TISMIR

Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription

DATASET ARTICLE

HANS-ULRICH BERENDES
ABHIRUP SAHA
BEN MAMAN
VLORA ARIFI-MÜLLER
MEINARD MÜLLER

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

Orchestral music poses significant challenges for automatic music transcription (AMT) due to its dense polyphony, diverse instrument timbres, and varied acoustic conditions. While AMT research for solo instruments and chamber music has advanced considerably, orchestral transcription remains underexplored, primarily because of the lack of high-quality, time-aligned datasets. In this work, we make three main contributions. First, we introduce the Beethoven Symphony Excerpt Dataset (BSED), a carefully curated, publicly available benchmark for evaluating orchestral AMT systems. BSED contains 20 short excerpts from Beethoven's nine symphonies, each provided in five audio versions: four distinct concert recordings and one synthetic rendition, with an approximate total duration of 37 min. For each excerpt, we supply symbolic scores in multiple formats (PDF, MusicXML, Sibelius, MIDI, CSV) alongside high-quality note-level annotations obtained through robust, manually verified score–audio alignment, refined with transcription-based onset features. Second, to facilitate model development, we release the Beethoven Symphony Dataset (BSD), a large-scale orchestral training set comprising 62 h of public-domain recordings with time-aligned annotations derived from digital scores and structurally verified. While less controlled than BSED, BSD offers rich diversity across performances, conductors, and acoustic environments. Third, we establish a baseline for orchestral AMT by training an instrument-agnostic note-level transcription model on BSD and evaluating it on both BSED and the independent PHENICX dataset. Together, BSED and BSD address a critical gap in orchestral music information retrieval resources and provide a foundation for advancing AMT research toward more robust and generalizable systems.

CORRESPONDING AUTHOR:
Hans-Ulrich Berendes
International Audio Laboratories Erlangen, Germany
hans-ulrich.berendes@audiolabs-erlangen.de

KEYWORDS:
automatic music transcription, orchestral music, dataset, score–audio alignment

TO CITE THIS ARTICLE:
Berendes, H.-U., Saha, A., Maman, B., Arifi-Müller, V., & Müller, M. (2026). Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription. *Transactions of the International Society for Music Information Retrieval,* 9(1), 405–422.
DOI: https://doi.org/10.5334/tismir.343

1 INTRODUCTION

Orchestral music has been at the heart of Western art music for more than three centuries, evolving from the smaller ensembles of the Baroque era into the large, versatile orchestras of today (Del Mar, 1983). Built around the core families of strings, woodwinds, brass, and percussion, and often enriched with additional instruments, the orchestra offers a wide range of timbres, dynamics, and textures. Composers have used these possibilities to create works of remarkable timbral complexity and expressive depth, influencing musical traditions and reaching audiences worldwide. Among them, Ludwig van Beethoven occupies a central place. Born in 1770, he emerged as a pivotal figure whose nine symphonies bridged the Classical and Romantic eras and redefined the expressive and structural possibilities of orchestral composition (Solomon, 1998). His innovations in form, orchestration, and thematic development have made these works cornerstones of the concert repertoire.

From a music information retrieval (MIR) perspective, orchestral music offers both exciting opportunities and significant challenges. Tasks such as score-to-audio alignment (Müller, 2021), automatic music transcription (AMT) (Benetos et al., 2019), instrument separation and recognition (Cano et al., 2019; Krause and Müller, 2023), expressive performance analysis (Lerch et al., 2020), structural segmentation (Nieto et al., 2020), and music synthesis (Maman et al., 2025) are complicated by the orchestra's dense polyphony, wide timbral range, and the variability introduced by different interpretations and acoustic environments. While substantial progress has been made for solo instruments and chamber ensembles, often supported by high-quality, time-aligned datasets, advances in the orchestral domain have been far more limited. A major factor is the scarcity of publicly available resources that combine real-world orchestral recordings with full multi-instrument scores that are accurately aligned. Therefore, many studies have relied on synthetic audio (Garcia-Martinez et al., 2025), which fails to capture the subtleties of real performance, or have focused on smaller ensembles where annotation is more straightforward (Balke et al., 2025; Duan et al., 2010; Li et al., 2019). As a result, data-driven methods for orchestral AMT and related MIR tasks still lack the domain-specific resources required to address the full complexity of the orchestral repertoire.

Building on these observations, we developed resources to address this need by enabling reliable evaluation of transcription systems and by providing both a high-quality benchmark dataset for testing and a large-scale corpus for training orchestral AMT models. In particular, this work makes three main contributions, which we outline below.

Our first contribution is the Beethoven Symphony Excerpt Dataset (BSED), a benchmark designed to capture the richness and variability of orchestral performance while maintaining precise, high-quality annotations. BSED comprises selected short passages from Beethoven's nine symphonies, each available in multiple real-world performances as well as in a synthetic reference version. The dataset provides symbolic scores in multiple formats, accompanied by note-level annotations obtained through manually verified score–audio alignment. Additionally, we refine this initial alignment on the note level with onset-based features. This combination offers a controlled and thoroughly validated test bed for the systematic evaluation of orchestral transcription systems. Figure 1 illustrates the content of the BSED.

Our second contribution is a large-scale training dataset for AMT, the Beethoven Symphony Dataset (BSD). BSD contains 63 h of public-domain recordings of all nine Beethoven symphonies, paired with time-aligned annotations that preserve the structural choices made in each performance. Although the BSD undergoes less manual verification compared to the BSED due to its large scale, its breadth of conductors, ensembles, and recording conditions exposes models to a wide range of timbres, acoustics, and interpretive styles, making it a valuable resource for developing robust and generalizable orchestral AMT systems.

Our third contribution is a baseline study that demonstrates the practical value of these resources. We trained an instrument-agnostic AMT model on BSD and evaluated it on BSED as well as on an independent orchestral dataset. This experiment highlights the benefits of using domain-specific orchestral data, while also revealing the challenges that remain for achieving high transcription accuracy in complex, real-world orchestral settings. Together, these contributions establish a foundation for advancing research in orchestral AMT and related MIR tasks.

The datasets can be accessed on Zenodo.[1] Additionally, we provide a GitHub repository with code related to the dataset, enabling the recreation of the dataset and annotations from the source files.[2] We welcome community contributions and feedback on potential errors via the GitHub repository.

The remainder of this article is organized as follows. Section 2 reviews related work. Section 3 describes the BSED benchmark. Section 4 presents the BSD training dataset. Section 5 reports on the baseline AMT experiments. Section 6 concludes the paper and outlines directions for future research.

2 RELATED WORK

2.1 SYMBOLIC SHEET MUSIC–ENCODING

Creating reliable symbolic score datasets is a long-standing challenge in MIR, particularly when combined with orchestral recordings. While MIDI provides

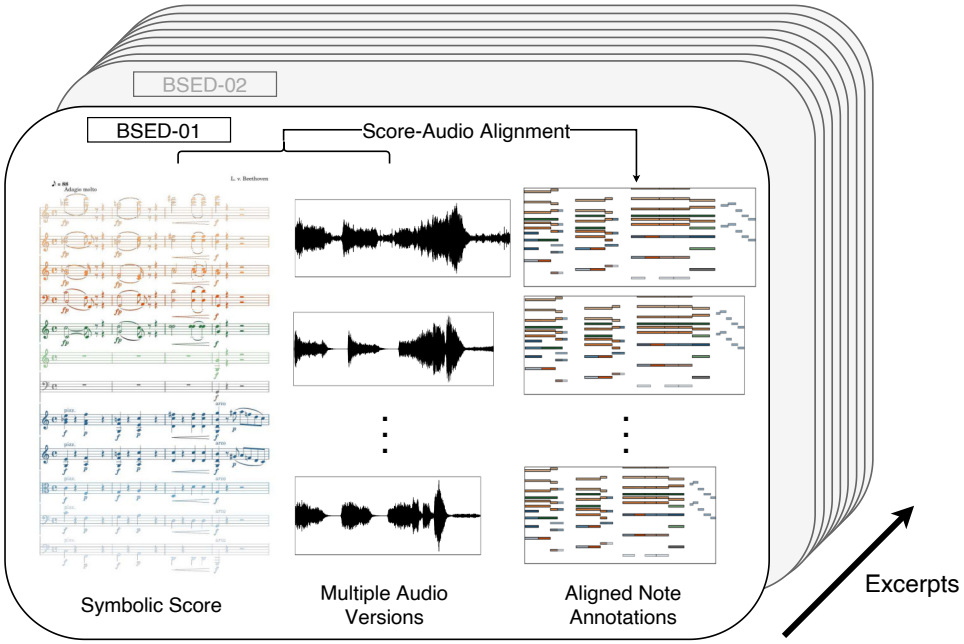

Figure 1 Overview of the Beethoven Symphony Excerpt Dataset. The dataset contains 20 short symbolic score excerpts from Beethoven's symphonies, each paired with five audio versions from different performances, including one synthetic version. Shown here as an example is the first excerpt only. All excerpts feature precisely aligned note-level annotations with pitch and instrument labels, offering a controlled, well-verified benchmark for testing orchestral automatic music transcription systems.

a basic note-level representation, richer formats such as MusicXML, **kern, MEI, MuseData, or LilyPond capture articulations, ornaments, and structural markers, though each comes with trade-offs between readability, interoperability, and machine processing (Good, 2001; Huron, 2002; Nienhuys and Nieuwenhuizen, 2003; Selfridge-Field, 1997). For orchestral repertoire, several large collections exist. The MuseData database[3] includes many Western classical music scores, including all nine Beethoven symphonies, which we use as a foundation in this work. Related resources include KernScores (Sapp, 2005) and the crowdsourced MuseScore library. The OpenScore Orchestra dataset,[4] which is based on Gotham et al. (2022), also provides the score of Beethoven's symphonies but was developed independently of our efforts.

2.2 DATASETS FOR AMT

Beyond symbolic encoding, a wide range of datasets has been developed for AMT, using different strategies to obtain annotations. These datasets can be broadly categorized into solo-instrument and multi-instrument collections, further subdivided into multi-channel (facilitating track-wise annotation) and full-mix recordings. Annotation strategies include capturing note data with specialized hardware (e.g., MIDI-ready piano), aligning scores to recordings, or using transcription models. Audio material may come from existing recordings, purpose-made performances, or synthesized scores. While synthesized audio enables perfect annotations, the gap between synthetic and real recordings can hinder generalization (Maman and Bermano, 2022; Zehren et al., 2024).

Solo-instrument datasets. The piano is the most extensively studied instrument for AMT. Datasets such as MAESTRO (Hawthorne et al., 2019), MAPS (Emiya et al., 2010), and SMD (Müller et al., 2011) provide time-aligned annotations, often via a MIDI-ready piano. The ASAP dataset (Foscarin et al., 2020) adds aligned sheet music. More recent resources include PiJAMA (Edwards et al., 2023) and ATEPP (Zhang et al., 2022), which rely on automatic transcription for annotation. Beyond piano, datasets exist for guitar (Riley et al., 2024b; Xi et al., 2018), bass (Abeßer and Schuller, 2017), violin (Tamer et al., 2022), and drums (Weber et al., 2025).

Multi-instrument datasets. In Western classical music, multi-instrument datasets typically focus on chamber ensembles and small orchestras. Annotation is usually performed via score–audio alignment (e.g., MusicNet, Thickstun et al. (2017)) or through synchronized multi-track recordings (e.g., Bach10 (Duan et al., 2010); URMP (Li et al., 2019); ChoraleBricks (Balke et al. 2025)). While multi-track data allows flexible re-mixing and semi-automated annotation, such datasets are usually small due to the high cost of controlled recording. Synthetic collections such as Slakh (Manilow et al., 2019) and EnsembleSet (Sarkar et al., 2022) provide a larger scale but lack real-world performance characteristics.

Orchestral datasets. Resources for full orchestra remain scarce. The PHENICX Anechoic dataset (Miron et al., 2016; Pätynen et al., 2008) and the Beethoven Anechoic dataset (Böhm et al., 2021) provide separately recorded orchestral parts under controlled conditions but are limited in duration (ca. 10–20 min). More recently, the SynthSOD dataset (Garcia-Martinez et al., 2025) has

offered large-scale synthetic orchestral data. However, to date, no dataset has combined existing real orchestral recordings with aligned score information—a gap the present work seeks to address.

2.3 SCORE–AUDIO ALIGNMENT AND STRUCTURAL INCONSISTENCIES

Closely related to dataset creation is the task of score–audio alignment, which maps the symbolic timeline of a score to the physical timeline of a performance. A common approach is *sequence alignment*, where methods such as dynamic time warping (DTW) (Müller, 2021; Müller et al., 2021) warp the score axis to match the recording. Chroma and onset features are typically used for robustness across conditions, while more recent approaches employ transcription-based features (Maman and Bermano, 2022; Zeitler et al., 2024a).

While sequence alignment provides reliable measure-level mappings, it cannot capture fine-grained timing deviations. For instance, cases where multiple notes are notated at the same measure positions but played with slightly different timing cannot be resolved by sequence alignment. *Note-level alignment* addresses this by matching individual note onsets (and sometimes offsets) to their precise timing in the audio. Applications include piano (Niedermayer and Widmer, 2010), orchestral recordings (Miron et al., 2014), and MIDI-to-MIDI alignment (Peter et al., 2023). Recent refinements again leverage transcription-based features for improved precision (Riley et al., 2024a; Saha et al., 2026).

A major challenge is handling *structural inconsistencies* between notated scores and performances, such as omitted or repeated sections or mismatched measure numbering. Fremerey et al. (2010) laid the foundation for structure-aware synchronization by modeling repeats and jumps explicitly. Building on this idea, subsequent work has proposed strategies such as excluding non-matching pairs (Thickstun et al., 2017), adapting performances to a reference structure (Zeitler et al., 2024b), encoding measure numbering for cross-version transfer (Gotham et al., 2023), or extending alignment algorithms to account for skipped sections (Grachten et al., 2013). Together, these approaches show that robust alignment requires not only temporal accuracy but also explicit treatment of structural variation. These issues are also relevant in orchestral settings, where interpretation of symphonies is shaped by the conductor, both in overall tempo and tempo changes, as well as in decisions about which repetitions are performed.

2.4 AMT

AMT aims to convert audio into symbolic note representations, with methods ranging from frame-level pitch detection to full score generation (Benetos et al., 2019). Recent research has focused on note-level transcription, which identifies the onsets and offsets of individual notes. Onset-based approaches, first shown to be highly effective for piano (Hawthorne et al., 2018; Kong et al., 2021), have since been extended to strings, winds, and small ensembles (Chang et al., 2024; Gardner et al., 2022; Kim et al., 2023; Riley et al., 2024a; Tamer et al., 2023; Wu et al., 2024). Despite this progress, AMT for orchestral music remains largely unexplored: most studies concentrate on solo instruments or chamber ensembles, and systematic evaluations on orchestral datasets are rare. One exception is Bittner et al. (2022), who applied an instrument-agnostic transcriber to orchestral stems from the PHENICX dataset, though this was limited to isolated instrument groups rather than complete recordings. The dataset proposed in this work addresses this gap by providing a foundation for systematic note-level transcription research on orchestral music.

3 BSED: CURATED TEST SET

3.1 OVERVIEW

The BSED is based on 20 short score excerpts from Beethoven's symphonies and is tailored for the evaluation of AMT models. For each excerpt, we provide symbolic score data together with five corresponding audio renditions: four concert recordings and one synthetic version, yielding 100 renditions in total. For each rendition, we provide note-level annotations obtained through score–audio synchronization methods. The design of the dataset enables robust evaluation of AMT models across both different musical content and diverse acoustic conditions.

Table 1 provides an overview of the excerpts. Each excerpt is specified by an identifier, `BSED-ID`, where the `ID` ranges from `01` to `20`. The excerpts are drawn from various movements across the symphonies. In total, the dataset contains 6342 symbolic note events, which correspond to 31,710 annotated note events within the audio versions. The total duration of all audio versions is approximately 37 min, with individual excerpt durations ranging between 17 and 30 seconds. We first discuss the musical content of the excerpts, followed by a more technical description of the symbolic data. Afterwards, we discuss the audio data and the alignment process.

3.2 THE EXCERPTS IN DETAIL

Twenty excerpts were selected from the corpus of Beethoven's Symphonies to represent the diverse musical content. However, we purposefully excluded the fourth movement of Symphony No. 9, as it uniquely features choir and unpitched percussion, instead focusing on the more consistent orchestral instrumentation found across the other symphonies. Ten out of the 20 excerpts were chosen manually by the authors to capture musically meaningful phrases and iconic themes, many of which are movement openings, such as the opening of the Fifth Symphony. The remaining 10 excerpts were selected at random to introduce unbiased diversity. As a result,

BSED-ID	MovementID	Measures	# Notes	Avg. Dur. [s]	BSED-Real				BSED-Synth
					Split 1	Split 2	Split 3	Split 4	Split 5
BSED-01	Op021-01	1–4	159	26.8	Karajan1963	Ansermet1956	Keilberth1958	Blomstedt1976	Synth
BSED-02	Op021-02	6–16	256	21.8	Krips1962	Paray1959	Nowlin2019	Blomstedt1976	Synth
BSED-03	Op036-01	0–3[5]	73	17.4	Schuricht1948	Leibowitz1961	Schubert2005	Blomstedt1978	Synth
BSED-04	Op036-01	35–48	556	19.5	Paray1959	Karajan1963	Steinberg1964	Blomstedt1978	Synth
BSED-05	Op036-03	1–26	401	17.0	Leibowitz1961	Paray1959	Schubert2005	Blomstedt1978	Synth
BSED-06	Op055-01	1–17	319	21.5	Pfitzner1929	Ormandy1961	Krips1958	Blomstedt1979	Synth
BSED-07	Op055-02	143–148	320	23.1	Markevitch1957	Klemperer1961	CzechNat2012[6]	Blomstedt1979	Synth
BSED-08	Op055-03	231–264[7]	234	21.8	Steinberg1955	Klemperer1961	Steinberg1963	Blomstedt1979	Synth
BSED-09	Op060-01	1–6	72	28.5	Weingartner1952	Karajan1962	Keilberth1959	Blomstedt1979	Synth
BSED-10	Op060-02	98–101	279	21.6	Weingartner1952	Karajan1955	Keilberth1959	Blomstedt1979	Synth
BSED-11	Op067-01	1–21	244	21.2	Furtwaengler1943	Bernstein1963	Karajan1962	Blomstedt1977	Synth
BSED-12	Op067-04	110–120	367	16.7	Melgelberg1937	Schuricht1949	Jochum1955	Blomstedt1977	Synth
BSED-13	Op068-01	1–26	300	27.7	Hurst1963	Lehmann1956	Toscanini1956	Blomstedt1977	Synth
BSED-14	Op068-05	194–201	497	16.9	Kubelik1960	Hurst1963	Cluytens1960	Blomstedt1977	Synth
BSED-15	Op092-02	1–18	171	30.2	Monteux1964	Bernstein1960	Karajan1962	Blomstedt1975	Synth
BSED-16	Op092-04	31–49	525	16.5	Monteux1953	Ormandy1945	Krip1961	Blomstedt1975	Synth
BSED-17	Op093-01	1–20	769	22.2	Leibowitz1961	Ansermet1956	Keilberth1958	Blomstedt1978	Synth
BSED-18	Op093-02	30–35	322	17.8	Leibowitz1961	Ansermet1956	Keilberth1958	Blomstedt1978	Synth
BSED-19	Op125-01	374–385	376	21.7	Klemperer1956	Furtwaengler1955	Leibowitz1961	Blomstedt1980	Synth
BSED-20	Op125-03	3–7	102	27.3	Furtwaengler1955	Fricsay1958	Leibowitz1961	Blomstedt1980	Synth
Total:			**6342**	**37:14 min**	**7:26 min**	**7:36 min**	**6:58 min**	**7:46 min**	**7:29 min**

Table 1 Overview of the excerpts in the BSED with their corresponding audio versions. The measure numbers are inclusive and refer to the Urtext edition (Del Mar, 2020), of which we only use full measures (except for the anacrusis in `BSED-03`). The five audio versions comprise four real performances (BSED-Real) as well as a synthetic rendition (BSED-Sytnh).

these 10 excerpts are not necessarily complete musical phrases. We made sure to have a diverse test dataset in terms of musical content. We will show this at the end of this section.

The performed durations of the individual excerpts are in the range of 15–30 s. For each excerpt, the corresponding measure numbers from the score can be found in Table 1. The measure numbers refer to the Urtext edition by Del Mar (2020), which serves as a basis for the BSED. We use inclusive measure ranges (e.g., measures 1–4 refer to the first four measures of the movement) and only selected complete measures from the score, except for the opening anacrusis of `BSED-03`, which is denoted as measure 0 to remain consistent with the numbering in the score edition.

Figure 2 shows the distribution of total notes by instrument, reflecting overall trends in Beethoven's orchestration. Twelve instruments are represented, grouped into four families: strings, woodwinds, brass, and percussion (timpani only). The violin is by far the most frequently used instrument, while the trombone appears least often, with only 23 notes in total. Some instruments occur in multiple variants (e.g., bass trombone, contrabassoon, piccolo flute), but, for simplicity, we treat these as single instruments here. In the scores, however, they are labeled with their correct specific names.

Figure 3 breaks down the number of notes per excerpt by instrument, color-coded by instrument. Across all excerpts, there are 6342 notes, with individual excerpts ranging from 72 to 769 notes. These variations stem primarily from differences in note density, i.e., number of onsets per time, and, not excerpt duration, as actual playing durations are all within a range of roughly 15–30 s (see Table 1). While most excerpts feature a mix of instrument families, some are dominated by a single group. For example, `BSED-03` contains mainly winds, while `BSED-15` contains mainly strings.

Figure 4 displays the overall pitch distribution, color-coded by instrument. Pitches range from a Contrabass C1 (MIDI number 24) to a Violin G6 (MIDI number 91), with

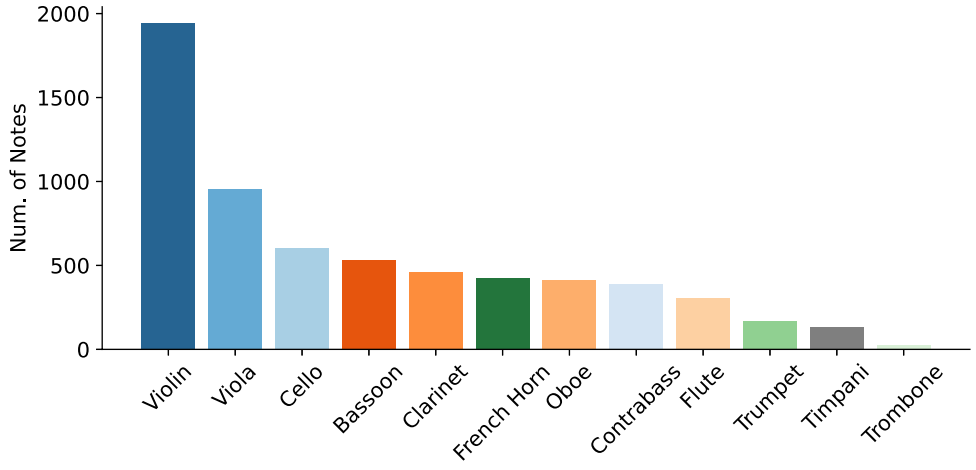

Figure 2 Total number of notes on the symbolic level per instrument in the BSED.

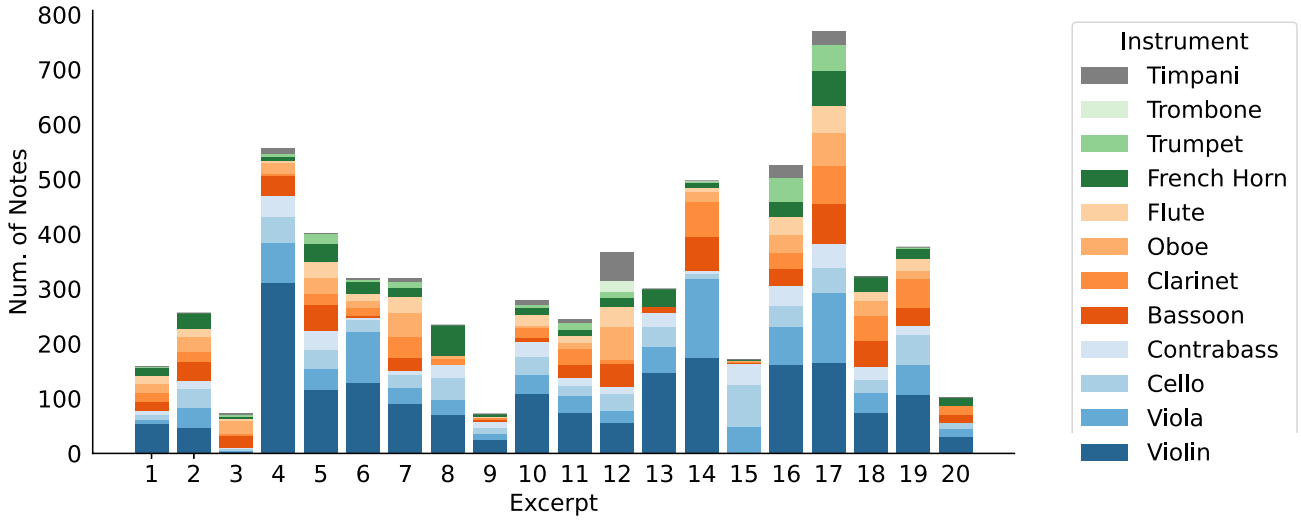

Figure 3 Number of notes on the symbolic level of each BSED excerpt grouped by instrument. Instruments are ordered first by instrument family and then by their total number of notes across all excerpts within each family.

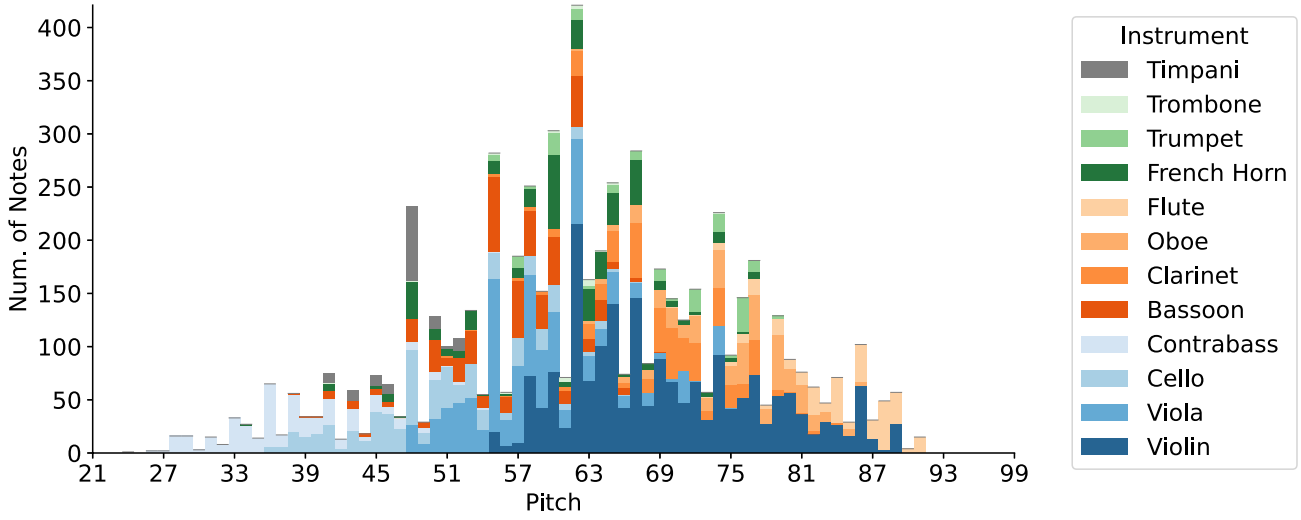

Figure 4 Number of notes on the symbolic level per pitch in the BSED scores grouped by instrument.

D4 (MIDI number 62) being the most frequently occurring pitch.

3.3 SYMBOLIC DATA

The symbolic layer of the BSED contains digital representations of the sheet music for the selected excerpts in various formats. Each score file is named according to the pattern

`BSED – ID_Beethoven_opNr – mvmtNr`

where `BSED-ID` is the unique identifier for each excerpt, `opNr` is the opus number of the corresponding symphony, and `mvmtNr` is the movement number from which the excerpt was taken. Thus, for instance, `BSED-01_Beethoven_Op-021-01` refers to the first excerpt, taken from Symphony No. 1 (Op. 21), first movement. Across the nine symphonies, there are 37 movements in total. To uniquely identify each movement, we use the combination `opNr-mvmtNr`, denoted as `movementID`, for the remainder of this work. The scores are based on the Urtext edition by Del Mar (2020) and were manually created by the authors using the Sibelius notation software.[8] The accuracy of the scores was cross-checked by expert annotators using the Urtext edition as a reference, with further validation performed through score sonification. Despite the availability of existing encodings, such as the aforementioned MuseData database, we created the scores manually to ensure high quality and consistent encoding across all BSED excerpts.

The symbolic scores are provided in the folder `1_ScoreData` in the folllowing five formats:

1. Sheet music in the form of Portable Document Format (PDF) files
2. Sibelius files (dedicated files for Sibelius 8 and Sibelius Ultimate)
3. Music Extensible Markup Language (MusicXML) files exported from Sibelius with the *Dolet 8 plugin for Sibelius*[9]
4. Note lists in `.csv` format
5. MIDI files

The note event lists are derived from the MusicXML files and serve as a MIDI-like encoding enriched with score-specific information. Each note event with a distinct onset is represented by a single row in the `.csv` file; tied notes (i.e., consecutive occurrences of the same pitch connected by a slur and performed as a single sustained note) are merged into one event.

Each note event comprises several attributes (columns in the `.csv`). Onset and offset are first specified as measure positions (`start_meas`, `end_meas`), encoded as single real-valued numbers following Zeitler et al. (2024b). The integer part denotes the measure number, and three decimal digits indicate the relative position within the measure (e.g., in 3/4 time, the second beat of measure 137 is encoded as 137.333). In addition, physical onset and offset times in seconds (`start_sec` and `end_sec`) are provided, computed from the annotated tempo in the score. Further attributes include `pitch` (MIDI pitch number), `instrument`, `velocity` (derived from score dynamics), and `part` (e.g., Violin 1 or Violin 2). A comprehensive description of all attributes is available in `04_Extra/note_encoding_info.md`. The provided MIDI files are directly derived from the note event lists using physical time, pitch, instrument, and velocity information.

Since the dataset is designed for note-level AMT rather than sheet music layout, we apply several simplifications at the notation level (PDF, Sibelius, MusicXML) to ensure consistent processing with software libraries. Dynamic and expression markings are standardized (e.g., *crescendo* indicated by hairpins instead of textual annotations), abbreviations for repeated notes—common in string parts—are written out explicitly, and timpani trills (occurring in `BSED-12`) are realized as individual notes. We provide two versions of the sheet music files: one in concert pitch to avoid parsing errors with respect to sounding pitch, and one retaining the original transposing notation for readability.

3.4 AUDIO DATA

For each of the score excerpts, we provide five different versions, indexed with a `splitID` from 1 to 5, as can be seen in Table 1. Splits 1–4, collectively referred to as BSED-Real, are taken from concert recordings, while split 5, also referred to as BSED-Sytnh, is synthesized using the NotePerformer software[10].

Of the four versions from BSED-Real, splits 1–3 are taken from public domain recordings from the International Music Score Library Project (IMSLP)[11] with varying conductors, while split 4 is always taken from a CD collection of performances conducted by Herbert Blomstedt, which is not in the public domain. The NotePerformer software used for BSED-Synth is a sample-based playback engine that incorporates articulation markings from the score to produce realistic and expressive output directly from the Sibelius scores. However, our primary objective in the synthesized excerpts is to have perfect ground-truth note annotations rather than maximum realism; therefore, we disable NotePerformer's *humanization* feature as well as fermatas in the score, as both would otherwise introduce timing deviations between the annotations and the audio. The total duration of the audio data is 37 min and 14 s.

The audio files stored in the folder `2_Audio` are named according to the format

`BSED – ID_splitID_Beethoven_movementID_versionID`

where `BSED-ID` is the unique ID of the score excerpt and `splitID` denotes the audio split indexed from 1 to 5, as discussed above. The `versionID` typically consists of the

conductor's last name with the four-digit recording year, but, in the case of the synthetic version, we use `Synth`. For example, as can be seen in Table 1,

`BSED – 01_1_Beethoven_Op021 – 01_Karajan1963`

is the first excerpt taken from Op. 21, first movement, cut from a recording of Karajan in 1963.

The recording excerpts were cut manually from the full recordings to ensure the correct start and end of the excerpts. We provide the audio files in the folder `2_Audio` as stereo `.wav` files with a sampling rate of 44.1 kHz. However, some recordings are only upsampled from lower sampling rates and therefore do not contain the full bandwidth, and some only contain a duplicated second channel, instead of proper stereo recordings.

3.4.1 Audio quality

The recording years in the dataset span from 1929 to 2019, leading to substantial variation in audio quality, ranging from digitized vinyl with audible noise to high-quality CD recordings. Such diversity can be valuable for evaluation but can also pose limitations in some cases. Therefore, the five audio splits we provide are approximately ordered by audio quality, from low (`splitID` 1) to high (`splitID` 5). The synthesized versions naturally have no quality degradations and therefore represent split 5. The Blomstedt version, sourced from a CD recording, consistently has a very high recording quality and is therefore always assigned `splitID` 4. The remaining `splitIDs` 1–3 are ordered by relative quality within each excerpt, based on an informal expert listening test. This ordering is excerpt-specific and does not indicate absolute ratings of quality. For instance, recordings within split 1 may range from very poor to moderate quality across different excerpts. An evaluation for each `splitID` independently might yield interesting insights into the robustness against noise for a given transcription model.

3.4.2 Tuning

Upon closer inspection, we found that the recordings have quite diverse tuning reference frequencies. Previous work has found that large tuning deviations can be prohibitive for training AMT models (Edwards et al., 2024). Therefore, we provide two versions of the audio recordings: (1) the natural recording and (2) a version with A4 = 440 Hz reference pitch. To get the 440 Hz version, we employ tuning estimation from the `libfmp` library (Müller and Zalkow, 2021) and pitch shifting with Rubber Band[12]. Because the `libfmp` tuning estimator only captures deviations within a semitone, we rely on transcription-based heuristics to identify and shift larger tuning deviations. The synthesized renditions are always synthesized with A4 = 440 Hz as the reference pitch.

3.5 SCORE–AUDIO ALIGNMENT

To obtain physically aligned note annotations, we apply a two-stage score–audio alignment procedure to map musical onset times (given in measures) to physical onset times (given in seconds).

As discussed in Section 3.5, we distinguish between sequence-level and note-level alignment. Sequence-level alignment provides a robust global mapping from musical time (score domain) to physical time (audio domain), assigning all notes with the same notated onset to a common physical time and thereby preserving structural coherence, which suffices, e.g., for score-following applications (Dannenberg and Raphael, 2006). Note-level alignment refines this mapping by adjusting individual onsets to better match their realized performance timing, resolving local asynchronies and enabling precise note-level analysis. In the following, we describe the resulting semi-automated pipeline.

3.5.1 Sequence-level alignment

Sequence-level alignment computes a global, monotonic mapping between musical and physical time, represented by a DTW warping path. We use the DTW-based alignment procedure of the Sync Toolbox (Müller et al., 2021) and follow Zeitler et al. (2024a) by employing features derived from a pretrained transcription model. Specifically, we use the transcriber of Maman and Bermano (2022), trained on MusicNet (Thickstun et al., 2017), which provides robust pitch-related activations for alignment across diverse orchestral recordings.

All 80 recording-based excerpts in BSED-Real were manually verified using sonifications in which the synchronized note events are superimposed onto the original recordings. If audible misalignments were detected, anchor points were inserted into the DTW procedure to explicitly link selected musical time positions to corresponding physical times. DTW was then re-run between consecutive anchor points while keeping the anchors fixed.

For the sequence-level alignment, we estimate a conservative worst-case residual alignment error of approximately 150 ms. In practice, alignment accuracy is typically substantially better, often within 50 ms or below, as confirmed by listening-based inspection. The sonifications are provided in `04_Extra/Sequence-Alignment_Sonification`.

3.5.2 Note-level alignment

Sequence-level alignment cannot resolve fine-grained temporal dispersion between notes that share the same musical onset time but are performed with slight timing differences. To address this limitation, we refine the physical onset times using transcription-based onset activations, using the same transcriber as that used for sequence alignment. The following approach is taken from Saha et al. (2026).

Starting from the sequence-aligned onsets, each onset is snapped to a strong activation within a ±150-ms window using a pitch-wise matching strategy inspired by graph-matching methods (Karp, 1980). If no activation above a threshold of 0.1 is found, the sequence-aligned onset time is retained. Only onsets are adjusted, while durations from the sequence alignment are preserved. The transcription model operates at a 32-ms frame resolution, which is an upper bound on the temporal precision.

3.5.3 Discussion

The two alignment variants reflect complementary perspectives. Sequence-level alignment preserves structural coherence and is less sensitive to local pitch-wise ambiguities. Note-level alignment emphasizes physically realized timing and is therefore closer to the performance domain.

Figure 5 illustrates this principle: sequence-level alignment assigns a single physical onset time to notes that share the same musical onset time, whereas note-level alignment can resolve small onset asynchronies between these notes. However, its quality ultimately depends on the transcription model providing sufficient onset cues for all notes.

We note that neither representation constitutes a perfect ground truth, as achieving such precision would require clean multitrack recordings or dedicated MIDI capture equipment for every instrument (e.g., a MIDI-ready piano). Instead, both should be understood as estimates from score-centric and performance-centric viewpoints, respectively. As we will show in Section 5, both provide a meaningful basis for training and evaluation. The note-level alignment in particular proves effective for training onset-based transcription models.

Aligned annotations are provided in `03_Note Annotations`, organized into Sequence-Alignment and Note-Level-Alignment. For each audio excerpt and alignment type, we provide a note-list `.csv` file and a corresponding MIDI file matching the audio filename. Compared to the unaligned symbolic note lists, onset and offset times now refer to physically aligned times in seconds.

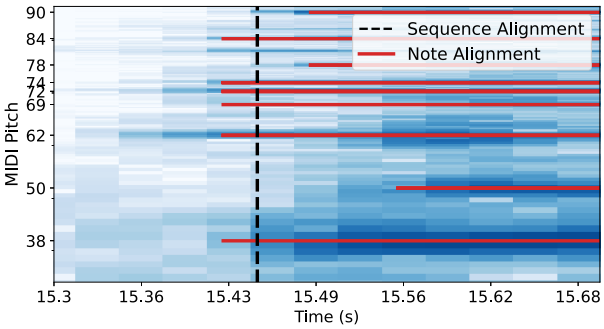

Figure 5 Sequence-level versus note-level alignment for a chord from `BSED-01`. All notes share the same musical onset time in the score. The spectrogram is shown at the transcription model's time resolution (hop size: 32 ms). The dashed black line indicates the common physical onset time assigned by sequence-level alignment, whereas the red boxes mark the individual physical onset times obtained by note-level alignment.

4 BSD: LARGE-SCALE TRAINING SET

While the BSED is carefully curated and verified, it is too small to train neural transcription models. To complement it and enable AMT training on orchestral data, we compile the Beethoven Symphony Dataset (BSD), covering all nine symphonies with 63 h of audio and aligned note-level annotations. Unlike the BSED, the BSD was created largely semi-automatically with less manual verification. The BSD is not intended as a comprehensive musicological resource to analyze Beethoven's symphonies in great detail. Instead, we focus on the aligned note-level annotations of the collected audio to enable transcriber training.

In the following, we describe the BSD and its creation process in greater detail. All data can be found within the folder `BSD`.

4.1 AUDIO DATA

We collected public-domain recordings of the Beethoven Symphonies from IMSLP. Table 2 gives an overview of the symphonies and the number of recordings collected for each movement. In total, we collected 415 recordings of single movements, resulting in 63 h of music. As already discussed previously, the fourth movement of Symphony No. 9 is an outlier in terms of instrumentation, which is why we report it separately in Table 2 and note that it may be advisable to exclude it when training transcribers on the more typical instrumentation of the other symphonies.

Each audio file name follows the same format as in the BSED, but without the `BSED-ID` and the `splitID` in front—for example, `Beethoven_Op021-01_Karajan 1963`. The audio, stored in the folder `02_AudioData`, follows the same format as the BSED and similarly is also available both in its original tuning and in a 'tuning-corrected' version (A4 = 440 Hz).

Similar to the BSED, the audio quality varies significantly. To give users more control over the data that they want to use, we provide quality ratings for each audio file from 4 (highest) to 1 (lowest). Ratings are derived using a semi-automated approach based on the Fréchet audio distance (FAD) (Kilgour et al., 2019). Specifically, we use a set of high-quality orchestral CD recordings as a reference and compute the FAD of each audio file relative to this set. The embeddings required for FAD computation are extracted using the TRILL model (Shor et al., 2020). We define decision thresholds based on the FAD for the quality categories through an informal listening

Opus	Name	# Mvmt. 1	# Mvmt. 2	# Mvmt. 3	# Mvmt. 4	# Mvmt. 5	Total Dur. [hh:mm]
Op. 21	Symphony No. 1 in C major	11	11	11	11	—	04:34
Op. 36	Symphony No. 2 in D major	8	8	8	8	—	04:19
Op. 55	Symphony No. 3 in E ♭ major	19	18	18	17	—	14:33
Op. 60	Symphony No. 4 in B ♭ major	5	5	5	5	—	02:45
Op. 67	Symphony No. 5 in C minor	16	15	15	13	—	07:50
Op. 68	Symphony No. 6 in F major	18	19	16	18	18	12:41
Op. 92	Symphony No. 7 in A major	17	17	16	17	—	10:05
Op. 93	Symphony No. 8 in F major	4	4	4	4	—	01:43
Op. 125	Symphony No. 9 in D minor	4	4	4	— [4]	—	02:59 + [01:33]
	∑	102	101	97	93 + [4]	18	61:28 + [01:33]

Table 2 Overview of the collected audio of the Beethoven Symphonies Dataset for each symphony and movement. The fourth movement of Symphony No. 9 is reported separately in brackets because of its distinct instrumentation. In total, the dataset comprises 411 + [4] movement recordings.

test. However, these ratings provide only a coarse indication of audio quality, without distinguishing between specific degradations (e.g., noise, clipping, compression). All quality ratings and corresponding FAD scores are provided in the file `quality_ratings_BSD.csv`.

4.2 SYMBOLIC DATA

Our annotations for the BSD are derived from the previously mentioned MuseData repository based on the work by Selfridge-Field (1997), which offers digitally encoded versions of all Beethoven symphonies, among others. The MuseData format was developed with a focus on producing human-readable sheet music and accurate notational detail. Unlike the BSED, we do not provide any sheet-music–oriented formats such as Sibelius or MusicXML, nor symbolic score representations in MIDI or note-list format and rather refer to the original MuseData repository. Instead, we only use the MuseData scores as a basis for deriving aligned note-level annotations of the collected recordings, including onset and offset time, pitch, and instrument.

4.3 SCORE–AUDIO ALIGNMENT

To align BSD note annotations with audio, we apply the same two-stage procedure as in BSED (Section 3.5): an initial sequence alignment followed by note-level refinement. For robustness, we manually annotated the start and end of each recording to serve as DTW anchor points.

In contrast to BSED, BSD involves full symphonies, requiring us to ensure structural consistency between score and audio. To this end, we create a set of possible unfoldings of the note annotations for each movement. Each set initially contains only a standard unfolding, where repetitions are expanded according to common conventions in classical scores. Each recording

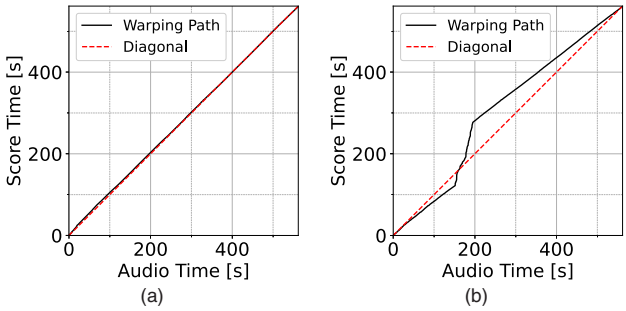

Figure 6 Warping paths between audio and score for two different unfoldings of the same piece: **(a)** Structurally equivalent pair **(b)** Structurally non-equivalent pair.

is then tested against all available unfoldings; if none match—that is, if no structural equivalence is found—a new unfolding is generated manually and added to the set. Since many recordings of the same movement share an identical structure, newly created unfoldings often apply to multiple cases, thereby reducing manual effort.

Structural equivalence is evaluated semi-automatically by comparing the DTW warping path to an ideal diagonal. Paths that remain close are accepted, while larger deviations trigger manual inspection. Such deviations may indicate structural differences but can also arise from local tempo changes or expressive timing, such as fermatas. Figure 6 illustrates this with two alignments of the same recording against structurally different scores.

Although this process ensures structural correctness, we do not manually verify local timing accuracy, and some errors may remain, particularly in lower-quality recordings. As a heuristic, paths close to the diagonal generally indicate good alignment, though this is neither a necessary nor sufficient condition. Stronger deviations can still correspond to valid alignments but require caution. For transparency, we provide the average

deviation from the diagonal (in seconds) for each piece in `warping_path_deviations.csv`.

After sequence alignment, we apply the same note-level refinement as in BSED, but with a wider snapping window (1 second compared to 150 ms for BSED) to accommodate larger initial alignment errors, since we do not set any anchor points for the BSD. For each recording, we provide three structurally matched annotation sets: (1) an unaligned version based on score tempo, (2) a sequence-aligned version via DTW, and (3) a fine-grained note-aligned version. All annotations are available both as note-list `.csv` files and as derived MIDI files. Since our transcriber used for note-level alignment is not trained on singing or unpitched percussion, we cannot generate note-level alignments for the fourth movement of Op. 125. For these recordings, we therefore provide only the initial sequence alignment.

4.4 DATA SPLITS BETWEEN BSD AND BSED

The recordings of which *splitIDs* 1–3 of the BSED were cut from also appear in the BSD. This creates an overlap between the training and test sets when using the full BSD for training and the BSED for testing. To prevent evaluating on data seen during training, we have to exclude some data from the BSD. There are, however, multiple ways to handle this removal, which has been noted before in datasets containing multiple versions of the same piece (Schreiber et al., 2020; Weiß et al., 2021; Zeitler et al., 2024b).

To address this issue, we build on the definitions of Schreiber et al. (2020) and define several splitting strategies that exclude recordings from the BSD in different ways. These strategies differ in how strictly and at what level they prevent overlap. At the movement level, recordings with the same `movementID` as a BSED excerpt are removed to ensure that no movement appears in both training and test sets. At the version level, recordings are excluded to avoid shared acoustic conditions between training and test data. Concretely, we define the following splits (excluding only BSD recordings unless noted otherwise):

- **Movement Split:** Excludes all recordings with the same `movementID` as a BSED excerpt.
- **Version Split—Whole Symphony:** Excludes all recordings with the same `versionID` as a BSED excerpt.
- **Version Split—Single Movement:** Excludes all recordings with both the same `versionID` and `movementID` as a BSED excerpt.
- **Neither Split:** Excludes all recordings with either the same `versionID` or the same `movementID` as a BSED excerpt.
- **Version Split—Reduced BSED:** Does not exclude BSD recordings but restricts evaluation to BSED `splitIDs` 4 and 5 (Blomstedt and Synthetic), which do not overlap with BSD `versionIDs`.

Data Split	# Train Rec.	# Test Excerpts
Movement Split	193 + [4]	100
Version Split—Whole Symphony	176 + [1]	100
Version Split—Single Movement	351 + [4]	100
Neither Split	85	100
Version Split Reduced BSED	411 + [4]	40

Table 3 Overview of different data split strategies for the BSD, when testing on BSED. Each split defines how recordings are excluded, with the corresponding number of training recordings and test excerpts shown. '+ [4]' indicates recordings of Op. 125-04 separately as in Table 2.

Table 3 provides an overview of these approaches along with the resulting number of training movement recordings and test excerpts. The choice of split depends on the evaluation goal: *version-level* tests generalization across acoustic conditions, while *movement-level* targets generalization across musical content. The *neither split* enforces the strictest separation but greatly reduces training data. For scenarios including additional cross-dataset evaluation, we suggest the *Version Split—Single Movement* as a balanced compromise. Note that the *Version Split—Reduced BSED* differs substantially from the full BSED in audio quality and proportion of synthetic audio, making results not directly comparable. To facilitate reproducibility, we provide Python functions in `data_splits.py` that implement these splitting strategies and return the corresponding train and test file lists. Further details can be found in the inline documentation.

4.5 ANNOTATION LIMITATIONS

The parsing process of the MuseData scores used for the training data introduces several limitations related to the quality and completeness of the annotations, which must be considered: (1) Dynamic markings are sometimes inaccurate or inconsistent; therefore, we fill the velocity parameter with a default value of 64 in the note list and MIDI files. (2) Pizzicato indications for the strings are not preserved, and such notes are treated as regular bowed string notes. (3) Grace notes are missing in the annotations. Please note that these limitations only apply to the training data and not to the BSED, since the BSED scores are created manually based on the printed sheet music.

5 INSTRUMENT-AGNOSTIC TRANSCRIPTION

In the following, we demonstrate the utility of our datasets and establish a baseline for instrument-agnostic note-level transcription in orchestral music. Specifically, we first analyze training and evaluation on our datasets with different alignment types and, second, benchmark transcription performance against existing

models and datasets. Although existing transcription models, such as Chang et al. (2024), are trained on diverse datasets and can generalize to orchestral content to some extent, to our knowledge, this is the first study to evaluate note-level transcription on full orchestral mixes.

5.1 MODELS AND EVALUATION

Our transcription models are based on the Onsets & Frames architecture (Hawthorne et al., 2018), which has been shown to handle multi-instrument settings such as chamber music (Maman and Bermano, 2022) but has not yet been applied to full orchestral data. In contrast with Maman and Bermano (2022), we train our models from scratch on a fixed set of labels without any iterative label refinement to demonstrate that our dataset can be used directly for training, without additional alignment adaptation steps. For training on BSD, we use the *Version Split—Single Movement* with a small subset of held-out movements for validation and use pitch shift-augmentation of ±2 semitones to make the transcribers more robust across tonality.

Our evaluation focuses on instrument-agnostic note-level transcription using the `mir_eval` library (Raffel et al. (2014)), computing precision, recall, and F1-score with a 50-ms onset tolerance, without considering note offset or instrument. We evaluate on both BSED and PHENICX. For BSED, we evaluate on the versions pitch-shifted to the reference tuning of A4 = 440 Hz.

Instrument agnostic transcription for orchestral music can constitute an ill-posed problem, as multiple sources may produce overlapping notes of identical pitch (Berendes et al., 2026). We therefore treat the existing—potentially simplified—metric as a baseline in this work, while advocating for the development of more appropriate evaluation metrics in future research.

PHENICX plays an important role in our study: as a non-synthetic multi-track dataset, its annotations were performed on separate monophonic tracks, enabling fairly accurate and reliable note-level annotation. It therefore serves as a valuable cross-dataset evaluation for our transcriütion models. As we show, results on BSED are generally consistent with those obtained on PHENICX.

For PHENICX, we combine all instrument tracks and apply a reverberation of 0.3 s to resemble real recording conditions, enabling assessment on unseen composers and acoustic environments. Since multiple notes of identical pitch may coincide within a single frame in the PHENICX annotations, we retain only one instance per pitch per frame to match the capabilities of our frame-based models.

We additionally evaluate two pre-trained models from the literature: the instrument-aware transcriber from Chang et al. (2024), denoted `YMT3+`,[13] and the instrument-agnostic transcriber from Bittner et al. (2022), denoted `BasicPitch`.[14] Although `YMT3+` is designed for multi-instrument transcription, we evaluate it in an instrument-agnostic manner by discarding the predicted instrument labels. As for the PHENICX labels, overlapping notes with identical pitch within the same time frame are merged into a single note in order to ensure a fair instrument-agnostic evaluation.

5.2 IMPACT OF ALIGNMENT TYPE ON TRAINING AND EVALUATION

To assess annotation quality and systematically compare sequence- and note-level alignments in training and evaluation, we consider three alignment variants on BSD and BSED. As a baseline, we use sequence-aligned annotations derived from signal-processing features via the SyncToolbox (Müller et al., 2021), denoted *SeqSP*. We further employ our transcription-based sequence alignment, *SeqTr*, and, lastly, the refined note-level alignment, *Note*. Separate models are trained on BSD for each label variant and denoted by $\mathcal{T}$ with alignment-specific subscripts (i.e., $\mathcal{T}_{\text{SeqSP}}$, $\mathcal{T}_{\text{SeqTr}}$, $\mathcal{T}_{\text{Note}}$). For evaluation on BSED-Real, we analogously consider all three alignment types, denoted $\mathcal{A}_{\text{SeqSP}}$, $\mathcal{A}_{\text{SeqTr}}$, and $\mathcal{A}_{\text{Note}}$. Additionally, we evaluate on BSED-Synth and PHENICX using 'ground-truth' annotations, $\mathcal{A}_{\text{GT}}$, to assess alignment-independent performance and cross-dataset generalization. However, please note that, although PHENICX onset annotations are mostly correct within 50 ms, they may include slight temporal errors or ambiguous onset positions, e.g., for very soft onsets.

Table 4 reports mean note-level precision, recall, and F1-scores across transcribers and evaluation sets. The $\mathcal{T}_{\text{SeqSP}}$ model performs consistently worse across almost all evaluations except for $\mathcal{A}_{\text{SeqSP}}$. In particular, on BSED-Synth and PHENICX, it has F1-scores of 36.5% and 49.4%, respectively, versus 70.6% and 65.7% with $\mathcal{T}_{\text{Note}}$. This indicates that the $\mathcal{A}_{\text{SeqSP}}$ labels are weak targets for training and, by extension, also problematic for evaluation. In contrast, models trained on transcription-informed sequence alignment ($\mathcal{T}_{\text{SeqTr}}$) and note-level alignment ($\mathcal{T}_{\text{Note}}$) yield substantially higher overall performance. However, the two training strategies exhibit complementary behavior: $\mathcal{T}_{\text{SeqTr}}$ tends to achieve higher precision but suffers from comparatively low recall, whereas $\mathcal{T}_{\text{Note}}$ shows the opposite trend, with greater recall at the cost of reduced precision. While differences between the evaluation sets $\mathcal{A}_{\text{SeqTr}}$ and $\mathcal{A}_{\text{Note}}$ are on the order of only a few percentages, $\mathcal{T}_{\text{Note}}$ yields higher overall scores on PHENICX, reaching 65.7% compared to 60.4% for $\mathcal{T}_{\text{SeqTr}}$.

Overall, both transcription-informed alignment variants yield meaningful training targets, enabling models trained from scratch on BSD to generalize to PHENICX. However, the resulting transcribers exhibit opposite precision–recall trade-offs. Training on the note-aligned labels seems to enable slightly better cross-dataset generalization, as evident in the results on PHENICX. Evaluating on $\mathcal{A}_{\text{Note}}$ appears slightly more informative for note-level transcription, as it better reflects the fine-grained temporal structure the model is expected to capture, which we explore further in the next section.

	BSED-Real									BSED-Synth			PHENICX		
	$\mathcal{A}_{\text{SeqSP}}$			$\mathcal{A}_{\text{SeqTr}}$			$\mathcal{A}_{\text{Note}}$			$\mathcal{A}_{\text{GT}}$			$\mathcal{A}_{\text{GT}}$		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F
$\mathcal{T}_{\text{SeqSP}}$	72.2	34.2	43.8	57.9	39.9	37.7	73.1	36.1	45.9	77.6	26.9	36.5	79.8	36.0	49.4
$\mathcal{T}_{\text{SeqTr}}$	50.0	38.8	43.2	81.7	61.7	69.2	84.9	63.9	71.6	**88.6**	60.2	**70.2**	**79.8**	49.0	60.4
$\mathcal{T}_{\text{Note}}$	36.5	38.2	37.1	66.6	69.2	67.3	70.5	73.3	71.3	73.8	**69.0**	**70.6**	74.1	**59.4**	**65.7**

Table 4 Note-level precision (P), recall (R), and F1-score (F1) (%) (onset only) for different evaluation sets. Rows represent transcribers trained on BSD with different alignment methods, and columns correspond to evaluation sets with the same three alignment methods applied to BSED-Real, as well as BSED-Synth and PHENICX dataset (which do not rely on the alignment strategies). The best scores for BSED-Synth and PHENICX are boldfaced.

5.3 NOTE-LEVEL TRANSCRIPTION BASELINE

In the following, we focus on training and evaluation using our note-level–aligned labels $\mathcal{A}_{\text{Note}}$ to establish a note-level transcription baseline. We assess note-level transcription performance on BSED and PHENICX across multiple transcribers, including the pre-trained models `YMT3+` and `BasicPitch`. We train three models from scratch: $\mathcal{T}_{\text{MN}}$, trained on MusicNet with MusicNetEM labels (Maman and Bermano, 2022); $\mathcal{T}_{\text{BSD}}$, trained on BSD with note-level alignments; and $\mathcal{T}_{\text{BSD+MN}}$, trained on the combined datasets.

Table 5 reports note-level precision, recall, and F1-scores across all transcribers on BSED and PHENICX. `BasicPitch` and `YMT3+` perform substantially below the other models. Training on BSD yields higher performance compared to MusicNetEM across both BSED and PHENICX, with a substantial increase on BSED (from 60.6% to 71.4%, +10.8 percentage points). When the two datasets are combined, the best overall results are achieved, showing consistent gains over MusicNetEM: the F1-score on BSED rises from 60.6% to 70.8%, while, on PHENICX, it improves from 64.9% to 69.0%.

Notably, $\mathcal{T}_{\text{BSD}}$ is as good as $\mathcal{T}_{\text{BSD+MN}}$ on BSED but worse on PHENICX, suggesting slight overfitting when training exclusively on BSD. On PHENICX, $\mathcal{T}_{\text{BSD+MN}}$ achieves the highest scores overall, demonstrating that combining BSD with MusicNet improves cross-dataset generalization.

The performance spread across transcribers is notably larger on BSED than on PHENICX, which may reflect the challenging acoustic conditions of some of the older recordings in BSED, where models trained on clean audio data or non-orchestral data tend to struggle.

Figure 7 shows the mean F1-score of $\mathcal{T}_{\text{BSD+MN}}$ across onset tolerances for all evaluation subsets. For a high onset tolerance, all performance metrics are close together. For stricter onset tolerances, we see that the F1-scores are slightly lower in general, but also the differences between them become larger. This is especially relevant when comparing the two BSED-Real alignments, where we can see that $\mathcal{A}_{\text{Note}}$ yields a circa four percentage points higher F1 score, compared to $\mathcal{T}_{\text{SeqTr}}$.

Test Set	Transcriber	Note (onset only)		
		P	R	F1
BSED	`BasicPitch`	32.3	23.5	24.9
	`YMT3+`	47.4	39.9	42.3
	$\mathcal{T}_{\text{MN}}$	67.6	57.3	60.6
	$\mathcal{T}_{\text{BSD}}$	71.1	**72.5**	**71.1**
	$\mathcal{T}_{\text{BSD+MN}}$	**73.1**	70.4	70.8
PHENICX	`BasicPitch`	54.1	34.7	41.8
	`YMT3+`	70.3	48.3	57.0
	$\mathcal{T}_{\text{MN}}$	77.5	56.3	64.9
	$\mathcal{T}_{\text{BSD}}$	74.1	59.4	65.7
	$\mathcal{T}_{\text{BSD+MN}}$	**78.5**	**62.0**	**69.0**

Table 5 Note-level transcription results for the BSED and PHENICX test sets in % with an onset tolerance of 50 ms. For BSED, we use the $\mathcal{A}_{\text{Note}}$ annotations.

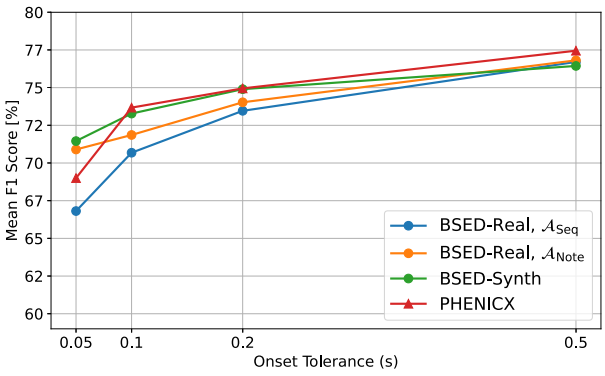

Figure 7 Mean F1-scores for note-level transcription (onset only) of $\mathcal{T}_{\text{BSD+MN}}$ across evaluation onset tolerances for BSED-Synth, BSED-Real with sequence-aligned ($\mathcal{A}_{\text{Seq}}$) and note-level–aligned ($\mathcal{A}_{\text{Note}}$) annotations, and PHENICX.

Overall, our experiments confirm that training on BSD yields strong transcription performance and combining it with MusicNet further improves generalization.

6 CONCLUSIONS

Despite significant progress in AMT, orchestral music remains largely underexplored in transcription research. In this work, we introduced the BSED, a manually curated and verified benchmark for evaluating orchestral AMT. BSED comprises 20 score excerpts from Beethoven's symphonies, each paired with five audio versions, including four real recordings and one synthetic version, totaling approximately 37 min of audio. We provide two complementary annotation variants obtained via score–audio alignment: a DTW-based sequence-level alignment that preserves score structure and a refined note-level alignment based on transcription features that better captures the realized timing of individual note onsets.

As a second contribution, we introduced the BSD, a large-scale training set consisting of 63 h of Beethoven symphony recordings with aligned note annotations obtained using the same methodology. Using BSD, we train instrument-agnostic transcription models and establish baseline results on BSED, achieving an onset-only F1-score of 71%. Together, BSED and BSD provide a unified resource for orchestral AMT, with BSED serving as a high-quality test set including real-world recordings and BSD providing a foundation for training and improving transcription models.

Future work includes extending these resources beyond Beethoven to cover a broader range of composers and orchestral styles and further improving alignment and annotation accuracy in challenging passages and low-quality recordings. A key next step is to move beyond instrument-agnostic transcription toward instrument-aware AMT, leveraging the provided instrument labels and incorporating richer targets such as offsets and expressive performance parameters. Finally, we see a need to refine the evaluation methodology for orchestral AMT by reporting performance across multiple onset tolerances and developing task-dependent protocols that better reflect annotation uncertainty and the intended downstream applications. We invite community feedback and contributions to improve and extend both datasets and tooling.

NOTES

1. https://zenodo.org/records/20344500.
2. https://github.com/groupmm/bsd-code.
3. https://wiki.ccarh.org/wiki/MuseData.
4. https://github.com/MarkGotham/Hauptstimme.
5. Measure 0 denotes the Anacrusis.
6. The conductor of the recording is not known.
7. Excluding measures in Volta Bracket 1.
8. https://www.avid.com/sibelius.
9. https://www.musicxml.com/dolet-plugin/dolet-plugin-for-sibelius/.
10. https://www.noteperformer.com/.
11. https://imslp.org/.
12. https://github.com/breakfastquay/rubberband.
13. We select the best-performing variant with the checkpoint name YPTF.MoE+Multi (noPS).
14. We use the implementation from https://github.com/spotify/basic-pitch, which goes slightly beyond the corresponding research paper.

ACKNOWLEDGMENTS

The International Audio Laboratories Erlangen are a joint institution of the Friedrich–Alexander–Universität Erlangen–Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander Universität (FAU). The hardware is funded by the German Research Foundation (DFG).

We thank Yigitcan Özer for collecting audio data for the BSD and Fatemeh Hosseini for helping with annotation tasks.

FUNDING SOURCES

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 500643750 (MU 2686/15-1).

ETHICAL STATEMENT

The curation of music datasets raises important considerations regarding copyright and intellectual property. For BSD, only historical recordings that are considered to be in the public domain under applicable European Union copyright regulations were included. For BSED, only short audio excerpts are distributed, relying on the quotation and citation provisions for scientific research purposes under European copyright law. The datasets, associated metadata, and annotations are released under a CC-BY-NC-SA 4.0 license. No formal ethics approval or human participant consent was required for this study, as it exclusively involved the collection and annotation of pre-existing musical material and associated metadata.

DATA ACCESSIBILITY

To foster reproducibility and further research, we provide access to BSED and BSD, as well as related resources.

DATASET

Both the BSD and BSED, including audio recordings, score data, annotations and related metadata, are made freely accessible on Zenodo: https://zenodo.org/records/20344500.

CODE REPOSITORY

The accompanying code repository, which can be used to recreate the dataset processing pipelines, can be found on GitHub: https://github.com/groupmm/bsd-code.

DEMOS

Demonstrations showcasing dataset applications can be accessed at https://audiolabs-erlangen.de/resources/MIR/2026-BSED-BSD.

COMPETING INTERESTS

The last author, Meinard Müller, is currently an Editor in Chief for the TISMIR journal.

AUTHORS' CONTRIBUTION STATEMENT

Hans-Ulrich Berendes and Meinard Müller are responsible for the initial conceptualization of the dataset. Data collection was carried out by Hans-Ulrich Berendes and Vlora Arifi-Müller. Data annotation was performed by Hans-Ulrich Berendes, Vlora Arifi-Müller, and Abhirup Saha. Hans-Ulrich Berendes and Abhirup Saha implemented the corresponding codebase. Hans-Ulrich Berendes, Abhirup Saha, Ben Maman, and Meinard Müller are responsible for the experiment design. Hans-Ulrich Berendes wrote the first draft of the paper, and all co-authors of this paper actively participated in the preparation of the final manuscript.

AUTHOR AFFILIATIONS

Hans-Ulrich Berendes
International Audio Laboratories Erlangen, Germany

Abhirup Saha
International Audio Laboratories Erlangen, Germany

Ben Maman
International Audio Laboratories Erlangen, Germany

Vlora Arifi-Müller
International Audio Laboratories Erlangen, Germany

Meinard Müller
International Audio Laboratories Erlangen, Germany

REFERENCES

Abeßer, J., and Schuller, G. (2017). Instrument-centered music transcription of solo bass guitar recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(9), 1741–1750. 10.1109/taslp.2017.2702384.

Balke, S., Berndt, A., and Müller, M. (2025). ChoraleBricks: A modular multitrack dataset for wind music research. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *8*(1), 39–54. 10.5334/tismir.252.

Benetos, E., Dixon, S., Duan, Z., and Ewert, S. (2019). Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, *36*(1), 20–30. 10.1109/msp.2018.2869928.

Berendes, H.-U., Saha, A., Müller, M., and Maman, B. (2026). Unison notes in multi-instrument polyphonic music transcription: Challenges for evaluation. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, Dresden, Germany.

Bittner, R. M., Bosch, J. J., Rubinstein, D., Meseguer-Brocal, G., and Ewert, S. (2022). A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore (pp. 781–785).

Böhm, C., Ackermann, D., and Weinzierl, S. (2021). A multi-channel anechoic orchestra recording of Beethoven's Symphony no. 8 op. 93. *Journal of the Audio Engineering Society*, *68*(12), 977–984. 10.17743/jaes.2020.0056.

Cano, E., FitzGerald, D., Liutkus, A., Plumbley, M. D., and Stöter, F. (2019). Musical source separation: An introduction. *IEEE Signal Processing Magazine*, *36*(1), 31–40. 10.1109/msp.2018.2874719.

Chang, S., Benetos, E., Kirchhoff, H., and Dixon, S. (2024). YourMT3+: Multi-instrument music transcription with enhanced transformer architectures and cross-dataset STEM augmentation. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, London, UK (pp. 1–6).

Dannenberg, R. B., and Raphael, C. (2006). Music score alignment and computer accompaniment. *Communications of the ACM*, *49*(8), 38–43. 10.1145/1145287.1145311.

Del Mar, J. (2020). *Die neun Symphonien*. Bärenreiter.

Del Mar, N. (1983). *Anatomy of the Orchestra*. University of California Press.

Duan, Z., Pardo, B., and Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 2121–2133. 10.1109/tasl.2010.2042119.

Edwards, D., Dixon, S., and Benetos, E. (2023). PiJAMA: Piano jazz with automatic MIDI annotations. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *6*(1), 89–102. 10.5334/tismir.162.

Edwards, D., Dixon, S., Benetos, E., Maezawa, A., and Kusaka, Y. (2024). A data-driven analysis of robust automatic piano transcription. *IEEE Signal Processing Letters*, *31*, 681–685. 10.1109/lsp.2024.3363646.

Emiya, V., Badeau, R., and David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1643–1654. 10.1109/tasl.2009.2038819.

Foscarin, F., McLeod, A., Rigaux, P., Jacquemard, F., and Sakai, M. (2020). ASAP: A dataset of aligned scores and performances for piano transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (pp. 534–541).

Fremerey, C., Müller, M., and Clausen, M. (2010). Handling repeats and jumps in score-performance synchronization. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands (pp. 243–248).

Garcia-Martinez, J., Diaz-Guerra, D., Politis, A., Virtanen, T., Carabias-Orti, J. J., and Vera-Candeas, P. (2025). SynthSOD: Developing an heterogeneous dataset for orchestra music source separation. *IEEE Open Journal of Signal Processing*, *6*, 129–137. 10.1109/ojsp.2025.3528361.

Gardner, J., Simon, I., Manilow, E., Hawthorne, C., and Engel, J. H. (2022). MT3: Multi-task multitrack music transcription. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual.

Good, M. (2001). MusicXML: An internet-friendly format for sheet music. In *Proceedings of the XML Conference and Exposition*, Orlando, FL, USA.

Gotham, M., Hentschel, J., Couturier, L., Dykeaylen, N., Rohrmeier, M., and Giraud, M. (2023). The 'Measure Map': an inter-operable standard for aligning symbolic music. In *Proceedings of the International Conference on Digital Libraries for Musicology (DLfM)*, Milano, Italy (pp. 91–99).

Gotham, M., Song, K., Böhlefeld, N., and Elgammal, A. (2022). Beethoven x: Es könnte sein! (it could be!). In *Proceedings of the Conference on AI Music Creativity*. Online.

Grachten, M., Gasser, M., Arzt, A., and Widmer, G. (2013). Automatic alignment of music performances with structural differences. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil (pp. 607–612).

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J. H., Oore, S., and Eck, D. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (pp. 50–57).

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. A., Dieleman, S., Elsen, E., Engel, J. H., and Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

Huron, D. (2002). Music information processing using the Humdrum toolkit: Concepts, examples, and lessons. *Computer Music Journal*, *26*(2), 11–26. 10.1162/014892602760137158.

Karp, R. M. (1980). An algorithm to solve the m x n assignment problem in expected time O(mn log N). *Networks*, *10*(2), 143–152. 10.1002/net.3230100205.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Graz, Austria (pp. 2350–2354).

Kim, H., Park, J., Kwon, T., Jeong, D., and Nam, J. (2023). A study of audio mixing methods for piano transcription in violin-piano ensembles. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Greece (pp. 1–5). Rhodes Island.

Kong, Q., Li, B., Song, X., Wan, Y., and Wang, Y. (2021). High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Transactions of Audio, Speech, and Language Processing*, *29*, 3707–3717. 10.1109/taslp.2021.3121991.

Krause, M., and Müller, M. (2023). Hierarchical classification for instrument activity detection in orchestral music recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 2567–2578. 10.1109/taslp.2023.3291506.

Lerch, A., Arthur, C., Pati, A., and Gururani, S. (2020). An interdisciplinary review of music performance analysis. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *3*(1), 221–245. 10.5334/tismir.53.

Li, B., Liu, X., Dinesh, K., Duan, Z., and Sharma, G. (2019). Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, *21*(2), 522–535. 10.1109/tmm.2018.2856090.

Maman, B., and Bermano, A. H. (2022). Unaligned supervision for automatic music transcription in the wild. In *Proceedings of the International Conference on Machine Learning (ICML)*, Baltimore, MD, USA (pp. 14918–14934).

Maman, B., Zeitler, J., Müller, M., and Bermano, A. H. (2025). Multi-aspect conditioning for diffusion-based music synthesis: Enhancing realism and acoustic control. *IEEE Transactions on Audio, Speech, and Language Processing*, *33*, 68–81. 10.1109/taslp.2024.3507553.

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

Miron, M., Carabias-Orti, J. J., Bosch, J. J., Gómez, E., and Janer, J. (2016). Score-informed source separation for multichannel orchestral recordings. *Journal of Electrical and Computer Engineering*, *2016*, Article ID 8363507. 10.1155/2016/8363507.

Miron, M., Carabias-Orti, J. J., and Janer, J. (2014). Audio-to-score alignment at the note level for orchestral recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan (pp. 125–130).

Müller, M. (2021). *Fundamentals of Music Processing: Using Python and Jupyter Notebooks* (2nd ed.). Springer Verlag.

Müller, M., Konz, V., Bogler, W., and Arifi-Müller, V. (2011). Saarland music data (SMD). In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, USA.

Müller, M., Özer, Y., Krause, M., Prätzlich, T., and Driedger, J. (2021). Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization. *Journal of Open Source Software (JOSS)*, *6*(64), 3434:1-4. 10.21105/joss.03434.

Müller, M., and Zalkow, F. (2021). libfmp: A Python package for fundamentals of music processing. *Journal of Open Source Software (JOSS)*, *6*(63), 3326:1-5. 10.21105/joss.03326.

Niedermayer, B., and Widmer, G. (2010). A multi-pass algorithm for accurate audio-to-score alignment. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands (pp. 417–422).

Nienhuys, H.-W., and Nieuwenhuizen, J. (2003). Lilypond, a system for automated music engraving. In *Proceedings of the XIV Colloquium on Musical Informatics*, Firenze, Italy (pp. 664–665).

Nieto, O., Mysore, G. J., Wang, C., Smith, J. B. L., Schlüter, J., Grill, T., and McFee, B. (2020). Audio-based music structure analysis: Current trends, open challenges, and applications. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *3*(1), 246–263. 10.5334/tismir.54.

Pätynen, J., Pulkki, V., and Lokki, T. (2008). Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica*, *94*(6), 856–865. 10.3813/aaa.918104.

Peter, S. D., Peter, E. S. D., Cancino-Chacón, C. E., Foscarin, F., McLeod, A. P., Henkel, F., Karystinaios, E., and Widmer, G. (2023). Automatic note-level score-to-performance alignments in the ASAP dataset. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *6*(1), 27–42. 10.5334/tismir.149.

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., and Ellis, D. P. W. (2014). MIR_EVAL: A transparent implementation of common MIR metrics. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan (pp. 367–372).

Riley, X., Edwards, D., and Dixon, S. (2024a). High resolution guitar transcription via domain adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* Seoul, South Korea (pp. 1051–1055). 10.1109/ICASSP48485.2024.10446182.

Riley, X., Guo, Z., Edwards, A. C., and Dixon, S. (2024b). GAPS: A large and diverse classical guitar dataset and benchmark transcription model. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, CA, USA (pp. 611–617).

Saha, A., Berendes, H.-U., Müller, M., and Maman, B. (2026). Snapping matters: Context-aware onset refinement for automatic music transcription. In *Proceedings of the International Computer Music Conference (ICMC)*, Hamburg, Germany (pp. 339–346).

Sapp, C. S. (2005). Online database of scores in the humdrum file format. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, London, UK (pp. 664–665).

Sarkar, S., Benetos, E., and Sandler, M. (2022). EnsembleSet: A new high quality dataset for chamber ensemble separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India (pp. 625–632).

Schreiber, H., Weiß, C., and Müller, M. (2020). Local key estimation in classical music audio recordings: A cross-version study on Schubert's Winterreise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain (pp. 501–505).

Selfridge-Field, E (Ed.). (1997). *Beyond MIDI: The Handbook of Musical Codes*. MIT Press.

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chaumont Quitry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. (2020). Towards learning a universal non-semantic representation of speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Shanghai, China (pp. 140–144).

Solomon, M. (1998). *Beethoven*. Schirmer Trade Books.

Tamer, N. C., Özer, Y., Müller, M., and Serra, X. (2023). High-resolution violin transcription using weak labels. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy (pp. 223–230).

Tamer, N. C., Ramoneda, P., and Serra, X. (2022). Violin etudes: A comprehensive dataset for f0 estimation and performance analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India (pp. 517–524).

Thickstun, J., Harchaoui, Z., and Kakade, S. M. (2017). Learning features of music from scratch. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France.

Weber, P., Uhle, C., Müller, M., and Lang, M. (2025). STAR drums: A dataset for automatic drum transcription. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *8*(1). 10.5334/tismir.244.

Weiß, C., Zalkow, F., Arifi-Müller, V., Müller, M., Koops, H. V., Volk, A., and Grohganz, H. (2021). Schubert Winterreise dataset: A multimodal scenario for music analysis. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, *14*(2), 25:1–18. 10.1145/3429743.

Wu, Y., Wei, W., Li, D., Li, M., Yu, Y., Gao, Y., and Li, W. (2024). Harmonic frequency-separable transformer for instrument-agnostic music transcription. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Canada (pp. 1–6). Niagara Falls.

Xi, Q., Bittner, R. M., Pauwels, J., Ye, X., and Bello, J. P. (2018). GuitarSet: A dataset for guitar transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France (pp. 453–460).

Zehren, M., Alunno, M., and Bientinesi, P. (2024). Analyzing and reducing the synthetic-to-real transfer gap in music information retrieval: the task of automatic drum transcription. *CoRR*, abs/2407.19823.

Zeitler, J., Maman, B., and Müller, M. (2024a). Robust and accurate audio synchronization using raw features from transcription models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, CA, USA (pp. 120–127).

Zeitler, J., Weiß, C., Arifi-Müller, V., and Müller, M. (2024b). BPSD: A coherent multi-version dataset for analyzing the first movements of beethoven's piano sonatas. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *7*(1), 195–212. 10.5334/tismir.196.

Zhang, H., Tang, J., Rafee, S. R. M., Dixon, S., and Fazekas, G. (2022). ATEPP: A dataset of automatically transcribed expressive piano performance. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India (pp. 446–453).

TO CITE THIS ARTICLE:
Berendes, H.-U., Saha, A., Maman, B., Arifi-Müller, V., & Müller, M. (2026). Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription. *Transactions of the International Society for Music Information Retrieval, 9*(1), 405–422. DOI: https://doi.org/10.5334/tismir.343

Submitted: 4 October 2025 **Accepted:** 29 April 2026 **Published:** 24 July 2026