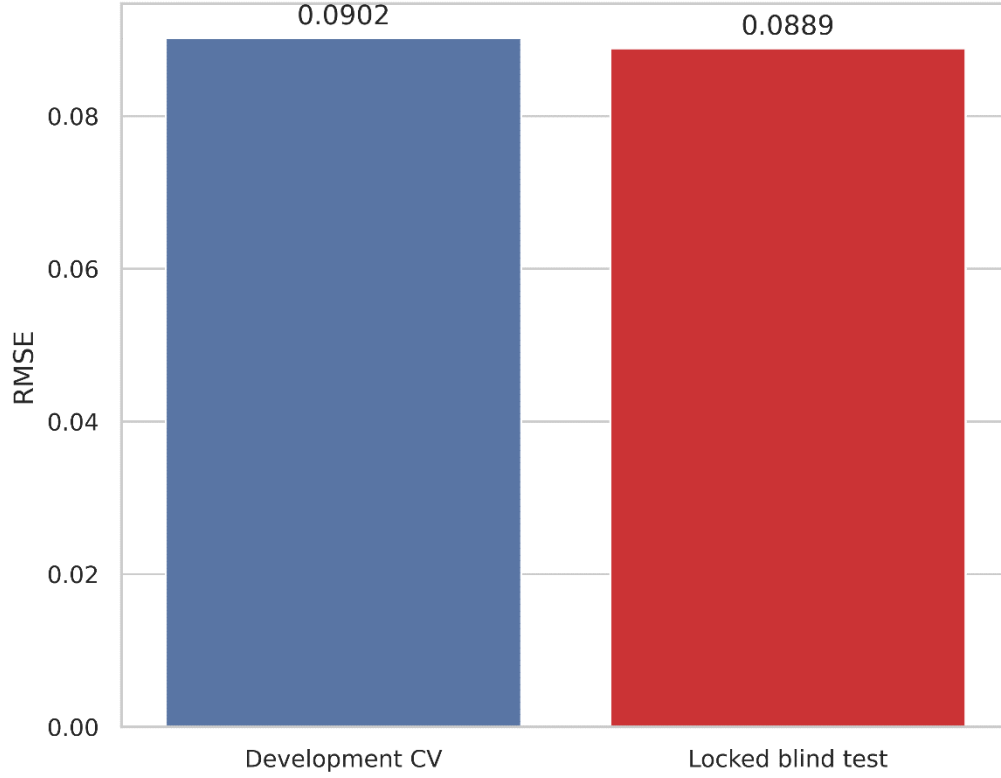


Figure 6. Development and locked-set RMSE of the RCG-DMAF-Net ensemble

The similarity of development and locked-set errors suggests that there was not much degradation of the model after the model was frozen. The mean development RMSE was 0.0902 and the mean locked-set RMSE for the two targets was around 0.0889. This does not mean that all the data in the partitions was easier, as each partition contained different amounts of target variance and composition. It obviously does not, however, reflect a significant degradation on the locked composition-segment groups[22].

4.4 Paired Group-Bootstrap Comparison

A set of 5,000 paired bootstrap samples for each of the 20 composition-segment groups (C-20) were extracted from a total of 20 groups of locked compositions to test for the numerical advantage of ExtraTrees over RCG-DMAF-Net. Positive differences in table 9 are associated with an increase in RMSE for RCG-DMAF-Net.

Table 9. Paired group-bootstrap comparison between ExtraTrees and RCG-DMAF-Net

Target	RCG-DMAF-Net minus ExtraTrees RMSE	95% confidence interval	Probability RCG- DMAF-Net was better
Interpretation Doubtful Convincing	0.00196	−0.00165 to 0.00534	0.1416
Music Making Pure Expressive	0.00346	−0.00524 to 0.01317	0.2594

Both 95% confidence intervals contained the value of zero. The locked groups available were therefore not statistically sufficient to differentiate between the statistical RMSE of RCG-DMAF-Net and ExtraTrees at the 95% level. This result should not be taken as a sign of equivalence as ExtraTrees was numerically stronger, and was preferred in most bootstrap samples. Instead, there was not a clear difference between the grouped sample at the chosen confidence level[23].

4.5 Controlled Ablation of RCG-DMAF-Net

Contributions by each of the representation and fusion mechanisms were isolated in the development-only ablation analysis. The mean performance of the sequence-only modelling is the lowest among the other model types, as seen in Table 10, with $R^2 = 0.3140$ and $RMSE = 0.0975$. The RMSE and mean R^2 values for tabular-only neural pathway were 0.0922 and 0.3868, respectively. All three multimodal configurations showed further improvement, with a mean R^2 of between 0.4155 and 0.4210. The Controlled development-only ablation of RCG-DMAF-Net components is depicted in **Figure 7**.

Table 10. Development-only RCG-DMAF-Net ablation results averaged across both primary targets

Architecture variant	Active parameters	Mean R^2	Mean MAE	Mean RMSE	Mean residual gate
Simple fusion	74,851	0.4210	0.0706	0.0895	Not applicable
Full RCG-DMAF-Net	96,839	0.4169	0.0707	0.0899	0.3937

Architecture variant	Active parameters	Mean R^2	Mean MAE	Mean RMSE	Mean residual gate
Ungated residual fusion	96,453	0.4155	0.0705	0.0900	Fixed at 1.000
Tabular only	13,058	0.3868	0.0731	0.0922	Fixed at 0.000
Sequence only	49,219	0.3140	0.0772	0.0975	Fixed at 1.000

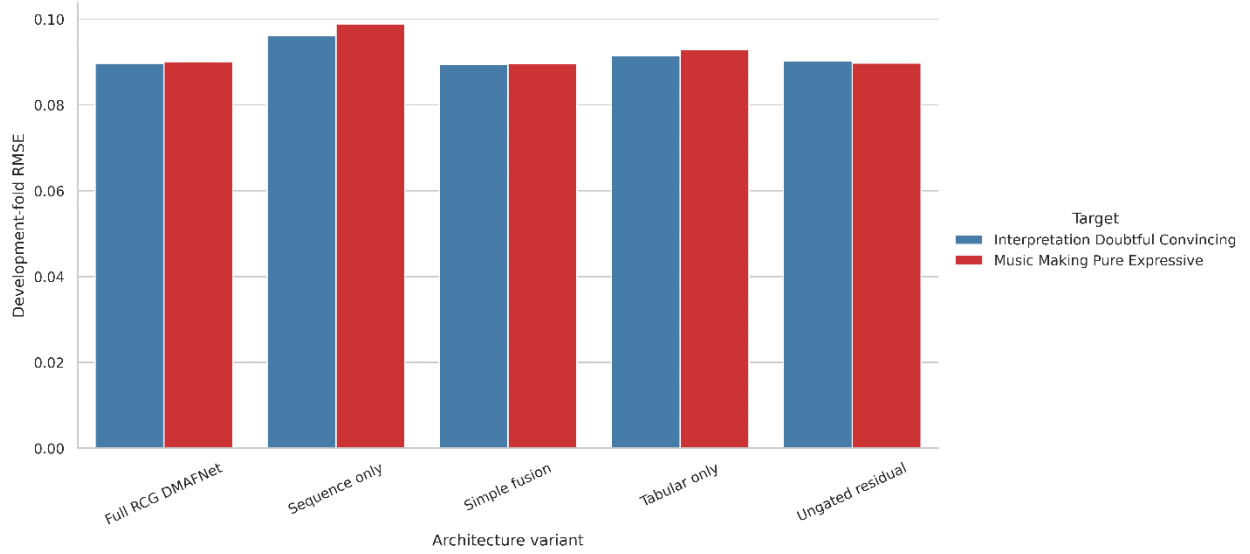


Figure 7. Controlled development-only ablation of RCG-DMAF-Net components

The lowest mean RMSE value was achieved by simple fusion with value 0.0895, closely followed by full RCG-DMAF-Net with value 0.0899 and ungated residual fusion with value 0.0900. These differences were not statistically distinguishable using paired group-bootstrap analysis for either primary target. The overall model was not found to be significantly better than simpler multimodal fusion. Conversely, the full RCG-DMAF-Net was much better than the sequence-only modelling. The RMSE difference between the full and the sequence only models was -0.00876 with a 95% interval from -0.01302 to -0.00467 for Music Making Pure Expressive. The difference for

Interpretation Doubtful Convincing was -0.00656 , with a 95% interval of $(-0.00992, -0.00303)$. The full model numerically had better result for both targets than the tabular-only pathway, but did not show a positive result in the confidence intervals[24].

The results show that the main architectural advantage is that of multimodal representation. The mean value of the target-specific residual gates was 0.3937 , suggesting that the target-specific correction sequence was not applied in its entirety. But the existing collection does not prove that cross-gating alone is the easiest method of increasing the accuracy of prediction over all the other simpler fusion methods. The contribution of RCG-DMAF-Net is thus a structured and interpretable residual-fusion formulation and not a general accuracy gain.

4.6 SHAP-Based Interpretation

SHAP analysis was performed on models that were built for the development set only, in this case the strongest baseline model ExtraTrees. **Figure 8 (a-b)** shows the feature level effects for the two main targets. Positive values of SHAP increased the predicted perceptual score, and the negative values of SHAP reduced the predicted perceptual score. Colour is the intrinsic value of the feature.

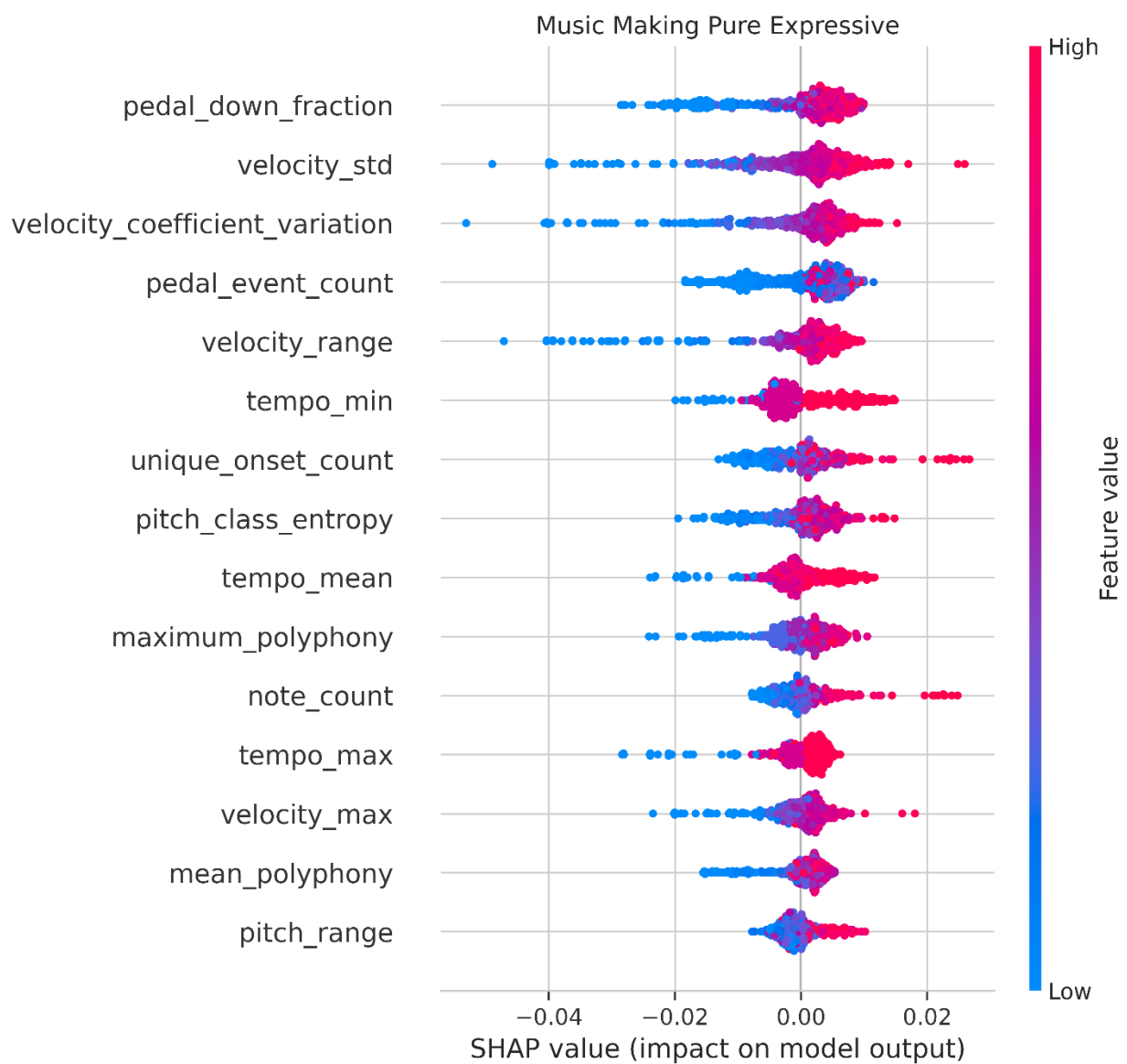


Figure 8a. SHAP summary for Music Making Pure Expressive using the development-only ExtraTrees model

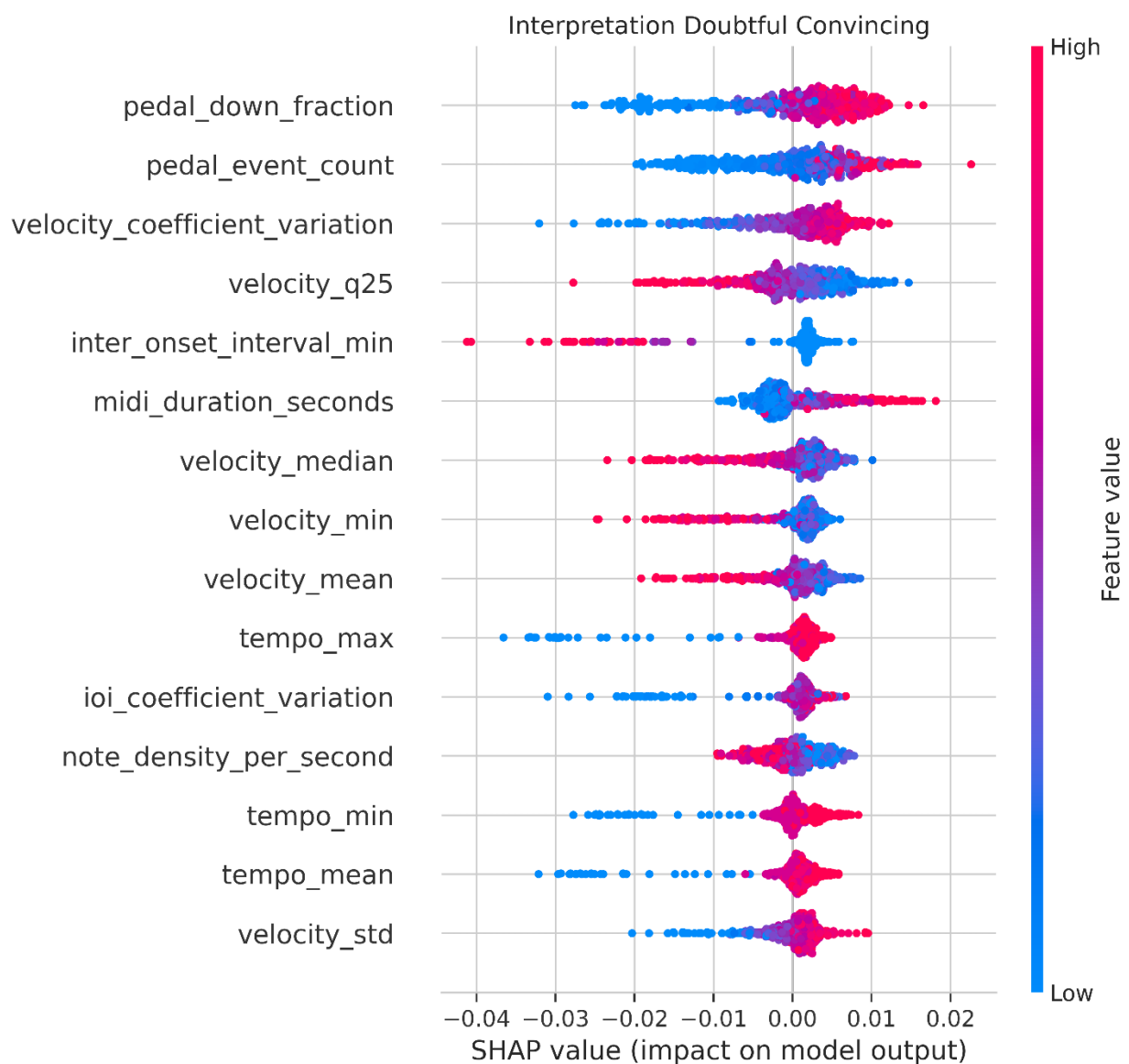


Figure 8b. SHAP summary for Interpretation Doubtful Convincing using the development-only ExtraTrees model

Pedal-down fraction and velocity standard deviation and velocity coefficient of variation were the top three most influential descriptors by mean absolute SHAP value. Other factors included: Pedal-event count, velocity range, minimum tempo, unique-onset count, pitch-class entropy, mean tempo and maximum polyphony. Increasing the pedal-down fraction and the velocity variation tended to move the predictions toward the expressive end of the continuum, but the distributions showed

that not all of the pedal-down fraction and velocity variation effects were monotonic for all observations[25].

Table 11 shows Ten leading SHAP features for each primary target. Pedal-down fraction and pedal-event count were the top two descriptors for Interpretation Doubtful Convincing. Lower-quartile velocity, minimum inter-onset interval, performance duration, median velocity, minimum velocity, mean velocity and maximum tempo followed. The high frequency of both pedal and velocity descriptors on both targets indicates that high level perceptual judgement carries information in sustained resonance and dynamic shaping. The significance of onset and tempo variables for the persuasiveness of a message also suggests that perceived interpretative coherence is created by temporal organization[26].

Table 11. Ten leading SHAP features for each primary target

Target	Rank	Feature	Mean absolute SHAP
Music Making Pure Expressive	1	Pedal-down fraction	0.006367
Music Making Pure Expressive	2	Velocity standard deviation	0.006356
Music Making Pure Expressive	3	Velocity coefficient of variation	0.005894
Music Making Pure Expressive	4	Pedal-event count	0.005637
Music Making Pure Expressive	5	Velocity range	0.004950
Music Making Pure Expressive	6	Minimum tempo	0.004736
Music Making Pure Expressive	7	Unique-onset count	0.004553
Music Making Pure Expressive	8	Pitch-class entropy	0.004081
Music Making Pure Expressive	9	Mean tempo	0.003705
Music Making Pure Expressive	10	Maximum polyphony	0.003504
Interpretation Doubtful Convincing	1	Pedal-down fraction	0.006727

Target	Rank	Feature	Mean absolute SHAP
Interpretation Doubtful Convincing	2	Pedal-event count	0.006194
Interpretation Doubtful Convincing	3	Velocity coefficient of variation	0.005121
Interpretation Doubtful Convincing	4	Lower-quartile velocity	0.004491
Interpretation Doubtful Convincing	5	Minimum inter-onset interval	0.004001
Interpretation Doubtful Convincing	6	Performance duration	0.003693
Interpretation Doubtful Convincing	7	Median velocity	0.003639
Interpretation Doubtful Convincing	8	Minimum velocity	0.003297
Interpretation Doubtful Convincing	9	Mean velocity	0.002990
Interpretation Doubtful Convincing	10	Maximum tempo	0.002849

The global feature importance scores for the models trained on the development set are given in Appendix A for the target Interpretation Doubtful–Convincing (**Figure A1**) and Music Making Pure Expressive (**Figure A2**). These complementary plots agree that the following predictor groups were among the most influential: pedalling behaviour, velocity variation, onset structure, tempo and polyphony. The explanations were only produced from observations from the development set, leaving the evaluation set locked out of model interpretation, model development and selection.

The fitted model (and not causal musical mechanisms) is the reason for the SHAP values. Predictive information can be shared among timing, velocity and pedal descriptors, and a high SHAP value does not necessarily indicate that the manipulation of one of these features would have a direct effect on the perceived expressivity of the other[27].

4.7 Practical Implication

Based on the results, three main conclusions can be made. Indeed, engineered MIDI descriptors proved to be very effective representations for the two high-level perceptual goals. ExtraTrees had the best locked-set performance, and shown that carefully designed global statistics are competitive with more complicated neural networks when there are moderate numbers of independent musical groups. Second, sequence information proved to be helpful in conjunction with global descriptors, but not as helpful when used alone. DMAF-Net significantly outperformed the sequence-only pathway and the ablation results demonstrated that all the multimodal configurations were better than the baselines of BiLSTM and Transformer. In this regard, this is in line with the central representation hypothesis that local note-event evolution and global performance statistics offer complementary information[28].

Third, complexity did not predict to the best of advantage. While RCG-DMAF-Net approached the leading tree ensembles and outdid XGBoost on the locked set in convincingsness, it was not able to outperform all tree baselines. There was also no statistically significant improvement of the ablation over simple or ungated fusion. The results do not support the view that neural fusion is always a good thing. Instead, RCG-DMAF-Net offers a principled approach to adapt a sound tabular estimate based on temporal evidence but subject to target-specific control over the correction. The study also highlights the need for the leakage-aware evaluation. Random observation-level splitting may lead to over-optimistic estimates since the performances cover a significant part of the same musical content in the score segments. The locked grouped comparison and composition-segment-disjoint development folds were more challenging to assess for generalization to unseen excerpts. The moderate R^2 values from this protocol are thus more believable than the high classification scores from random sampling or discretized continuous labels. Pedalling, onset structure, tempo and polyphony were always informative from a musical point of view. These results are consistent with the idea that the capacity to express and interpret conviction takes the form of an integrated control of resonance, dynamics and shaping in time.

Concurrently, perceptual ratings are subjective, and can be influenced by musical context, which is not completely captured by the MIDI descriptors. Consequently, the results are to be viewed as predictive evidence rather than general rules of expressive performance in PercePiano.

5 Limitations and Future Research

The results must be considered in the light of certain limitations. The first is that the dataset originally consisted of 1,202 performance segments, but these observations were for 99 composition-segment groups. The group-disjoint design allowed for a sound evaluation of unseen excerpts, but further validation across other composers, performers and performance traditions would enhance generalizability.

Secondly, the models were based on symbolic MIDI data. It was a representation that allowed serious analysis of timing, pitch, velocity, duration, polyphony and pedalling, but not the acoustic properties of recording and room response, or the timbre of instruments. Future studies might integrate sequences of MIDI and score data with pre-trained audio representations in a leakage controlled multimodal approach.

Third, RCG-DMAF-Net was optimized for two level-2 perceptual targets. Numerically it was the strongest result of the multimodal formulation with ExtraTrees, but the difference among the multimodal fusion variants was small and the result of the multimodal formulation was superior in terms of competitive locked-set performance and improvement over sequence-only learning. To find out when residual cross-gating is most beneficial, more testing needs to be done on more datasets and other perceptual dimensions.

Lastly, the SHAP values do not represent musical mechanisms, but rather, they represent predictive associations. Future research could be based on the integration of model explanations and expert musicological analysis, controlled listening experiments and uncertainty-aware prediction. These extensions would aid in the creation of reliable systems for providing feedback on education and real-time assessment of piano performances.

6 Conclusion

In this study, the expressivity of piano performance was reformulated as continuous perceptual regression based on the official PercePiano dataset. Variable-length sequences of eight attributes, representing 1,202 performance segments, and 51 MIDI descriptors, were engineered. All observations of a common excerpt from the same composition were stored in the same partition to minimize leakage due to the repeated performance of the same excerpt. Four group-disjoint folds comprising 958 observations were used for model development and hyperparameter selection, while 244 observations from 20 previously unseen groups were reserved as a locked evaluation set. The locked observations were not used for hyperparameter tuning or model selection. The study introduces RCG-DMAF-Net, Residual Cross-Gated Dual-Branch MIDI Attention Fusion Network. The main prediction was obtained from its tabular pathway, and an attention-weighted bidirectional LSTM encoded note-event sequences and provided a target-specific gated residual correction. The architecture was tested with Ridge regression, Random Forest, ExtraTrees, XGBoost, BiLSTM, Transformer, DMAF-Net and MT-DMAF-Net models. RCG-DMAF-Net had $R^2 = 0.4765$ and $RMSE = 0.0908$ on the locked set for Interpretation Doubtful Convincing and $R^2 = 0.4581$ and $RMSE = 0.0870$ for Music Making Pure Expressive. The highest numerical R^2 values were obtained for the ExtraTrees with values 0.4930 and 0.4979 respectively. Pairwise group-bootstrap intervals for the RMSE differences between ExtraTrees and RCG-DMAF-Net, however, contained zero, suggesting that differences in the available grouped observations were statistically non-significant at 95% confidence level. The results show that the proposed architecture performed competitively with the strong structured-data ensembles but this does not justify a claim to outperform all the baselines.

Multimodal representation was the most important advantage identified with controlled ablation. The complete model greatly outperformed the sequence-only learning approach for both targets, while its performance was only slightly worse than the simple fusion approach and ungated residual fusion approach, and significantly less so than the tabular-only pathway. The SHAP analysis also revealed that the pedalling behaviour, velocity variability, onset structure, tempo and polyphony were consistently linked to model predictions. These results relate predictive performance to musically interpretable features and are non-causal. As a whole, the study shows the necessity of designing expressive piano representation carefully and validating it through

leakage aware methods to achieve reliable computational evaluation of performance. Sequential MIDI information can be used together with global engineered descriptors, but a complex architecture does not necessarily provide better generalization on moderately sized data sets. The combination of a proposed residual fusion model, strong classical baselines, composition-segment-disjoint evaluation, controlled ablation, group-bootstrap inference and reproducible evidence provides a transparent foundation for future multimodal piano-performance research.

Declarations

Funding: The authors received no funding for this research.

Competing interests: The authors declare that they have no competing financial or non-financial interests.

Author contributions Statement: G.J. and D.J. contributed to the study conception and design, methodology, data preparation, formal analysis, validation, interpretation of the results, and manuscript preparation. Both authors reviewed and approved the final manuscript.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Data and code availability

The original data analysed in this study are publicly available in the PercePiano repository at <https://github.com/JonghoKimSNU/PercePiano> and are described at <https://doi.org/10.1038/s41598-024-73810-0>. Repository commit e672299b8661f63802cd5cd73bb3c76218a13029 was used for the present analysis.

References

1. Shen, L. *et al.* Visuotactile multimodal fusion with flexible piezo-triboelectric sensors for high-fidelity piano keystroke dynamics decoding. *Chemical Engineering Journal* 547, 180292 (2026).
2. Wu, D. Utilizing AI and machine learning in vocal training programs: A contemporary approach. *Learn. Instr.* 102, 102278 (2026).
3. Zhang, D. A hybrid architecture combining physical modeling and neural networks for piano sound synthesis. *Systems and Soft Computing* 8, 200435 (2026).
4. Widmer, G. & Goebel, W. Computational Models of Expressive Music Performance: The State of the Art. *J. New Music Res.* 33, 203–216 (2004).
5. Seyedjavad, M., Naeeni, S. O. & Saneie, M. Flow velocity pattern around trapezoidal piano key side weirs. *Flow Measurement and Instrumentation* 76, 101847 (2020).
6. Prasetyo, Y. T. *et al.* Preference analysis on piano music school attributes among young pianists: A conjoint analysis approach. *Acta Psychol. (Amst)*. 262, 106070 (2026).
7. Murtaza, N., Akbar, Z., Ahmed Pasha, G. & Alquraish, M. Piano Key Weirs in Modern Hydraulics: A state-of-the-art review. *Flow Measurement and Instrumentation* 107, 103113 (2026).
8. Hao, J. Online piano learning game design method: Piano music style recognition based on CRNNH. *Entertain. Comput.* 50, 100645 (2024).
9. Shi, J., Wan, N. & Ibrahim, R. Analysis of Piano Online Teaching System Based on Maximum Logarithm MPA Algorithm Technology. *International Journal of Web-Based Learning and Teaching Technologies* 19, (2024).
10. Ghatas, Y., Fayek, M. & Hadhoud, M. A hybrid deep learning approach for musical difficulty estimation of piano symbolic music. *Alexandria Engineering Journal* 61, 10183–10196 (2022).
11. Minato, A. *et al.* Improving Data-Driven Coffee Machine Learning Through IoT Integration. *IFAC-PapersOnLine* 59, 1–6 (2025).

12. Ramoneda, P., Suzuki, M., Maezawa, A. & Serra, X. Difficulty-aware score generation for piano sight-reading. *Expert Syst. Appl.* 321, 132088 (2026).
13. Roushangar, K., Panahi, A. & Farajpour, I. Simultaneous prediction of discharge coefficients for sharp and broad crested trapezoidal side weirs using machine learning models. *Flow Measurement and Instrumentation* 111, 103473 (2026).
14. Wei, X. & Sun, S. Optimization of Piano Performance Teaching Mode Using Network Big Data Analysis Technology. *International Journal of Information and Communication Technology Education* 20, (2024).
15. Zhang, Q. Research on the Application of Network Sensing Technology and Virtual Reality Technology in Piano Teaching. *International Journal of Cognitive Informatics and Natural Intelligence* 19, (2025).
16. Ramoneda, P. *et al.* Combining piano performance dimensions for score difficulty classification. *Expert Syst. Appl.* 238, 121776 (2024).
17. Lian, Y. An AI-Enhanced Piano Improvisation Teaching System. *Journal of Cases on Information Technology* 28, (2026).
18. Satkunarajah, P., Power, S. D. & Zendel, B. R. Single-trial EEG-based classification reveals instrument-specific timbre perception via traditional machine learning classifiers. *Neuroimage* 336, 122010 (2026).
19. Lv, C. Development and application of piano accompanying system based on New Fingerprint algorithm. *Array* 26, 100387 (2025).
20. Cancino-Chacón, C. E., Grachten, M., Goebel, W. & Widmer, G. Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities* Volume 5-2018, (2018).
21. Mehri, Y., Mehri, M. & Nasrabadi, M. Application of machine learning models in predicting discharge coefficient of side B-type piano key weir. *Flow Measurement and Instrumentation* 100, 102687 (2024).

22. Zheng, Y. & Wang, L. Application of entertainment virtual technology based on network information resources in piano teaching. *Entertain. Comput.* 50, 100675 (2024).
23. Hamed, A. K. & Elshaarawy, M. K. Hybrid data-driven machine learning approach for forecasting discharge coefficient of Piano Key Weirs. *Flow Measurement and Instrumentation* 107, 103050 (2026).
24. Tian, W., Isleem, H. F., Hamed, A. K. & Elshaarawy, M. K. Enhancing discharge prediction over Type-A piano key weirs: An innovative machine learning approach. *Flow Measurement and Instrumentation* 100, 102732 (2024).
25. Zhou, D. & Nataliaia, S. Generative adversarial network for adaptive piano accompaniment. *Systems and Soft Computing* 7, 200289 (2025).
26. Xiao, L. Deep Learning-Assisted Personalized Piano Teaching in Blended Learning Environments. *International Journal of Mobile and Blended Learning* 18, (2026).
27. Guo, H. Piano harmony automatic adaptation system based on deep reinforcement learning. *Entertain. Comput.* 52, 100706 (2025).
28. Park, J. *et al.* Piano performance evaluation dataset with multilevel perceptual features. *Sci. Rep.* 14, 23002 (2024).

Appendix A

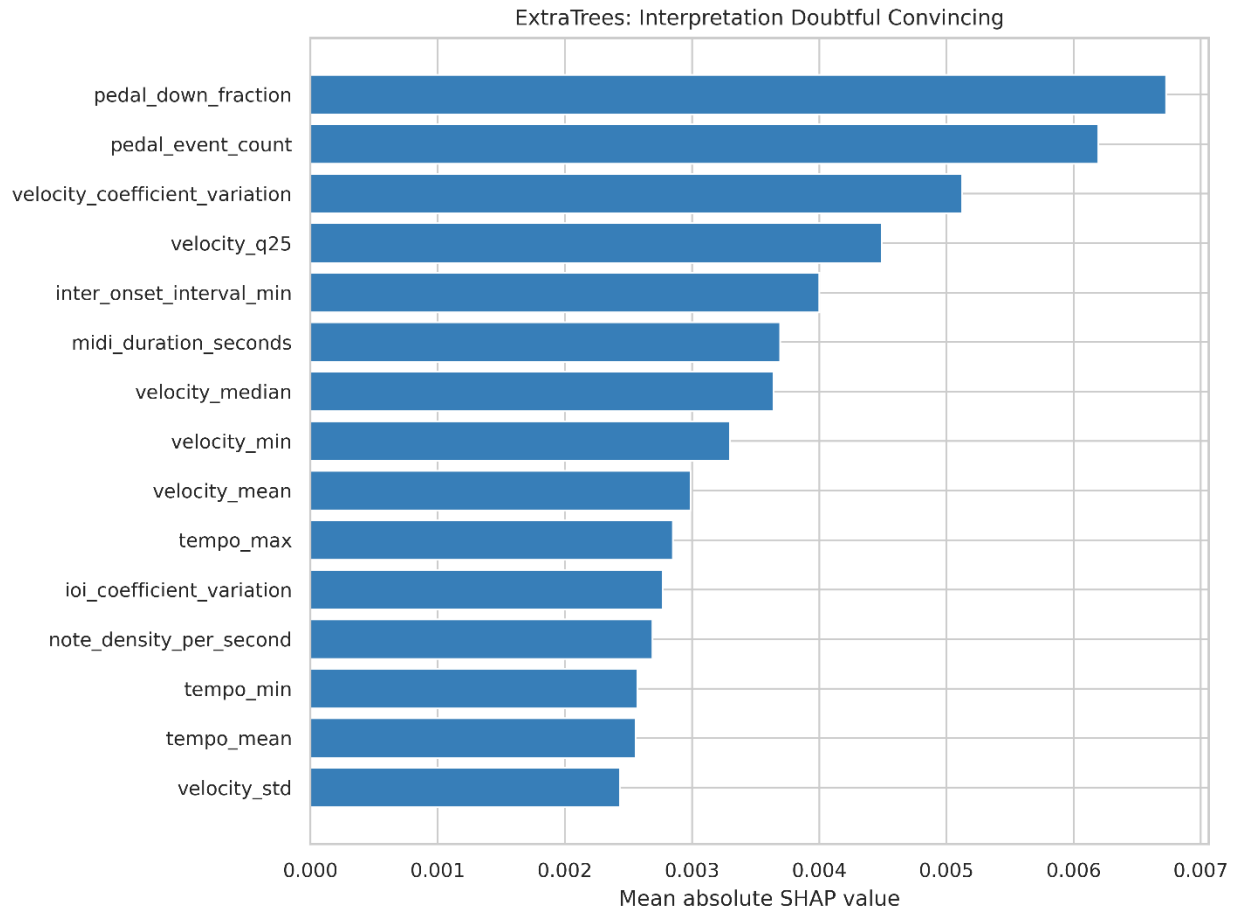


Figure A1. Development-only ExtraTrees feature-importance ranking for the *Interpretation Doubtful–Convincing* target, based on mean absolute SHAP values.

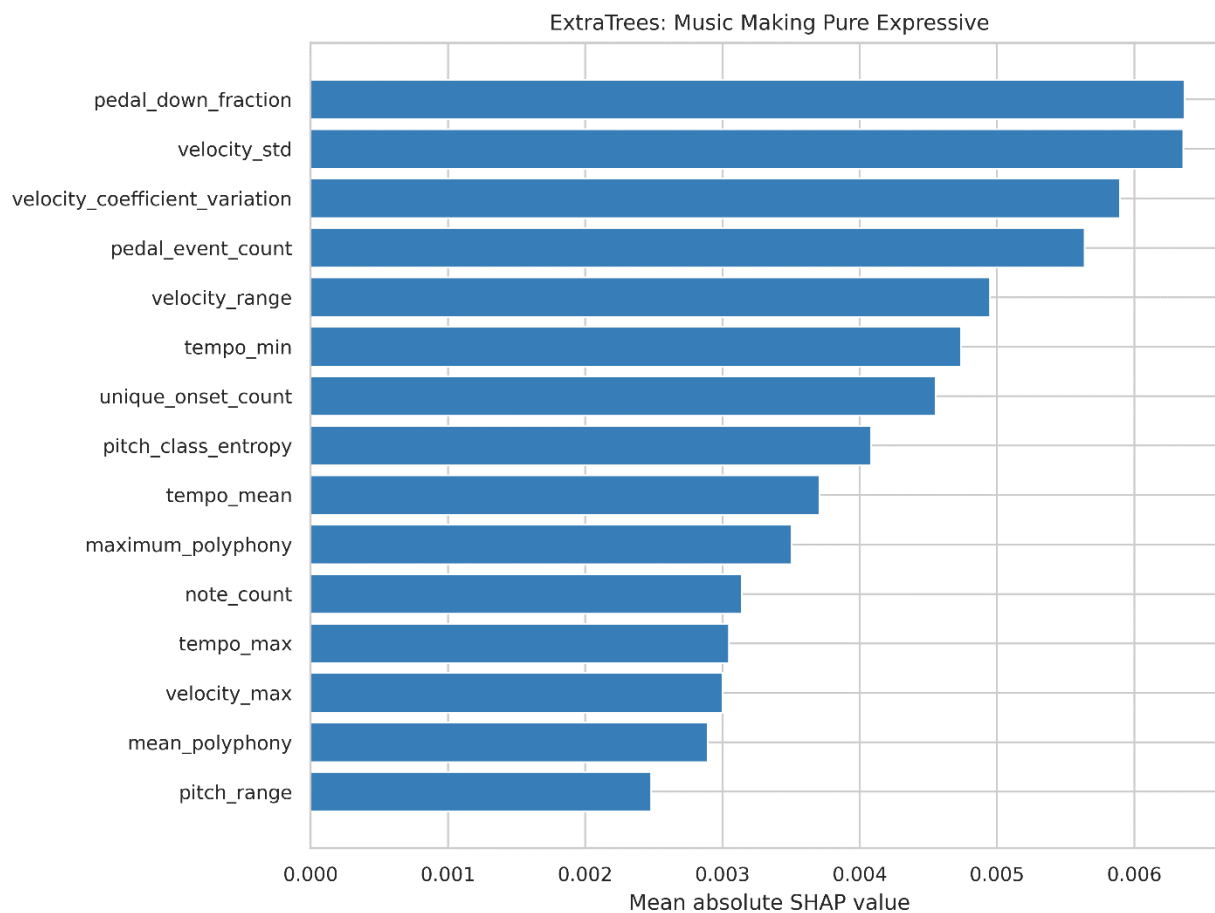


Figure A2. Development-only ExtraTrees feature-importance ranking for the *Music Making Pure Expressive* target, based on mean absolute SHAP values.