Introduction

Music Information Retreival (MIR) is a rapidly growing interdisciplinary field focused on extracting feature or making predictions from musical representations. Music content analysis tools, such as MIRtoolbox (Lartillot et al., 2008; Lartillot & Toiviainen, 2007a), Essentia (Bogdanov et al., n.d., 2013), Librosa (McFee et al., 2015), jAudio (McEnnis et al., n.d.), Timbre Toolbox (Peeters et al., 2011), Humdrum (Huron, 2002), Music21 (Lartillot et al., 2008), and others extract a wide variety of these features (for instance, timing, timbre, pitch, and emotion features) — typically from audio files, but also from symbolic formats such as Musical Instrument Digital Interface (MIDI).

Despite their widespread use in research and industry, independent or third-party evaluations of these tools are relatively rare. The MIR research community has developed incentives like Music Information Retrieval Evaluation eXchange (MIREX) to formally assess a variety of music information retrieval tasks (Downie, 2008; Raś et al., 2010). Beyond MIREX, independent evaluations of music content analysis tools outside of the development process have answered key research questions relating to these tools and created resources for newcomers to orient themselves and select a tool suitable for their needs. Urbano et al. (2014) tested the effect of audio quality factors (sample rate, bit rate, codec, and frame size) on the robustness of chroma features and Mel-frequency cepstral coefficients (MFCCs) with 400 audio files across two notable extraction tools. Moffat et al. (2015) evaluated the diversity of features, user interface, data accessibility, and computational efficiency of ten MIR tools, but do not examine the effectiveness, accuracy, or consistency of their feature extraction. Kumar et al. (2015) tested the reliability of three pitch-based features in MIRtoolbox using a variety of instruments: however, it is unclear precisely what audio files were used and how representative they are of naturalistic

musical stimuli. Tools for modifying audio files have provided a framework for the systematic testing of music content analysis tools (in particular, Mauch and Ewert (n.d.)). In addition to these testing tools and evaluations, significant discussion has been devoted to the need for evaluation and testing of new extraction tools (Cunningham et al., 2012; Downie, 2004; Gómez, 2006; Sturm, 2016; Urbano et al., 2013). The difficulties associated with testing subjective musical features, or features without a ground truth (some outcome or true value to be modeled), are often cited as one of the main barriers to more substantial evaluations. Indeed, part of the motivation for efforts like MIREX is to overcome difficulties associated with evaluating musical features (Downie, 2008).

Similar to other computational and statistical tools, MIR algorithms are typically tested by comparing the output of an algorithm with a ground truth that reflects either the perceived reality of a feature or otherwise a physical or technical property of the acoustic sound wave. Synthesizing MIDI files can be particularly effective, as manipulating parameters such as tempo, pitch distribution, timbre, and dynamics can provide labeled data suitable for use as a ground truth in training and evaluation (Cataltepe et al., 2007; Hu & Dannenberg, n.d.; Raffel & Ellis, n.d.). In the case of symbolically notated music, such as western classical music, structural features or labels specified by the composer such as title-specified keys, key signatures, tempo markings, etc. can provide an additional source of ground truth for evaluating analogous extracted features.

However, many features of music are highly subjective: for instance, mode is a highly culturally informed phenomenon that differs even among experts (Delle Grazie et al., in review), and the perception of consonance and dissonance is largely informed by inculturation (Lahdelma et al., 2022). The absence of an objectively verifiable ground truth complicates evaluation of

algorithms, often relying on human labeled databases which can be challenging to develop or retrieve (Gómez, 2006), and are generally limited in their generalizability.

We propose a method for evaluating the effectiveness and validity of a MIR algorithm that compares versions of the same composition. Although in principle this method could be applied with many styles, instruments, and genres, here we focus on classical piano repertoire, which is particularly well-suited to this task. Western classical piano music is predominantly learned from symbolic notation, and it is generally understood that although some features might differ across performances of the same composition (i.e. tempo, timbre; here referred to as variant features), others should not (i.e., mode, pitch distribution; here referred to as invariant features). Comparing the consistency of feature extractions across multiple performances, or versions, of the same composition offers an opportunity to clarify the ability of MIR tools to focus on specific dimensions of interest, particularly in the case of invariant features. While cover song identification, the task of identifying different versions of the same composition has been the subject of significant study (see Zheng et al. (2023) for a review), we are not aware of any attempts to use different performances of the same composition as an evaluation technique.

To assess consistency in algorithmic analysis of related musical passages, here we analyze the first eight measures of 17 audio file versions of all 24 piano preludes from J.S. Bach's "The Well Tempered Clavier" (WTC) Book 1 — 408 files in total. The WTC is particularly suitable since its musical and historical significance has resulted in numerous professional recordings by renowned artists. A larger sample of versions is desirable for this method as outliers can disproportionately influence variability measures in smaller samples. Further, Bach's work is still considered relevant and influential today, and is also the subject of significant study in MIR (Chen & Su, 2021; Gotham et al., 2023), music cognition (Battcock &

Schutz, 2019, 2021, 2022), computational musicology (Lieck et al., 2020; Schmuckler & Tomovski, 2005; Temperley, 1999), and his compositions are frequently used in the context of music education.

Methods

Stimuli

We analyzed 17 unique albums, each of which contain all 24 preludes from J.S. Bach's "The Well Tempered Clavier" Book 1 (408 files). From each Compact Disc (CD), we extracted audio encoded in the Waveform Audio File format (.wav) at a sampling rate of 44100 Hz with 16 bit-depth. Each analyzed audio file includes the first eight measures of the prelude, as outlined in Battcock and Schutz (2019), however without the two second fade-out used in that study. Sixteen notable performers recorded these albums between 1934 and 2015 (see Table 1 for more details). 11 are piano performances and 6 are recorded on harpsichords. A subset of these albums are included in Palmer's analysis (Bach & Palmer, 2004) of the Well Tempered Clavier.

Table 1: Performances of the Well Tempered Clavier used, with details.

Performer	Year Recorded	Instrument	Label
Edwin Fischer	1934	Piano	EMI Records Ltd.
Wanda Landowska	1951	Harpsichord	BGM Music
Rosalyn Tureck	1953	Piano	Deutsche Grammophon
Jorg Demus	1956	Piano	MCA Records, Inc
Ralph Kirkpatrick	1963	Harpsichord	Deutsche Grammonphon
Martin Galling	1964	Harpsichord	Membran Music Ltd.
Malcolm Hamilton	1964	Harpsichord	Everest
Glenn Gould	1965	Piano	Sony Classical
Friedrich Gulda	1972	Piano	Decca
Sviatoslav Richter	1972	Piano	BMG Music
Gustav Leonhardt	1973	Harpsichord	BMG Classics
Joao Carlos Martins	1981	Piano	Labor Records
Anthony Newman	2000	Harpsichord	KHAEON World Music. Inc.

Performer	Year Recorded	Instrument	Label
Anthony Newman	2001	Piano	KHAEON World Music, Inc
Vladimir Ashkenazy	2004	Piano	Decca Record Company Ltd.
Daniel Barenboim	2006	Piano	Warner Classics
Pietro De Maria	2015	Piano	DECCA Music Group Limited

Extraction Tools

We evaluated the outputs of three music content analysis tools: Essentia 2.1 beta5 (Bogdanov et al., n.d., 2013), MIRtoolbox 1.8.1 (Lartillot et al., 2008; Lartillot & Toiviainen, 2007b) and Librosa 0.10.1 (McFee et al., 2015). All three are routinely used in research and industry applications at the time of writing. We selected these tools due to their prominence in the music information retrieval, music cognition, and empirical musicology literature (at the time of writing, Librosa has ~2300 citations on Google Scholar, whereas MIRtoolbox and Essentia respectively have ~1900 and ~640 citations). The implementation of these tools also varies making them more suitable for certain applications depending on the end-users' goals and experience with scripting languages: Librosa is a Python package; Essentia a C++ library with an extensive Python Application Programming Interface (API), and MIRtoolbox is implemented within MATLAB. To streamline our procedure, we performed all analyses in Python, using the MATLAB Engine API in the case of MIRtoolbox. Our implementations of the analysis tools are available at https://github.com/konradswierczek/Musical-Feature-Evaluation-with-Versions.

Features

For all analyses we used default or recommended settings to evaluate the baseline variability of each algorithm and to simulate "out-of-the-box" usage by the typical end-user. Although parameter optimization for specific use cases might increase the figure of merit of an algorithm, our focus here is the relative variability between tools and features rather than an

evaluation of accuracy. Future work may use this method to explore how parameter optimization influences the variability of these tools (see further discussion below). To facilitate readily quantifiable comparisons, we select features that are represented with a single numeric value. In order to establish a baseline understanding of the range of variability, we identify two thought to be invariant (relative mode; number of onsets) and two thought to be variant (spectral centroid; tempo) across performances (see Table 2).

Table 2: Four features suitable for evaluation using the Feature Evaluation with Versions Procedure.

	Spectral	Temporal
Invariant	Relative Mode	Number of Onsets
Variant	Spectral Centroid	Tempo

Mode

Modality, a key aspect of western musical structure, plays a crucial role in conveying musical emotion (Crowder, 1984; Gagnon & Peretz, 2003). Mode is generally defined as the pitch distribution and pitch order of a piece of music and is therefore a structural property that should not vary significantly between versions. Despite changes in pitch distributions between versions due to performer expressive timing and dynamics, in principle the extracted mode should remain relatively consistent. MIRtoolbox uses an audio file mode extraction procedure adapted from the Krumhansl-Schmuckler keyfinding algorithm (Krumhansl, 2001) with additional adaptations from Gómez (2006). This relative mode algorithm returns a value between -1 (minor) and 1 (major), the major and minor modes being the most frequently used in western tonal music. The underlying keyfinding algorithm traditionally relies on the pitch class distribution (PCD) of a musical score or other symbolic notation but relies on chroma features –

sometimes referred to as a chromagram or harmonic pitch class profile (HPCP) – when analyzing audio files. Only MIRtoolbox has a native implementation of the mode extraction algorithm (mirmode), however all three tools can extract chroma features. We therefore wrote a standalone Python version of the mode extraction algorithm from MIRtoolbox to accept chroma features extracted with any tool (see Figure 7 in supplementary materials for the values of the mirmode algorithm directly from MIRtoolbox plotted against the values of our reproduction using chroma features extracted with MIRtoolbox's mirchromagram algorithm (Pearson's r = 1)). The values are identical, indicating our mode extraction algorithm is successfully reproducing the MIRtoolbox mirmode algorithm. MIRtoolbox also implements a Constant-Q Transform that could be applied to the mode algorithm, not considered here, Essentia implements both a Constant-Q Transform and a Fourier Transform, and Librosa implements a Short-Term Fourier Transform, Constant-Q Transform, Constant-Q Transform with CENS, and Variable-Q Transform. Details on these extractors can be found in each of the toolbox's extensive documentations.

Onsets

Tracking onsets in an audio file is a useful mid-level feature for temporal and rhythmic analyses. It forms the basis of beat-tracking, tempo prediction, meter prediction, novelty metrics, and many other high-level features. However, onset analysis is useful here since while the length of versions may vary, the total number of onsets should not change between versions since pianists are playing from the same musical score. A notable exception to this is ornamentation, a common practice in baroque keyboard music where performers add unique elaborations at prescribed moments in the piece, which may cause slight differences in the number of onsets. Further, the speed at which elaborations such as trills are performed may lead to a difference in

the number of onsets. Although MIRtoolbox and Librosa implement only one algorithm for onset extraction (mirevents and onset_detection respectively), Essentia has six. The final numeric value for comparison of this metric is the number of onsets detected in the audio file.

Spectral Centroid

Spectral centroid is the weighted mean of frequency components in a signal, measured in Hertz (Hz), and is often used a simple predictor of the "brightness" of a sound (Klapuri & Davy, 2006) and more generally used in the classification of timbre. Since the timbre of a version is likely to vary depending on the instrument, acoustics, recording technology, and processing used, we would expect the spectral centroid to vary across versions. Each extraction tool used in this study only has one method for extracting spectral centroid.

Tempo

Tempo, measured in Beats per Minute (BPM), is the speed or pace of a piece of music. Extracting tempo is useful for genre and style classification (Tzanetakis & Cook, 2002), predicting emotional appraisal (Eerola et al., 2013), music theoretical analysis of versions (Bach & Palmer, 2004) and other tasks. Written compositions of classical music often have a BPM marking or a text annotation indicating a range of possible BPMs. However, performers can vary considerably in their choice of tempo and may also alter the tempo throughout a performance. Palmer's analysis of the Well Tempered Clavier (Bach & Palmer, 2004) reviews 13 performances and finds significant variation of tempo within pieces. We therefore expect tempo to vary between versions. MIRtoolbox and Librosa each have two methods for extracting tempo whereas we implement three methods from Essentia here.

Algorithm Selection

For all features except Spectral Centroid there are multiple methods for extraction, as described above. Here we select one method for each feature/tool combination. Using MIDI representations of the first eight measures of the 24 preludes from Bach's Well Tempered Clavier, we extract the MIDI tempo, number of onsets, and the mode using the same mode extraction algorithm discussed above on a pitch class distribution. We also synthesize audio files from the MIDI with a generic piano sound font and extract the same features as described above. For each feature/tool combination where more than one algorithm is available, we calculate the mean squared error from the MIDI features (Table 3). The lowest MSE within a feature/tool is selected for subsequent version analysis.

(a) Relative Mode

Tool	Algorithm	MSE
Essentia	cqt	0.032
Essentia	sftf	0.039
Librosa	cens	0.019
Librosa	cqt	0.019
Librosa	stft	0.021
Librosa	vqt	0.032

(b) Number of Onsets

Tool	Algorithm	MSE
Essentia	complex	2902.034
Essentia	flux	3059.708
Essentia	phase	3676.176
Essentia	melflux	4483.407
Essentia	rms	5221.162
Essentia	hfc	6737.353
Librosa	std	3234.676

(c) Tempo (BPM)

Tool	Algorithm	MSE
Essentia	percival	1682.449
Essentia	multifeature	4474.328
Essentia	degara	4530.455
Librosa	onsets	3544.288
Librosa	beattrack	5182.151
MIRtoolbox	metre	3078.521
MIRtoolbox	classical	4001.032

Table 3: Mean Squared Error of synthesized audio compared to MIDI for each feature and algorithm.

Version Variability Ratio

To facilitate comparison between features with dissimilar numeric scales, we propose a metric of relative variability. We compare the variability (here represented by the standard deviation) of all versions of one prelude (for instance, the C Major prelude) to the standard deviation of all versions of all 24 preludes. Values approaching zero would indicate little variability between versions of a piece, while values of 1 would indicate that variability between pieces is similar to that across the entire corpus. Conceptually, invariant features should have smaller variability ratios than variant features. Figure 1 outlines the process used to calculate the version variability ratio for a single prelude/tool/feature combination. We extend this process to all 24 pieces, 3 tools, and 4 features.

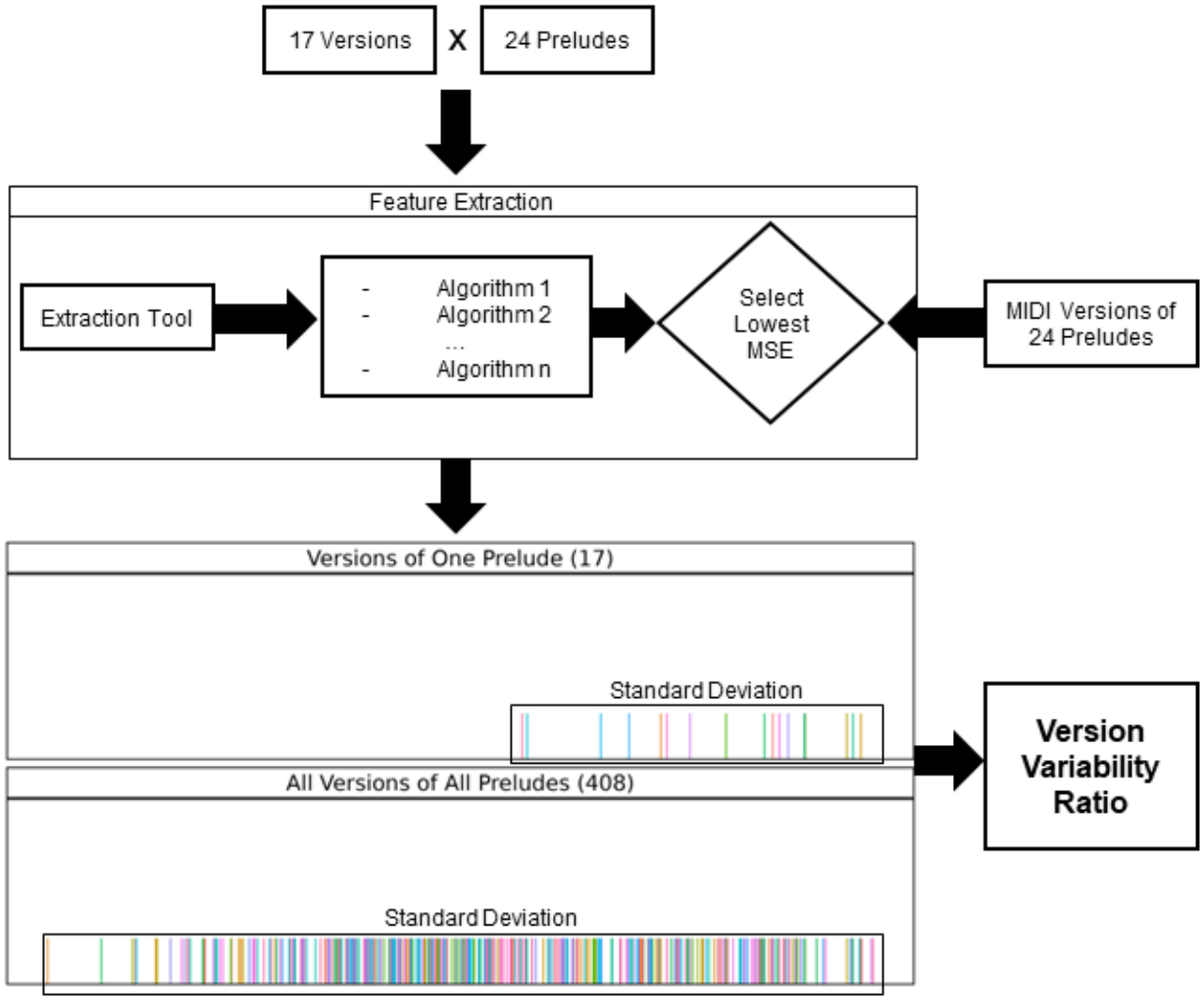

Figure 1: A schematic representation of the version variability ratio.

Questions

To demonstrate possible applications of this method, we identify three evaluation questions for this corpus and feature set. First, how do different implementations of the same feature differ in their consistency? For instance, is any implementation of an invariant feature less consistent than another? Second, are these features consistent with our variant/invariant classifications? That is, are the variant features less consistent across versions than the invariant

features? Finally, do mediating effects such as the instrument of performance or nominal mode (the mode defined by the composer in a title or key signature, importantly distinct from the extracted mode or perceived mode) influence how consistently extracted a given feature is across versions?

Results

Figure 2 shows the version variability ratio of each prelude across features and tools. Distributions of raw values are available in the supplementary materials. As it is dimensionless, the variability ratio provides an intuitive mechanism to assess consistency across both tools and features. The mean values of all preludes for a given feature/tool combination indicate the overall variability of that combination, while the variability of ratios across all pieces indicate the influence of prelude-specific factors (i.e., structural features).

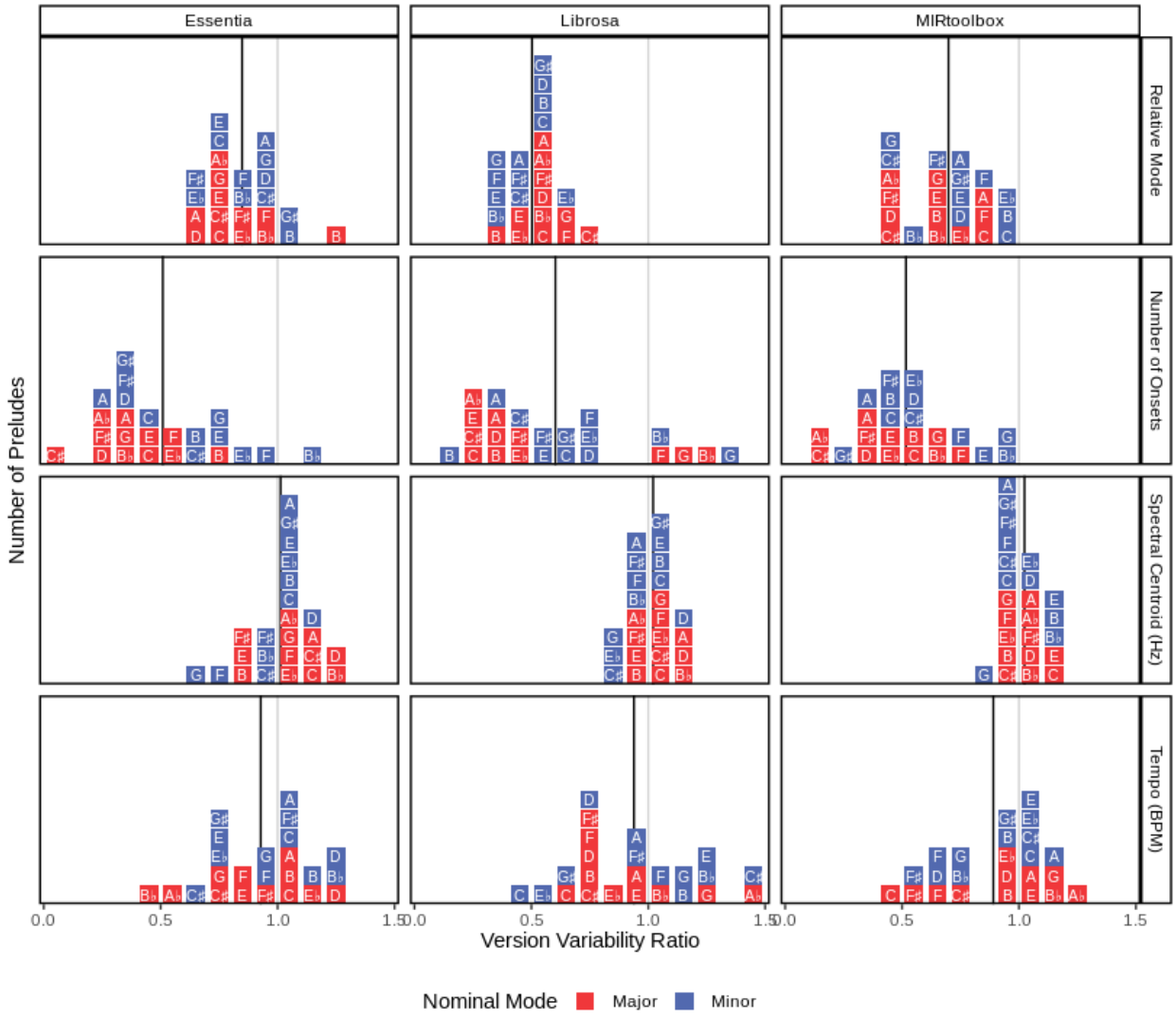

Figure 2: Version variability ratios for all 24 piano preludes. The gray lines indicate a ratio value of 1: the point of equal variability between versions of the same piece and the entire corpus. The black lines indicate the means of all ratios for a given feature/tool combination. Colour indicates the nominal mode: for instance, the C Major prelude corresponds to the red point with the letter 'C'.

To evaluate the three questions discussed above, we use permutation tests on the mean difference in the version variability ratios of the 24 preludes for each pairwise combination of factors due to the non-normality and unequal variances in this data. Permutation tests were performed in R using a customized function available at https://github.com/konradswierczek/Musical-Feature-Evaluation-with-Versions.

Question 1: Between-Tool Feature Variability

We performed permutation tests on pairwise combinations of tools within a given feature (Figure 4). Results revealed a significant difference between all tools for relative mode extractions. Librosa is significantly less variable than both MIRtoolbox and Librosa ($p < 0.01$), and MIRtoolbox is significantly less variable than Essentia ($p < 0.01$). No other significant differences between tools were found for any other feature.

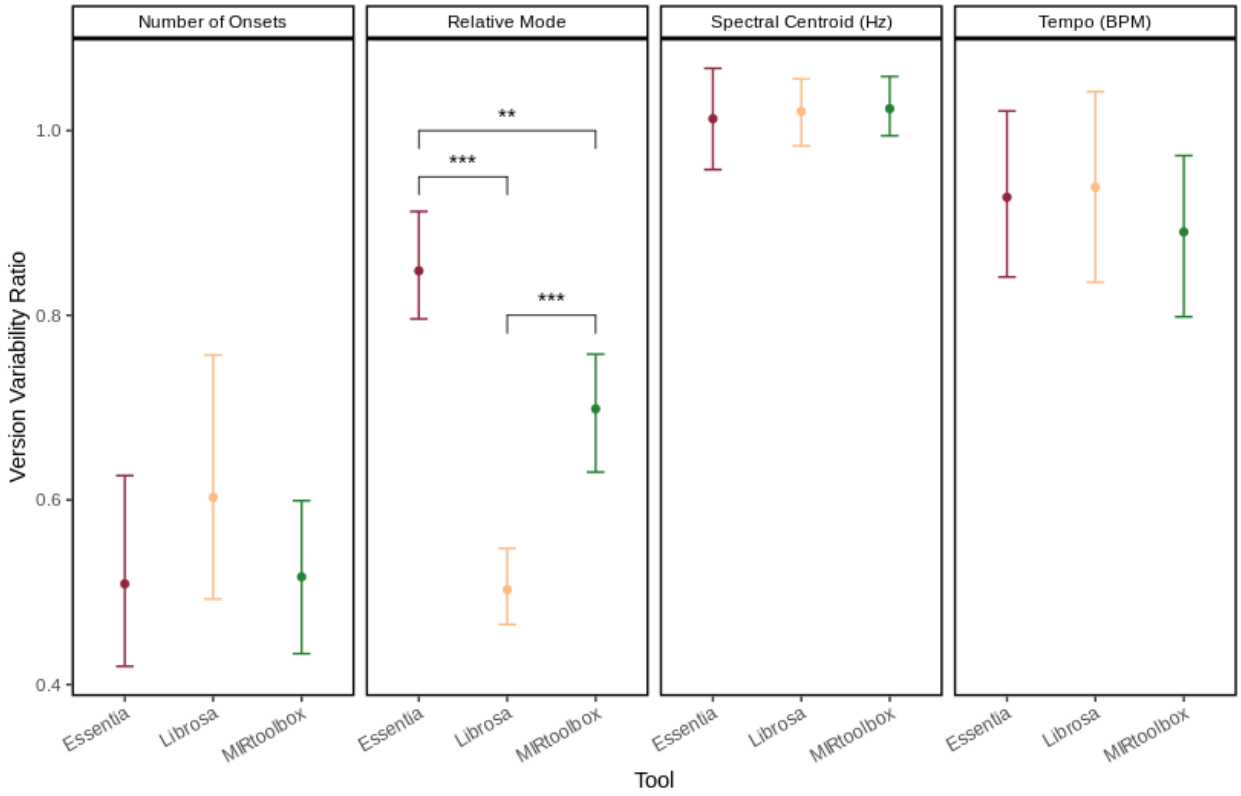

*Figure 3: Version variability ratios for feature/tool combinations. Dots indicate the mean ratio for that combination, plotted with 95% adjusted bootstrap adjusted percentile confidence intervals (R = 1000). Significance Codes: * < 0.05, ** < 0.01, *** < 0.001. Note that these are the same values as Question 2, grouped according to feature rather than tool.*

Question 2: Comparing Variability Between Features

Next, we evaluate the variability between features for each tool. In the same procedure as above, we conducted pairwise permutation tests on pairwise combinations of features for each tool. In respect to Essentia, the number of onsets is significantly less variable than all other features. However, there is no significant difference between tempo and relative mode or spectral centroid. In respect to Librosa, results revealed significant differences between all pairwise

comparisons of features except relative mode and the number of onsets (however, the version variability ratios of librosa indicate greater piece-level variability). Both invariant features (see Figure 4) are less variable than both variant features ($p < 0.001$). Within variant features, tempo is less variable than spectral centroid ($p < 0.05$). Finally, in respect to MIRtoolbox, results revealed significant differences between all pairwise comparisons of features. Specifically, both invariant features (see Figure 4) are less variable than both variant features ($p < 0.001$). Within variant features, tempo is less variant than spectral centroid ($p < 0.001$), while within invariant features, the number of onsets is less variable than relative mode.

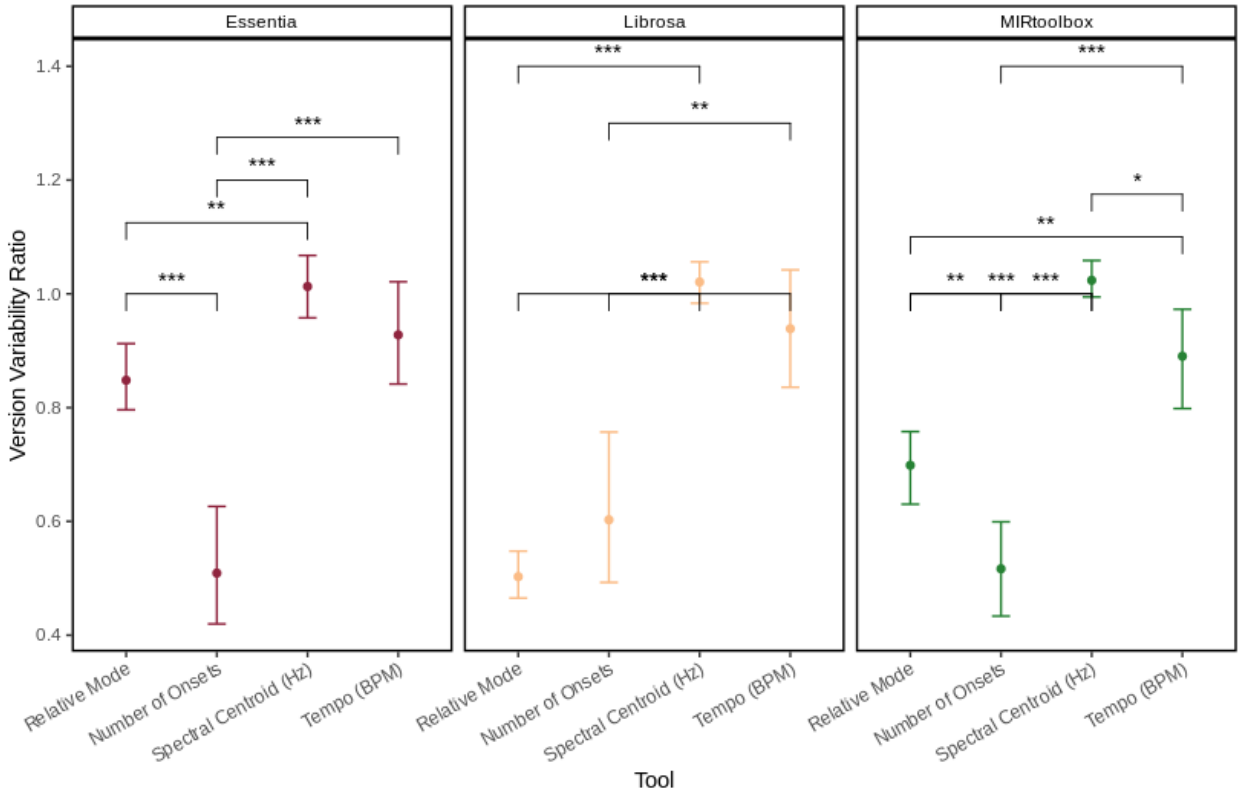

*Figure 4: Version variability ratios for feature/tool combinations. Dots indicate the mean ratio for that combination, plotted with 95% adjusted bootstrap adjusted percentile confidence intervals (R = 1000). Significance Codes: * < 0.05, ** < 0.01, *** < 0.001. Note that these are the same values as Question 1, grouped according to tool rather than feature.*

Question 3: Mediating Effects

We examine the effect of two potential mediating variables (instrument, nominal mode) on the variability of a given feature-tool combination. Permutation tests on the effect of instrument (piano or harpsichord) revealed no sigificant differences ($p > 0.05$) for any feature/tool combination (see Figure 5).

In respect to nominal mode, permutation tests revealed no significant differences ($p > 0.05$) for any feature/tool combination (see Figure 6).

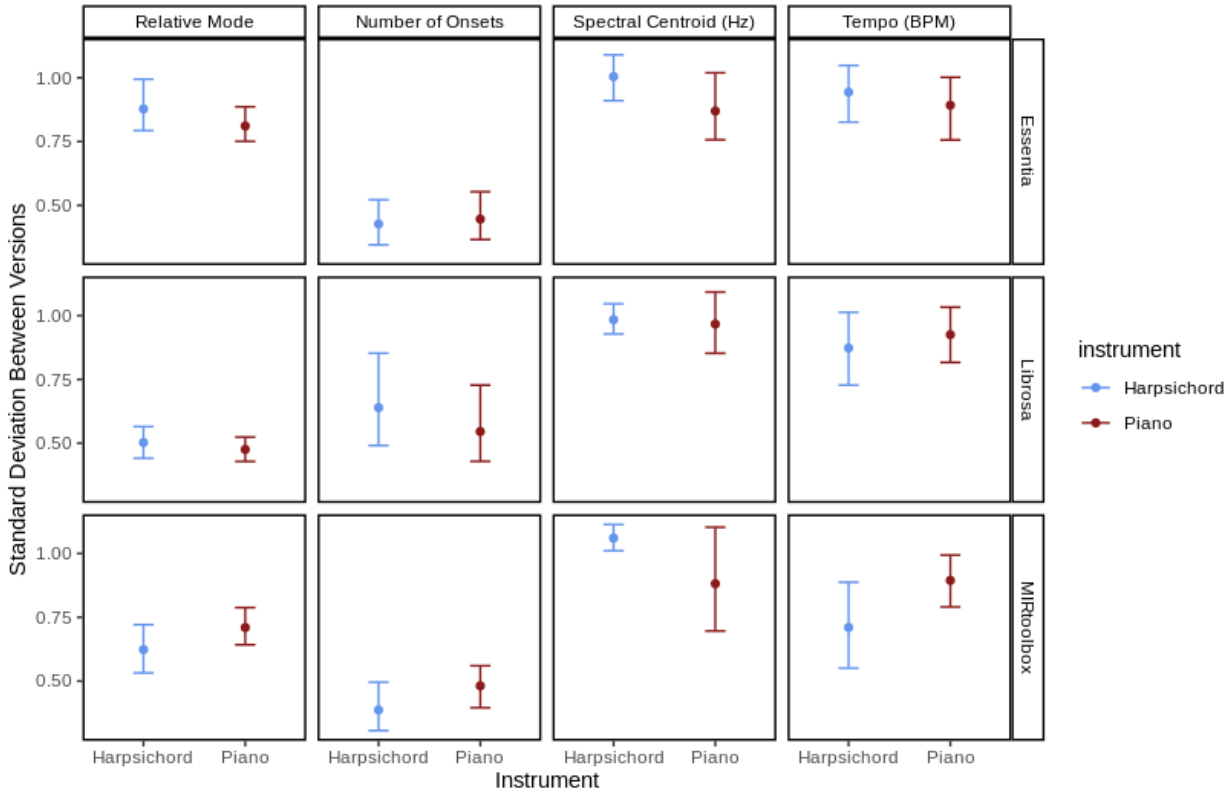

*Figure 5: Version variability ratios for feature/tool combinations, further subset by the instrument used in the performance. Dots indicate the mean ratio for that combination, plotted with 95% adjusted bootstrap adjusted percentile confidence intervals (R = 1000). Significance Codes: * < 0.05, ** < 0.01, *** < 0.001.*

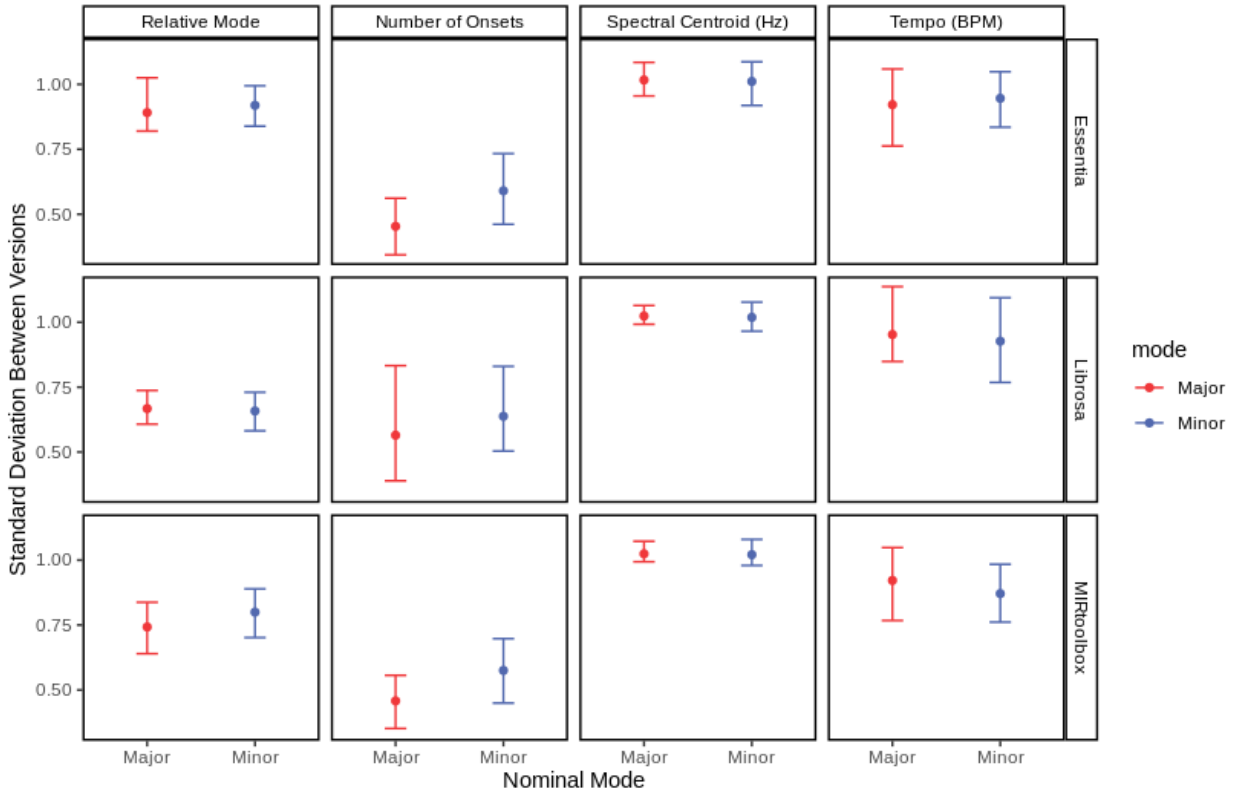

*Figure 6: Version variability ratios for feature/tool combinations, further subset by the nominal mode. Dots indicate the mean ratio for that combination, plotted with 95% adjusted bootstrap adjusted percentile confidence intervals (R = 1000). Significance Codes: * < 0.05, ** < 0.01, *** < 0.001.*

Discussion

Here we demonstrate a method for evaluating music content analysis tools without relying on ground truths. While our objective here has been to formalize and demonstrate this method, we begin our discussion with consideration of the question results.

Between-Tool Feature Variability

Tool comparison is a useful approach to examine the relative performance. It can provide a benchmark to improve upon, and allows for selecting an optimal tool for a given task. While each individual tool may be tested during it's development, independent or third-party tests on different data sets can help support the utility, accuracy or robustness of the tool. To this end, we applied our Version Variability Analysis predictions to this between-tool evaluation. Of particular interest here are the invariant features since we do not have a specific prediction about the degree of variability desired in variant features. Specifically, adopting an approach of minimizing variability between versions for invariant features is a useful benchmark. We find no differences in the variability between tools for the number of onsets, suggesting these three tools perform similarly, regardless of their accuracy or overall consistency. However, in respect to relative mode, we find that Librosa is the least variable option of the three tools examined. Since the primary difference between these three relative mode extractions is the method of chroma feature extraction, this likely means Librosa is least susceptible to pitch independent factors, such as timbre, despite being no less sensitive to timbre in spectral centroid evaluations.

Comparing Variability Between Features

A key assumption of this method is the distinction between variant and invariant features. By comparing the different features to one another, we are able test that assumption: here we compared the features to one another within each tool. Both Librosa and MIRtoolbox support

this assumption: invariant features are less variable than variant features. However, in respect to Essentia, relative mode is not significantly different from tempo (although it is significantly different from spectral centroid). When comparing the invariant features, relative mode is only less variable in the case of Librosa, with number of onsets consistently having the lowest ratios observed across all tools and features. This suggests that despite some variability being expected in the number of onsets due to ornamentation, relative mode is still susceptible to greater unwanted variability than number of onsets.

Mediating Effects

We present two mediating effects that might influence the extraction of these features: instrument, and nominal mode. While these are by no means exhaustive, they are an effective example. Both these features do not significantly change the variability of any of the feature/tool combinations. While we would not expect the performed instrument to influence the invariant features, it seems more likely that the performed instrument would influence the variant features: specifically spectral centroid as a measure of timbre. However, since we compare the versions within a piece to the entire corpus, the variability between specific tracks on an album likely have the same degree of variability as all tracks on all albums, explaining the lack of difference here. Further, in the case of tempo, there is no reason to believe the instrument a performer selects would have an impact on their tempo choices in any systematic fashion. Applying these analyses to other factors, such as structural features of the pieces, may account for the variability seen in the other two questions.

Version Variability Analysis

Beyond the specific results of the three questions considered above, Version Variability Analysis is a viable alternative to conventional accuracy evaluations. Specific applications include:

- Tool/Algorithm Selection: As described above, selection of invariant feature extraction tools can be facilitated by minimizing the variability ratio. For instance, the minimal variability relative mode algorithm in this study would be Librosa. This approach can also be applied to parameter optimization: for instance, in the case of relative mode, this procedure could be applied to selecting optimal weights in the Krumhansl Schmuckler keyfinding algorithm. In the case of variant features, while not explored here, prior distributions from similar musical samples can be used as a variability target (in the case of this sample Palmers analysis of tempi in the WTC may be appropriate (Bach & Palmer, 2004))
- Development and Testing: Applying a priori expectations of how features behave psychologically during development of music content analysis tools would improve these tools at the source. In conjunction with traditional accuracy assessments, this method may aid in identifying the factors that contribute to unwanted variability at a lower level in the processing chain. For instance, if the method by which chroma features are computed in a relative mode algorithm is highly variable, this can be diagnosed and localized with this method.

Limitations

Ornamentation, the practice of adding improvised embellishments at specific points in a piece of music, is a common feature in performances of Bach's music. These embellishments

may introduce small changes in onset distributions that would introduce variability. While these ornaments likely also influence the pitch distributions, they are generally diatonic and therefore should not change the mode evaluation. Despite this, relative mode is still more variable in two of three of the tools examined. While the version variability ratio is a useful metric, the exact range of these values remains unclear. While the invariant features are generally less variable than the variant features, it is unclear how extreme these values may be in a larger set of features.

Future Directions

To maintain the benefits of the WTC discussed above while avoiding the issue of ornamentation, future evaluations with this method might use popular composers from later era's such as Mozart, Beethoven, or Chopin who have similar compositions and are widely recorded but generally do not include ornamentation. This does not account for the speed at which performers execute trills and tremolos, but careful selection of pieces may avoid that confounding variable. Further, while we focus on classical piano compositions due to the nature of the variant and invariant features in this style, the assumptions about what features are variant and invariant likely can be adapted to different styles of music.

Differences in variability between pieces indicate that structural features may have an influence on the extraction of other seemingly unrelated features. Future work will examine the effect of structural features like tempo, dynamics, articulation, and pitch height on the extraction of features that should not be influenced by these changes (i.e., chroma features and onset detection). Further, differences between versions may point to the effect of performance and audio features on seemingly unrealted features. While previous work has shown that chroma and timbre features (MFCCs) are robust to changes in audio quality, it is unclear if this extends to

other features, or to manipulations in performance factors like acoustics and recording technology.

While we focus on three tools and four features here, the version variability ratio may be useful in applications outside of music. In principle, the concept of invariant and variant features is present in other domains (for instance, linguistics) making this a useful approach for evaluating non-musical feature extraction or prediction algorithms in other data types.

Conclusion

Here we introduce a novel method for evaluating the output of music content analysis tools without relying on ground truth, while applying understanding of music perception and practice. Since these tools are used frequently both in research and industry contexts, it is important to develop frameworks for evaluating these tools to ensure that the extracted features have construct validity, and to ensure they truly measure what we think they are measuring: in the case of strictly musical features, some approximation of the human experience. In a broader sense, we aim to contribute to a rich tradition of improving the efficacy of MIR through evaluative frameworks.

References

Bach, J. S. (1963a). *Bach: The Well-Tempered Clavier Book I [Recorded by R. Kirkpatrick]*. [CD]. Germany: Polydor International.

Bach, J. S. (1963b). *The Well-tempered Clavier, Part 1 [Recorded by R. Kirkpatrick]*. [Vinyl]. Deutsche Grammophon.

Bach, J. S. (1964). *Bach: The Well-Tempered Clavier Book One and Book Two [Recorded by M. Hamilton]*. [Vinyl]. Everest.

Bach, J. S. (1973a). Bach: The Well-Tempered Clavier, Book I - BWV 846-869 [Recorded by A. Newman]. In *Bach: The Well-Tempered Clavier, Book I*. [Vinyl]. Columbia Records.

Bach, J. S. (1973b). *Bach: Well-Tempered Clavier Book 1 [Recorded by A. Newman]*. [Vinyl]. Columbia Masterworks.

Bach, J. S. (1987). *Bach: The Well-Tempered Clavier Book I [Recorded by W. Landowska]*. [CD]. BMG Music.

Bach, J. S. (1989). *Bach: The Well Tempered Clavier Book I [Recorded by G. Leonhardt]*. [CD]. BMG Classics. (Original work published 1973).

Bach, J. S. (1992a). *J.S. Bach: The Well-Tempered Clavier Das Wohltemperierte Klavier [Recorded by S. Richter]*. [CD]. BMG Music. (Original work published 1970).

Bach, J. S. (1992b). *The Well Tempered Clavier, Book One [Recorded by J. Demus]*. [CD]. MCA Records, Inc. (Original work published 1956).

Bach, J. S. (1993). *Bach: The Well-Tempered Clavier I [Recorded by G. Gould]*. [CD]. Sony Classical. (Original work published 1963, 1964, & 1965).

Bach, J. S. (1994). *The Well Tempered Clavier, Book One [Recorded by J. Martins]*. [CD]. Labor Records. (Original work published 1964).

Bach, J. S. (1995). *Bach: The Well-Tempered Clavier Book I [Recorded by F. Gulda]*. [CD]. Decca Classics Production (Original work published 1972).

Bach, J. S. (1999). *Bach: The Well-Tempered Clavier Book 1 & 2 [Recorded by R. Tureck]*. [CD]. Deutsche Grammonphon. (Original work published 1953).

Bach, J. S. (2001). *Bach: The Well-Tempered Clavier Book I: Two complete recordings on piano and harpsichord [Recorded by A. Newman]*. [CD]. KHAEON World Music, Inc.

Bach, J. S. (2006a). *Bach: The Well Tempered Clavier Book I [Recorded by D. Barenboim]*. [CD]. Warner Classics.

Bach, J. S. (2006b). *Bach: The Well-Tempered Clavier, Book I - BWV 855 [Recorded by V. Ashkenazy]*. [CD]. DECCA Record Company Limited.

Bach, J. S. (2006c). *The Well-Tempered Clavier - BWV 846–893 [Recorded by M. Galling]*. [CD]. Membran Music Ltd.

Bach, J. S. (2007). *Bach: The Well Tempered Clavier [Recorded by E. Fischer]*. [CD]. EMI Records Ltd. (Original work published 1989).

Bach, J. S. (2015). *The Well-Tempered Clavier Book I [Recorded by P. De Maria]*. [CD]. DECCA Music Group Limited (Original work published 1722).

Bach, J. S., & Palmer, W. A. (2004). *The well-tempered clavier. Volume 1* (3rd ed). Alfred Pub. Co.

Battcock, A., & Schutz, M. (2019). Acoustically Expressing Affect. *Music Perception*, *37*(1), 66–91. https://doi.org/10.1525/mp.2019.37.1.66

Battcock, A., & Schutz, M. (2021). Individualized interpretation: Exploring structural and interpretive effects on evaluations of emotional content in Bach's Well Tempered

Clavier. *Journal of New Music Research*, *50*(5), 447–468. https://doi.org/10.1080/09298215.2021.1979050

Battcock, A., & Schutz, M. (2022). Emotion and expertise: How listeners with formal music training use cues to perceive emotion. *Psychological Research*, *86*(1), 66–86. https://doi.org/10.1007/s00426-020-01467-1

Bogdanov, D., Wack, N., Gomez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & Serra, X. (n.d.). *ESSENTIA: AN AUDIO ANALYSIS LIBRARY FOR MUSIC INFORMATION RETRIEVAL*.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & Serra, X. (2013). ESSENTIA: An open-source library for sound and music analysis. *Proceedings of the 21st ACM International Conference on Multimedia*, 855–858. https://doi.org/10.1145/2502081.2502229

Cataltepe, Z., Yaslan, Y., & Sonmez, A. (2007). Music Genre Classification Using MIDI and Audio Features. *EURASIP Journal on Advances in Signal Processing*, *2007*(1), 036409. https://doi.org/10.1155/2007/36409

Chen, T.-P., & Su, L. (2021). Attend to Chords: Improving Harmonic Analysis of Symbolic Music Using Transformer-Based Models. *Transactions of the International Society for Music Information Retrieval*, *4*(1), 1–13. https://doi.org/10.5334/tismir.65

Crowder, R. G. (1984). Perception of the major/minor distinction: I. Historical and theoretical foundations. *Psychomusicology: A Journal of Research in Music Cognition*, *4*(1-2), 3–12. https://doi.org/10.1037/h0094207

Cunningham, S. J., Bainbridge, D., & Downie, J. S. (2012). *THE IMPACT OF MIREX ON SCHOLARLY RESEARCH (2005 – 2010)*.

Downie, J. S. (2004). The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future. *Computer Music Journal*, *28*(2), 12–23. https://doi.org/10.1162/014892604323112211

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(4), 247–255. https://doi.org/10.1250/ast.29.247

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00487

Gagnon, L., & Peretz, I. (2003). Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies. *Cognition and Emotion*, *17*(1), 25–40. https://doi.org/10.1080/02699930302279

Gómez, E. (2006). Tonal Description of Polyphonic Audio for Music Content Processing. *INFORMS Journal on Computing*, *18*(3), 294–304. https://doi.org/10.1287/ijoc.1040.0126

Gotham, M., Micchi, G., López, N. N., & Sailor, M. (2023). When in Rome: A Meta-corpus of Functional Harmony. *Transactions of the International Society for Music Information Retrieval*, *6*(1), 150–166. https://doi.org/10.5334/tismir.165

Hu, N., & Dannenberg, R. B. (n.d.). *A Bootstrap Method for Training an Accurate Audio Segmenter*.

Huron, D. (2002). Music Information Processing Using the Humdrum Toolkit: Concepts, Examples, and Lessons. *Computer Music Journal*, *26*(2), 11–26. https://doi.org/10.1162/014892602760137158

Klapuri, A., & Davy, M. (Eds.). (2006). *Signal processing methods for music transcription*. Springer.

Krumhansl, C. L. (2001). *Cognitive foundations of musical pitch* (1. issued paperb). Oxford Univ. Press.

Kumar, N., Kumar, R., & Bhattacharya, S. (2015). Testing reliability of Mirtoolbox. *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 710–717. https://doi.org/10.1109/ECS.2015.7125004

Lahdelma, I., Eerola, T., & Armitage, J. (2022). Is Harmonicity a Misnomer for Cultural Familiarity in Consonance Preferences? *Frontiers in Psychology*, *13*, 802385. https://doi.org/10.3389/fpsyg.2022.802385

Lartillot, O., & Toiviainen, P. (2007b). *A Matlab Toolbox for Musical Feature Extraction from Audio*. 8.

Lartillot, O., & Toiviainen, P. (2007a). *A Matlab Toolbox for Musical Feature Extraction from Audio*.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab Toolbox for Music Information Retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 261–268). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_31

Lieck, R., Moss, F. C., & Rohrmeier, M. (2020). The Tonal Diffusion Model. *Transactions of the International Society for Music Information Retrieval*, *3*(1), 153. https://doi.org/10.5334/tismir.46

Mauch, M., & Ewert, S. (n.d.). *THE AUDIO DEGRADATION TOOLBOX AND ITS APPLICATION TO ROBUSTNESS EVALUATION*.

McEnnis, D., McKay, C., Fujinaga, I., & Depalle, P. (n.d.). *JAUDIO: A FEATURE EXTRACTION LIBRARY*.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). *Librosa: Audio and Music Signal Analysis in Python*. 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

Moffat, D., Ronan, D., & Reiss, J. D. (2015). *An Evaluation of Audio Feature Extraction Toolboxes*.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916. https://doi.org/10.1121/1.3642604

Raffel, C., & Ellis, D. P. W. (n.d.). *LARGE-SCALE CONTENT-BASED MATCHING OF MIDI AND AUDIO FILES*.

Raś, Z. W., Wieczorkowska, A. A., & Kacprzyk, J. (Eds.). (2010). *Advances in Music Information Retrieval* (Vol. 274). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11674-2

Schmuckler, M. A., & Tomovski, R. (2005). Perceptual Tests of an Algorithm for Musical Key-Finding. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 1124–1149. https://doi.org/10.1037/0096-1523.31.5.1124

Sturm, B. L. (2016). The "Horse" Inside: Seeking Causes Behind the Behaviors of Music Content Analysis Systems. *Computers in Entertainment*, *14*(2), 1–32. https://doi.org/10.1145/2967507

Temperley, D. (1999). What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered. *Music Perception*, *17*(1), 65–100. https://doi.org/10.2307/40285812

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302. https://doi.org/10.1109/TSA.2002.800560

Urbano, J., Bogdanov, D., Herrera, P., Gomez, E., & Serra, X. (2014). *WHAT IS THE EFFECT OF AUDIO QUALITY ON THE ROBUSTNESS OF MFCCs AND CHROMA FEATURES?*

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, *41*(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Zheng, Y., Liu, J., & Zhang, W. S. (2023). Cover Song Identification Technologies: A Survey. *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, 38–42. https://doi.org/10.1145/3638884.3638891

Supplementary Materials

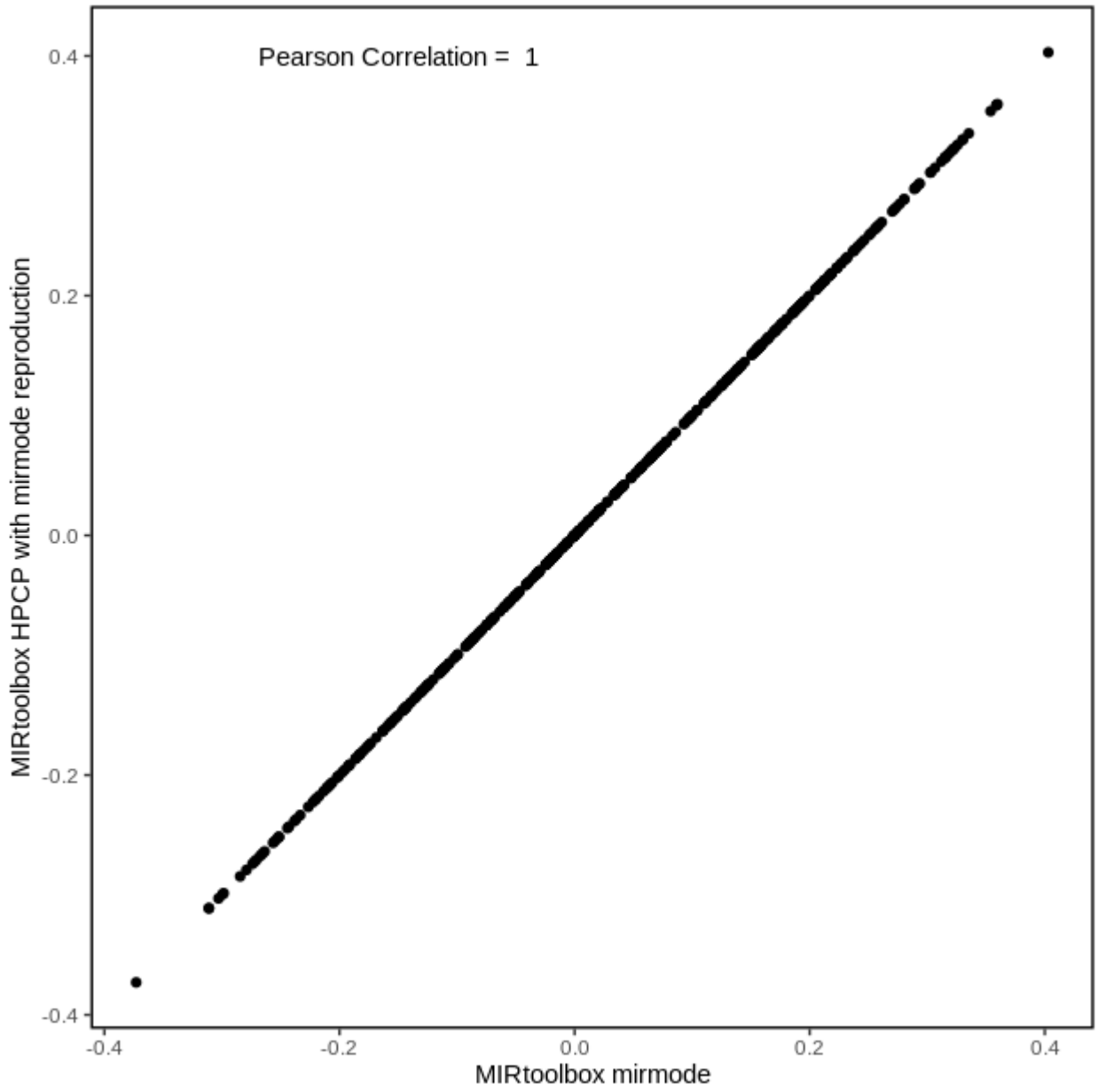

Figure 7: mirmode values for MIRtoolbox mirmode and MIRtoolbox mirchromagram HPCP with mirmode reproduction.

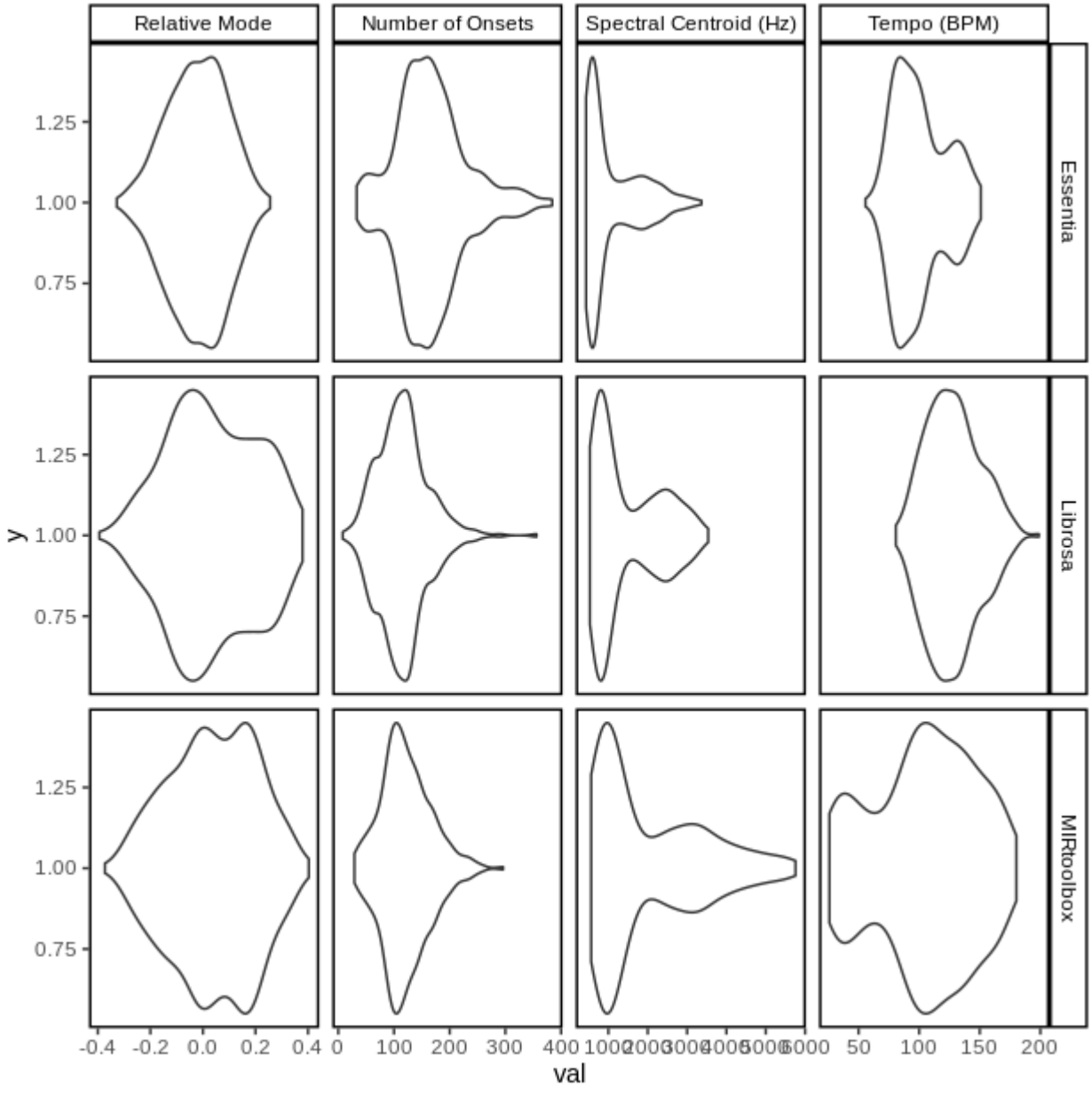

Figure 8: Distributions of original extracted values for each feature/tool combination.