Transactions of the International Society for Music Information Retrieval

Edwards, D., Dixon, S., and Benetos, E. (2023). PiJAMA: Piano Jazz with Automatic MIDI Annotations, *Transactions of the International Society for Music Information Retrieval*, V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

DATASET

PiJAMA: Piano Jazz with Automatic MIDI Annotations

Drew Edwards, Simon Dixon, Emmanouil Benetos

Abstract

Recent advances in automatic piano transcription have enabled large scale analysis of piano music in the symbolic domain. However, the research has largely focused on classical piano music. We present **PiJAMA** (**Pi**ano **J**azz with **A**utomatic **M**IDI **A**nnotations): a dataset of over 200 hours of solo jazz piano performances with automatically transcribed MIDI. In total there are 2,777 unique performances by 120 different pianists across 244 recorded albums. The dataset contains a mixture of studio recordings and live performances. We use automatic audio tagging to identify applause, spoken introductions, and other non-piano audio to facilitate downstream music information retrieval tasks. We explore descriptive statistics of the MIDI data, including pitch histograms and chromaticism. We then demonstrate two experimental benchmarks on the data: performer detection and generative modeling. The dataset, including a link to the associated source code is available at `REDACTED_FOR_ANONYMOUS_REVIEW`.

Keywords: PiJAMA, jazz piano, dataset, piano transcription, generative modeling

1. Introduction

Recent progress in automatic piano transcription has made it possible to conduct large scale analysis of piano music in the symbolic domain. Thus far, the research community has primarily focused on Western classical piano music. Given the status of classical music as the predominant genre of solo piano music, this is understandable. Our goal is to focus on what is arguably the second largest genre of solo piano music: jazz piano.

Jazz is sometimes referred to as America's classical music (see Taylor, 1975). With its roots in the spirituals of early African-Americans, jazz has developed over the past hundred years into a prominent cultural art form. The jazz piano tradition stretches back to the origins of jazz, with early ragtime and boogie woogie styles leading to stride piano. New styles continued to develop, such as swing, bebop, cool, hard bop, fusion, free, and modal. While much of this development took place in ensemble playing, the solo jazz piano tradition developed in parallel, with pianists seeking to incorporate these new styles into their own solo performances.

The depth and importance of jazz piano is in stark contrast to the lack of available datasets for its analysis and modeling. To this end, we present PiJAMA: Piano Jazz with Automatic MIDI Annotations. To our knowledge, this represents the first dataset at this scale solely focused on jazz piano music. By consulting jazz piano pedagogical works, international piano competitions, and jazz critics' publications, we identify 120 pianists to include in the dataset. For each pianist, we scrape as much of their solo piano discography as possible from YouTube. Using state-of-the-art algorithms for automatic piano transcription, MIDI annotations for each performance are computed and released. While there do exist many manual transcriptions of jazz piano performances, these can be difficult to digitize from sheet music. Just as automatic transcriptions is not be perfect, neither is optical music recognition (OMR). Furthermore, the transcriptions will lack the expressive timing of a performance.

In this article, we discuss the current state of piano datasets, describe the data collection methodology, explore the dataset via summary statistics, and demonstrate multiple machine learning experiments on the data. PiJAMA will be useful to researchers in music information research, computational musicology, and music performance studies, and in specific tasks such as automatic music transcription, performer identification, and cover song detection.

2. Related Work

This dataset and the methodology used in this research work are made possible by significant improvements in automatic piano transcription in recent years. In this section, we review related datasets and then discuss the current state-of-the-art in automatic piano transcription.

	MAPS			MAESTRO (v1)		
Model	Frame F1	Onset F1	On+offset F1	Frame F1	Onset F1	On+offset F1
Sigtia et al. (2016)	72.22	46.58	18.38	-	-	-
Hawthorne et al. (2018)	78.30	82.29	50.22	-	-	-
Hawthorne et al. (2019)	**84.91**	**86.44**	**67.43**	**90.15**	95.32	80.50
Kong et al. (2021)	82.78	82.40	56.59	89.71	**96.76**	82.47
Hawthorne et al. (2021)	-	-	-	88.00	95.95	**83.46**

Table 1: Overview of automatic piano transcription techniques and their performance on the datasets MAPS and MAESTRO. For Hawthorne et al. (2019), the MAPS results are from a training configuration with data augmentation, and the MAESTRO results are without augmentation. For Kong et al. (2021), the MAPS results were evaluated with the published checkpoint and the MAESTRO results are the published numbers. Note that this model was trained without data augmentation.

2.1 Datasets

2.1.1 MAPS

MAPS (MIDI Aligned Piano Sounds, Emiya et al., 2010) consists of MIDI files either algorithmically generated (e.g. common chords, random chords) or scraped from the Internet, and then rendered to audio by software instruments and Yamaha Disklavier. The scraped data are mostly classical performances with some "traditional" pieces as well. All MIDI files were taken from scores, with tempo curves manually edited to create a more realistic synthesized performance. In total the dataset has 65 hours of aligned audio and MIDI. Prior to MAESTRO, this was the most important dataset for piano transcription research.

2.1.2 MAESTRO

The MAESTRO dataset (Hawthorne et al., 2019) was created from performances at the International Piano e-Competition and consists of 1276 performances with a total duration of about 200 hours. All audio comes from Yamaha Disklavier pianos and all performances are Western classical. Although the 200 hours are unique performances, it is worth mentioning that there are many duplicate compositions in the dataset. If these duplicate pieces are removed, there are 84 hours of unique compositions.

2.1.3 GiantMIDI

GiantMIDI (Kong et al., 2022) is the most closely related work to our contribution. Similar to our approach, the dataset is compiled by scraping YouTube audio and using an automatic piano transcription algorithm. The metadata is sourced from the International Music Score Library Project (IMSLP), from which over 140,000 compositions are identified. Through a process of scraping and filtering, this set is narrowed down to roughly 10,000 audio files with automatically transcribed MIDI. All tracks in the dataset are classical piano pieces.

2.1.4 ATEPP

Another work taking a similar approach to us is ATEPP (Automatically Transcribed Expressive Piano Performance, Zhang et al., 2022). Whereas GiantMIDI sought to maximize diversity of composition, ATEPP seeks to find many different performances of the same compositions to allow analysis of the expressive differences between the individual pianists and performances. In total, they collect 11,742 performances by 49 pianists, covering 1580 distinct movements. They use the Spotify API to locate performances of pieces by famous Western classical composers and find YouTube URLs for the various performances. Once again, all tracks in the dataset are classical.

2.1.5 RWC Jazz Music Database

The RWC (Real World Computing, Goto et al., 2002) Music Database has a collection of audio and (unaligned) MIDI with a wide variety of instrumentations. Their RWC Jazz Music Database is one of the few datasets we could find with solo jazz piano performances. However, only five pieces are available with a total duration of less than twenty minutes.

2.1.6 Weimar Jazz Database & Dig That Lick

The Weimar Jazz Database (Pfleiderer et al., 2017) is a dataset of 456 transcriptions of monophonic jazz improvisations (about 13.5 hours). These were transcribed with a combination of manual (pitches, onsets, offsets, chords, beats) and automatic (dynamics and intonation) techniques. A subsequent project, *Dig That Lick* (Höger et al., 2019), introduced the DTL1000 database with 1736 (22.2 hours) monophonic jazz solos that were automatically extracted with a CRNN-based algorithm. These datasets represent two of the most significant efforts to provide a computational analysis of real, professional jazz performance. But the focus is entirely on monophonic transcriptions: no solo jazz piano transcriptions exist in the dataset.

2.2 Automatic Piano Transcription

Automatic piano transcription has seen dramatic improvement over the past decade, owing primarily to the use of deep learning algorithms. Due to its acoustic properties and the large amount of available training data, solo piano transcription has witnessed the greatest progress. We briefly review the recent literature leading up to the current state-of-the-art.

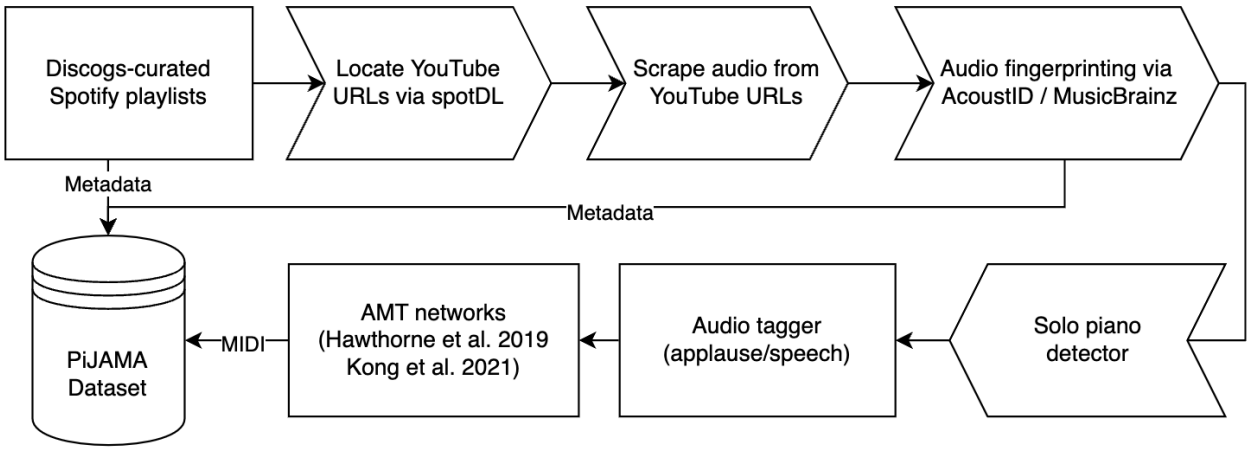

Figure 1: Diagram of the data collection process for the PiJAMA dataset. Stages with a filtering effect are represented with an arrow block symbol.

The first end-to-end neural network for polyphonic music transcription was put forth by Sigtia et al. (2016). Inspired by the success of deep learning in speech recognition, they design a transcription system composed of two parts: an *acoustic model* and a *language model*. They use a convolutional neural network for the acoustic model and a recurrent neural network for the language model. A follow-up result (Kelz et al., 2016) reached a new SOTA result on the MAPS dataset using a simple, single-CNN based model that predicts the pitches present in each frame.

A breakthrough occurred with the Onsets and Frames paper (Hawthorne et al., 2018). They use an acoustic model inspired by Kelz et al. (2016), and also employ a bi-directional LSTM to allow temporal correlations between frames to be learned by the model. However, the key insight of their research is to decompose the training objective into two parts: predicting note onsets and note frames. Then, new notes will only start if the onset detector predicts an onset of that pitch. This factorization of the output space seems to help the network learn better: the onset time series is relatively sparse compared to the frames (since notes tend to be held down for many frames). Their result more than doubles Kelz's previous SOTA F1-score (notes + offsets) on the MAPS dataset. When the Onsets and Frames model is trained on MAESTRO, the accuracy improves dramatically.

Hawthorne et al. (2018) mention that, "the current practice of using a 50ms tolerance for note onset correctness allows for too much timing jitter." Kong et al. (2021) sought to address this limitation in their approach. Their architecture is heavily inspired by Onsets and Frames, but they develop a novel strategy to regress on the precise note onsets and offsets within a single frame. This method is inspired by a paper from computer vision research called "You Only Look Once" (YOLO, Redmon et al., 2016). Here, an object recognition system is trained with a technique of splitting an image into grids. Then at each grid, the network predicts a distance from the coordinate of the grid to that of the object being recognized. Similarly, Kong et al. (2021) assign each frame a continuous value distance to its nearest onset, and the network is trained to predict this distance. Then, using a sliding window, the values from contiguous blocks of frames are geometrically combined to predict a precise onset time within a frame. Their system outperforms Onsets and Frames and achieves very realistic sounding MIDI due to the higher resolution of its predicted output. They expand their approach to predict sustain pedal onsets, which adds to the realism of the transcribed performances.

The most recent development in automatic piano transcription is the sequence-to-sequence Transformer architecture by Hawthorne et al. (2021). The authors criticise existing SOTA methods for piano transcription as being very domain-specific and having complex decoding schemes. Their research investigates using an "off-the-shelf" encoder-decoder Transformer. Its input sequence comprises the single spectrogram frames and its output sequence is a MIDI-like vocabulary, and the network learns to "translate" from the first representation to the latter. Their system achieves SOTA in F1-score incorporating onset time and offset time, but falls short of Kong's model when only measuring onset F1-scores. The authors soon after released an extended model that adds MIDI program number to the output vocabulary and attain SOTA in a variety of datasets for multi-instrument transcription (Gardner et al., 2022).

A comparison of results for transcription performance of the reviewed systems is given in Table 1. The significant progress in this area has made it possible to conduct large scale music analysis in the symbolic domain, which is the approach we follow in this research.

Table 2: Evaluation on transcribed solo jazz piano performances. Due to varying quality in the transcriptions, we report metrics for both 50- and 100-millisecond note onset tolerance. The results on RWC Jazz and Jazz Web show little improvement from the increased tolerance, whereas the metrics on the human labeled evaluation sets show significant improvement, suggesting greater misalignment in these sources.

		Hawthorne et al.		Kong et al.	
Dataset	#	Note F1 (50ms)	Note F1 (100ms)	Note F1 (50ms)	Note F1 (100ms)
RWC Jazz	4	0.932	0.938	0.909	0.91
Jazz Web	5	0.956	0.959	0.926	0.926
Joe Bagg	5	0.876	0.912	0.806	0.858
Daan Schreuder	8	0.889	0.910	0.865	0.881
per recording average	22	0.908	0.925	0.873	0.891

2.3 Evaluation on Solo Jazz Piano

One natural concern is the ability of these systems to generalize to jazz piano data. They were trained on nearly 200 hours of solo piano music and demonstrate strong performance on held-out test data, but it is possible that jazz music contains harmonic, melodic, and rhythmic features which may lead to lower transcription accuracy.

To verify the ability of these systems, we evaluate transcription performance on out-of-distribution solo jazz piano data. We use the following sources of data:

1. *RWC Jazz Music Database*: Four[1] solo piano performances.
2. *Jazz Web*[2]: A website with solo jazz piano tutorials of increasing difficulty. We only selected the advanced pieces for evaluation; five in total.
3. *Joe Bagg*[3] *Sheet Music Transcriptions*: These are sheet music transcriptions with a coarse alignment from `www.soundslice.com`. There are five transcriptions in total, including one from a recording in PiJAMA.
4. *Daan Schreuder*[4] *MIDI Transcriptions*: These are note-for-note MIDI transcriptions by a professional pianist. All eight of the transcriptions are from songs in the PiJAMA dataset.

All data sources required some pre-processing for evaluation. The Jazz Web only required a small time shift for optimal alignment, whereas the other three sources required dynamic time warping to properly align with the audio. The quality of the final transcriptions varied across the set, so we report both 50 and 100 millisecond metrics. The results are shown in Table 2. The results are comparable to the evaluation on MAPS in Table 1, and demonstrate that the models can generalize well to jazz performances. The Hawthorne et al. model was trained with data augmentation, and it is evident that this leads to significant improvement on out-of-distribution transcription accuracy.

3. Methodology

This section describes the methodology used in compiling the data and associated metadata. Data collection consisted of five stages: selecting pianists for inclusion, identifying solo piano albums for each artist, scraping them from YouTube, applying quality filtering, and finally automatically transcribing the performances (see Figure 1).

3.1 Pianist Selection

It is difficult to agree upon a definition of "jazz". The authors have their own listening preferences and an intuitive sense of what qualifies as solo jazz piano, but we took steps to have more objective criteria for inclusion in the dataset. This was mainly accomplished by referring to widely-used textbooks of jazz piano instruction. The first reference used is "The Jazz Piano Book" (Levine, 1989). At the end of the book, the author gives a detailed list of listening recommendations with specific performances of many professional jazz pianists. For every pianist mentioned in this list, we add them to our set of artists. However, this book was published in 1989, and so we also consult "Playing Solo Jazz Piano: A New Approach for Creative Pianists" (Siskind, 2020). In addition, we include past finalists of the Thelonious Monk Institute of Jazz International Piano Competition [5] and the American Pianists Association Jazz Competition [6]. Another authoritative source is the Live at Maybeck Recital Hall series of recorded concert performances. We attempt to include every artist who performs in this collection in our dataset.

3.1.1 Exceptional Inclusions

One issue that emerged from the data collection methodology was a large gender skew. Only 10% of the pianists in the dataset are female, and this is after an intentional effort was made to increase representation. This raises an ethical question of gender representation in jazz piano music which is beyond the scope of this paper to address, but this statistical imbalance is important to consider when using this dataset for downstream tasks.

To increase the number of female pianists, we add performances from the following pianists not found in the aforementioned sources: Eliane Elias, Renee Rosnes, Beegie Adair, and Lynne Arriale. In addition to these inclusions, we add performances from two young

modern prodigies of jazz piano: Joey Alexander and Justin Kauflin.

3.2 Metadata Curation

For each artist, we record the metadata of their name, gender, and year of birth. We then manually search for solo piano albums by each artist. This manual search was performed primarily by querying the Discogs[7] database of artist discographies. Albums were then cross-referenced on Spotify and compiled into playlists. The albums were organized into two playlists: "Live" and "Studio", corresponding to live performances in the former case and recording studio conditions in the latter.

3.3 YouTube Audio Search

Once the playlists were finalized, we used the open-source software spotDL [8] to correlate each Spotify URI with a YouTube URL. Following this, a combination of automatic and manual data validation was conducted. First, all tracks were audio fingerprinted with AcoustID's Chromaprint algorithm and searched against the MusicBrainz database (see Swartz, 2002). Tracks that were matched to the same metadata (artist, track, and album) were moved forward to the next stage of the quality pipeline. Any tracks that were unmatched or mismatched were manually inspected by opening the YouTube URL. False positive matches were corrected when possible. If no YouTube URL could be found manually, the track was excluded from the final dataset. Of the 3,023 songs in the curated playlists, 2792 remained at this stage.

3.4 Quality Filtering

3.4.1 Solo Piano Detection

Next, we used the solo piano detector from Kong et al. (2022) to identify and remove any non-piano performances. The authors use a threshold of 0.5 and report precision, recall, and F1-score of 89.66%, 86.67%, and 88.14%, respectively. Everything with an average score above 0.5 was automatically retained, which accounted for over 95% of the data. The remaining tracks were manually inspected. Ultimately, only 15 tracks were removed and each had an average score of less than 0.12.

3.4.2 Audio Tagging

The subsequent stage of the pipeline was running an audio tagging system on the data. For this, we utilized the Audio Spectrogram Transformer (AST, Gong et al., 2021), a multi-label tagging system trained on YouTube AudioSet. The labels in the AudioSet ontology vary in quality, so we used only three labels: *Music, Applause*, and *Speech*. We evaluate the AST on every track using non-overlapping one-second segments and record the scores for these three classes. For the live performances, this allowed us to remove spoken introductions and interjections, and to filter out applause at the beginning and end of the recorded track. For each track, we compute the start and end times of the performance with the following algorithm:

1. Divide the track into one-second segments
2. Let m, s, a be defined as the AST scores for classes *Music, Speech*, and *Applause*, respectively.
3. Define a segment as "clean" if either:
   (a) $m > \max(s, a)$ and $a < T_a$ and $s < T_s$, or
   (b) $\max(m, s, a) < T_{\text{rest}}$
4. Find the longest contiguous sequence of clean segments, and define the performance start and end times to coincide with this section.

Intuitively, this says that a second of clean piano music should yield AST class labels where music dominates and applause and speech are relatively small. Otherwise, if music is not the highest category, the applause and speech scores must be very low to permit inclusion. This latter criterion lets us cleanly deal with brief moments of rest in the musical performance. The thresholds ($T_a = 0.4$, $T_s = 0.5$, $T_{\text{rest}} = 0.1$) were manually tuned.

To approximate the accuracy of the estimated start and end times, we randomly select 25 studio and 25 live recordings and manually annotate the start and end times. The selected performances were not used in tuning the thresholds. An endpoint was considered correct if it was within 1 second of the annotation. All 50 endpoints of the studio performances are correctly predicted and 47 endpoints of the live recordings are correct, for a total accuracy of 97%. Only one endpoint deviated by more than 2 seconds, due to vocalizations of the pianist (Keith Jarrett) causing an earlier region of audio to be classified as speech.

3.5 Piano Transcription

For piano transcription, we use two state-of-the-art systems to generate MIDI from the audio. First we use Onsets and Frames by Hawthorne et al. (2018). This model has the advantage of being trained with data augmentation, and is more robust to the acoustic variation across PiJAMA. The second model is High-Resolution Piano Transcription with Pedals by Kong et al. (2021). This model has two main advantages: arbitrary time resolution and pedal prediction. Both models are used as-is with no fine-tuning for the jazz genre. With respect to the sustain pedal: both models are trained such that note durations are extended based on the pedal control signal, but the Kong model additionally predicts when a pedal is pressed. For each audio track, both MIDI transcriptions are made available.

3.5.1 Agreement Measure

Having computed the transcription output from both systems, we can measure the similarity of the transcriptions on a per-file basis. The motivation for this analysis is the question of whether high agreement is suggestive of a higher confidence transcription.

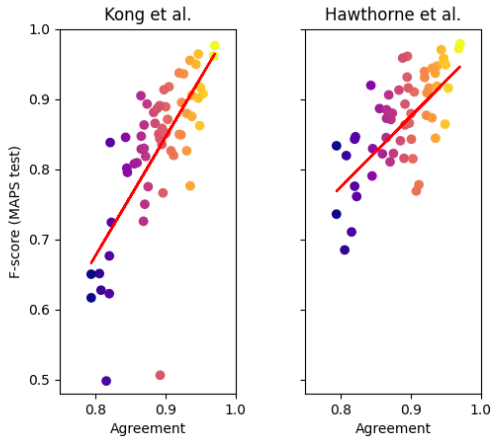

Figure 2: Scatter plots depicting the relationship between transcription agreement and note onset F1 score. Each data point is computed from a performance in the MAPS test set.

First we define our measure of agreement. For a pair of transcriptions (T_1, T_2), we define our agreement score A as

$$A(T_1, T_2) = \frac{F1_{onset}(T_1, T_2) + F1_{onset}(T_2, T_1)}{2} \quad (1)$$

where T_1, T_2 are transcriptions, $F1_{onset}(r, e)$ is the F1-score of note onset transcription accuracy with r as reference and e as the estimated transcription, as computed by `mir_eval`. Figure 2 shows the relationship between agreement scores and F1 scores. The Pearson correlation coeffecient between the agreement score A and the F1 score of each system is 0.72 (Kong et al.) and 0.50 (Hawthorne et al.). The lines of best fit have slopes 1.7 and 1.0, respectively.

We caution that these two systems use the same training data (although only Hawthorne et al. use data augmentation) and have similar architectures, so agreement will in general be high. Furthermore, both systems may share typical error modes (e.g. octave errors), which will increase agreement but hurt accuracy. However, these statistical results suggest that our agreement metric can provide a weak signal of transcription quality and confidence. In the absence of ground truth, as is the case for PiJAMA, this measure may be valuable for curating a subset of transcriptions with a higher expected accuracy. We provide the agreement measure as an additional feature in the metadata. Within the PiJAMA dataset, 1506 tracks (119 hours) have an agreement greater than or equal to 0.90.

4. Descriptive Statistics

The dataset has 120 different artists and 244 albums. During the original sourcing of data, 130 pianists were identified. Those absent in the final dataset either had no solo piano albums, either at all or on Spotify, or their performances were not available on YouTube. The total duration of audio is 223.6 hours and the total duration of performances (as determined by the start/end times computed above) is 219.4 hours. In total, there are 2,777 unique performances. On average, each track is 4 minutes and 50 seconds long and has 2,560 notes. There are 7,108,460 total note events in the transcribed MIDI.

4.1 Pitch Histograms

Figure 3 shows a pitch histogram of all note events in the dataset. Each bar represents a piano key and is colored accordingly. The mode of the distribution is middle C (C_4).

Individual differences between pianists are clearly observable even at this macroscopic level. Two examples of artist-specific pitch histograms are given in Figure 4. The first histogram of pianist Jessica Williams shows a strong preference for white keys. The second histogram is of pianist Erroll Garner. Two things are noticeable from this: (1) a greater preference for black keys and (2) a higher frequency of notes in the upper register. These statistics confirm what many jazz critics observe about Erroll Garner. For instance, in a recorded piano lesson by Dick Hyman[9], he imagines how Erroll Garner might take the Tchaikovsky composition "Song Without Words". Hyman says, "I think the first thing he might have done would be to put it in 4/4 time and then I think the second thing he might have wanted to do would be to play it in a different key, a key with more flats in it because his style kind of demanded that you grab a hold of those black keys."

4.2 Frequent Compositions

The musical tradition of jazz music contains a number of compositions referred to as "standards" that are frequently performed. Our data collection methodology does not capture the composer for every track, but we can approximate the frequency of compositions by grouping by track title. Using an exact string match would be too strict to group, so instead we map each song title to a simpler string. We first note that many songs have a suffix such as " - Live in .*" or "(Live)". So we remove any suffix starting with a dash character surrounded by two spaces, or any parenthetical suffix. Next, we remove any punctuation and whitespace. Finally, we map the string to lowercase. With this derived string, we compute frequencies across the dataset.

In the PiJAMA dataset, 51% of recordings are of compositions that only appear once. If we filter for tracks whose composition appears four or more times by our grouping logic, we get 879 tunes, or 32% of the dataset. Thus, PiJAMA may provide a useful dataset for cover song or jazz standard identification.

4.3 Class Imbalance and PiJAMA-30

It is much more common for jazz pianists to record in the trio format with bass and drum accompaniment,

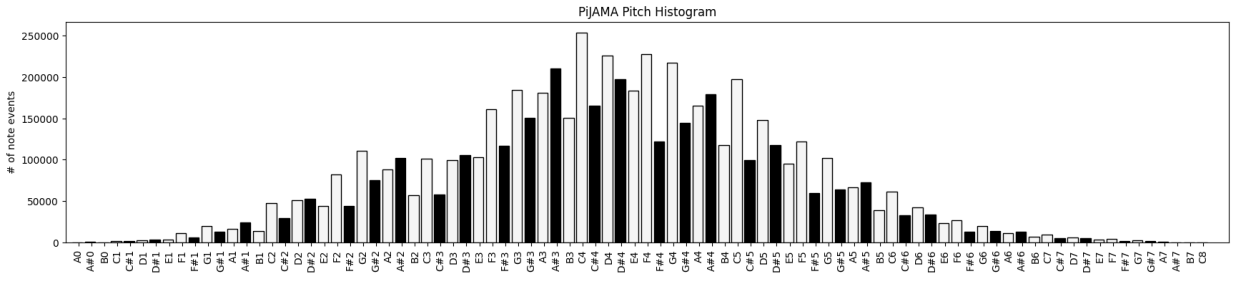

Figure 3: Pitch histogram of all note events in the PiJAMA dataset.

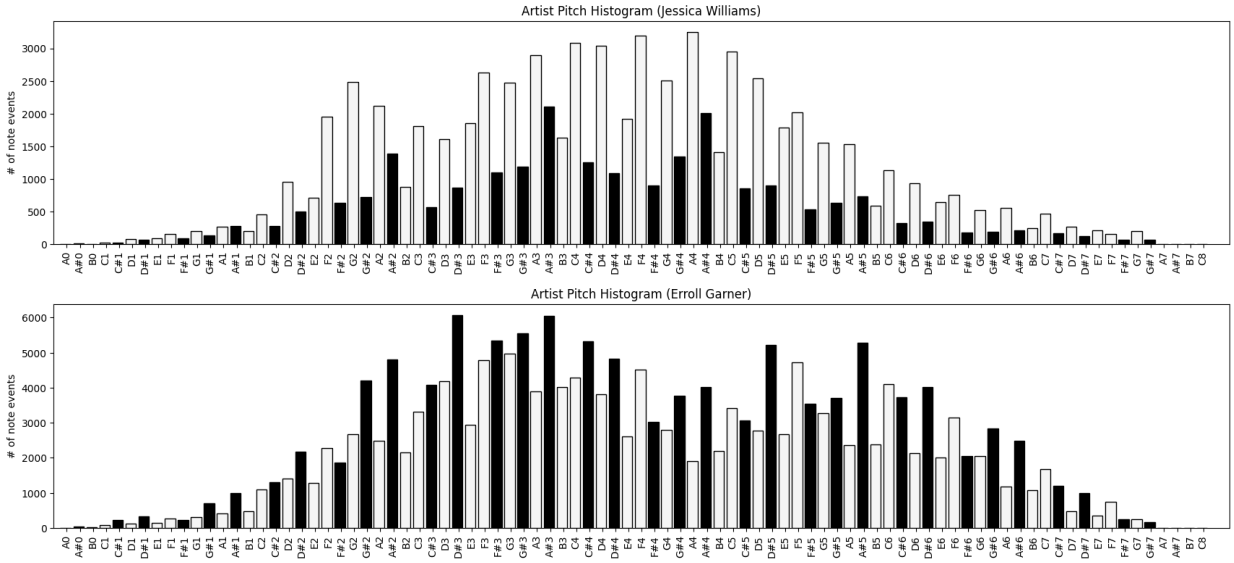

Figure 4: Pitch histograms from pianists Jessica Williams (above) and Erroll Garner (below).

Table 3: Most frequently repeated compositions in the PiJAMA dataset.

Frequency	Composition(s)
17	Body and Soul
13	All The Things You Are, Yesterdays
12	Sophisticated Lady
11	'Round Midnight
10	Blue Monk
9	Alone Together, Prelude to a Kiss, Sweet and Lovely
8	Someday My Prince Will Come, Jitterbug Waltz, Night and Day, My Funny Valentine, Darn That Dream, Someone to Watch Over Me, Don't Blame Me, Blue Bolero, I Should Care, Lush Life, Everything Happens To Me, In a Sentimental Mood, Con Alma

and thus many artists have very few solo piano albums. Some artists have no full length solo piano albums and their samples in the dataset come from solo piano tracks on albums with primarily group playing. On the other hand, there is a small handful of prolific solo jazz piano artists with immense output. As such, the dataset does exhibit a skew relating to the amount of audio per artist (see Figure 5). The four artists with the most audio are Dick Hyman, Brad Mehldau, Art Tatum, and Fred Hersch, with total durations (in hours) of 18.8, 8.9, 8.0, and 7.2, respectively. Meanwhile, some artists have only one track present in the collection. This class imbalance can be problematic for evaluating predictive models on the dataset, so we define a subset of the data with a more even distribution.

The set of artists included are the 30 pianists with the most data, and we refer to this as **PiJAMA-30**. This subset has one final modification of reducing the amount of data from Dick Hyman by excluding his Century of Jazz Piano (over 5 hours in duration). See Figure 6 for the distribution of durations in the PiJAMA-30 subset. This subset is more suitable for tasks such as pianist identification, which is explored in the next section. For other tasks, such as generative modeling, semi-supervised learning (e.g. using the data for training automatic piano transcription), and statistical analysis, the full PiJAMA dataset may be more appropriate.

4.4 Notes per Second

Figure 7 shows the artists in PiJAMA-30 sorted by notes per second (NPS). Note that this metric should not be directly associated with "playing speed" or tempo: fast, sparse sections may have a lower NPS than moderately paced chordal passages. The highest NPS in the dataset is Erroll Garner, a result of his predominantly chordal playing in both hands. Oscar Peterson and Art Tatum are both pianists well known for their dense playing and technical speed, and unsurprisingly they attain the next highest NPS scores.

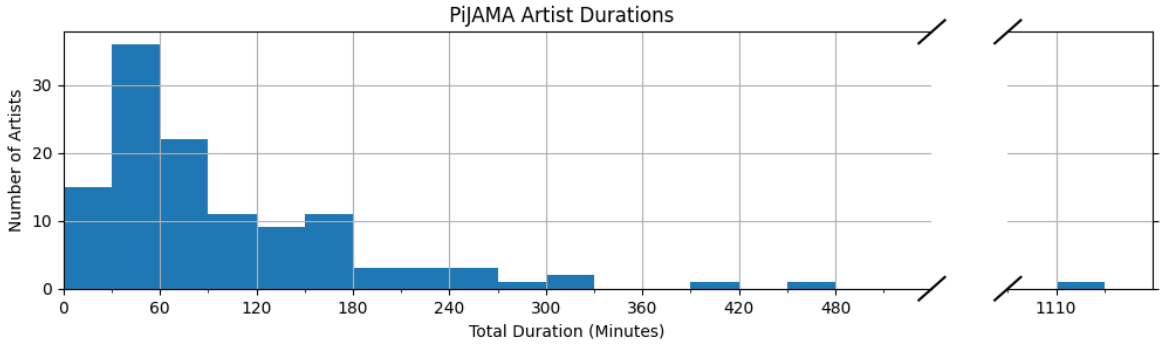

Figure 5: Histogram grouping the number of artists by their duration of performance data, in half-hour increments. One pianist (Dick Hyman) is an outlier with over 18 hours of solo piano recordings.

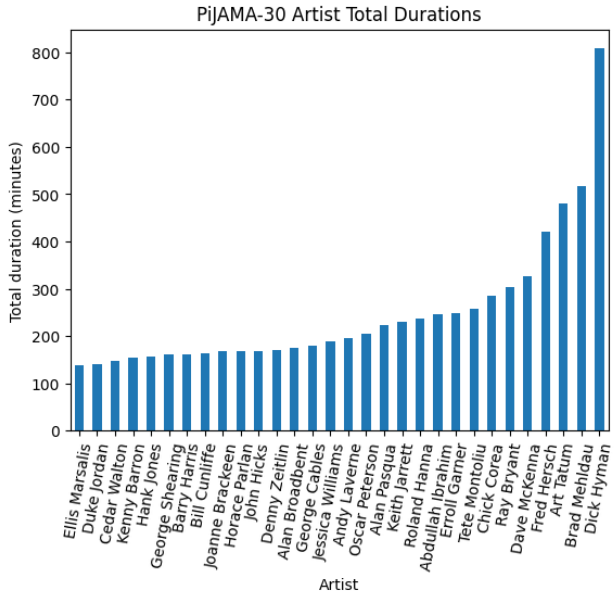

Figure 6: Total performance duration for each artist in the PiJAMA-30 subset.

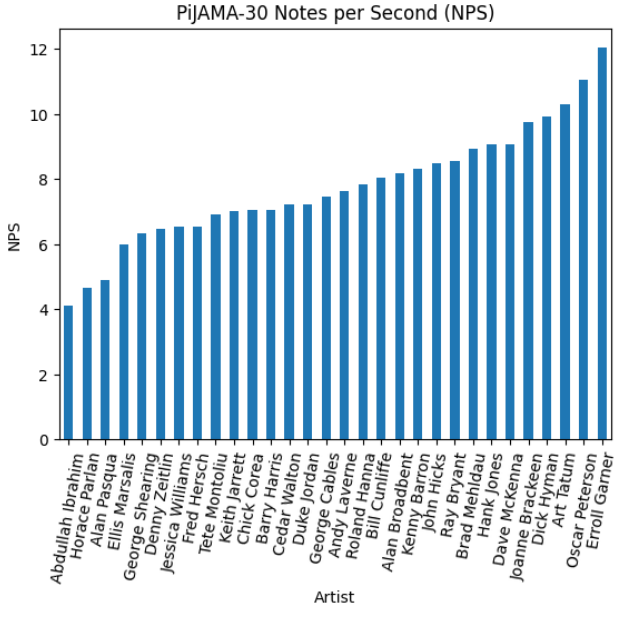

Figure 7: Bar plot of notes-per-second.

4.5 Chromaticism

Our next inquiry regards the degree of *chromaticism* during a performance. According to Forte (1962), "The chromatic expansion of tonality [...] is illustrated [...] by the substitution of a chromatic harmony for an expected diatonic harmony." Furthermore, he writes, "Notes which do not belong to the key [...] are called chromatic notes." Thus, a true measure of chromaticism would require knowledge of the key and chord at any point in a performance in order to be able to identify notes as belonging to the scale or not. We take a simpler approach as a rough approximation of how chromatic a performance is by introducing a measure we call sliding pitch class entropy (**SPCE**).

Consider a sequence of n note events

$$S = ((p_i, t_i) : i \in \{1, ..., n\})$$

where $p_i \in \{21, ..., 108\}$ is the pitch (as MIDI note number) and $t_i \in \mathbb{R}_{\geq 0}$ is the time of the note onset. First, we give the definition of pitch class entropy. It is the Shannon entropy of the normalized pitch class histogram. Formally,

$$\mathbf{1}_c(p) = \begin{cases} 1, & \text{if } p \mod 12 = c \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$f_c = \frac{\sum_{i=1}^{n} \mathbf{1}_c(p_i)}{n} \tag{3}$$

$$\mathbf{PCE}(S) = -\sum_{c=0}^{11} f_c \log(f_c) \tag{4}$$

Thus, $\mathbf{1}_c(p)$ indicates whether p has pitch class c, and f_c is the proportion of notes belonging to pitch class c. Some reference values for **PCE**: the maximum entropy is 2.4849 and drawing notes from a major scale with equal probability would be have entropy 1.9459.

The problem with this measure is that it has no notion of time. A pianist may play in a very diatonic style in a single performance with multiple modulations. Within each modulated section of music, the pitch class entropy could be low, but across the piece it

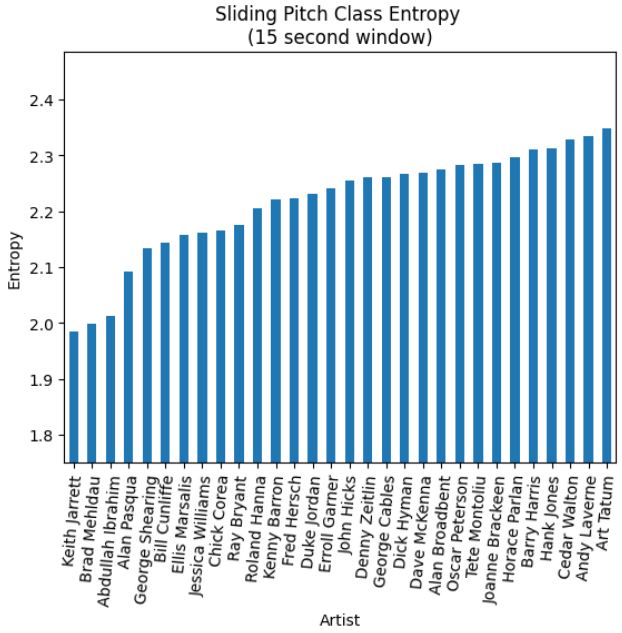

Figure 8: Bar plot of mean sliding pitch class entropy.

would be higher. Thus, **SPCE** is defined for a window size w:

$$\textbf{SPCE}(S) = \frac{1}{Z} \sum_{s=0}^{\lceil \max(t_i) \rceil - w} \textbf{PCE}(p_i : s < t_i < s + w) \qquad (5)$$

where s and w are in units of seconds and Z is the normalizing factor equal to the total number of windows. Note that this formulation corresponds to a hop size of one-second.

Music audio examples of low and high **SPCE** for a window size $w = 15$ can be found in the supplemental materials. In Figure 8 we plot the average **SPCE** for each artist in PiJAMA-30. The lowest scoring artist is Keith Jarrett, which quantitatively supports the analysis by Elsdon (2001), where he frequently comments on the predominantly diatonic style of Jarrett's improvisation:

- "Free improvisers who eschew any kind of reference to convention may consciously avoid diatonic material, something which does not apply to Keith Jarrett." (p. 138-139)
- "Jarrett's gospel style involves the deployment of diatonic progressions in such a way as to evidence a link with this kind of tradition of playing." (p. 159)
- "Chorales are one instance of a style which has increased in importance in Jarrett's improvisations over a number of years [...]. As is typical for a chorale, the harmony here is mainly diatonic." (p. 160-161)
- "The harmonic approach of such [folk] passages is predicated almost entirely on unaltered diatonic triads, moving often in sequential motion." (p. 163)

The highest scoring pianist is Art Tatum. For an excellent blog on Art Tatum's playing style and his extensive use of chromatic passages, see Bayley (2023).

5. Experiments

As a demonstration of the potential research utility of this dataset, we conduct a number of experiments. We consider a supervised learning task to predict the pianist from a short snippet of their playing. Then we briefly look at the task of generative modeling.

5.1 Performer Identification

We explore the task of identifying a pianist based on a short (15 second) snippet of piano performance. For this experiment, we use the PiJAMA-30 subset. All of our experiments use convolutional recurrent neural network (CRNN) architectures, largely inspired by the work of Kong et al. (2021). In all cases, training is performed with gradient descent using the Adam optimizer and loss is computed as negative log-likelihood.

5.1.1 Spectrogram CRNN

The first class of model operates on mel spectrograms from a 16 kHz resampling of the scraped audio. We use 229 mel bins, a hop size of 160, and frequency range of 30 Hz to 8 kHz. The network has five convolutional blocks followed by a two-layer Gated Recurrent Unit (GRU). Each convolutional block has two convolutional layers, each with a 3-by-3 kernel. Each convolutional layer is followed by a rectified linear unit (ReLU) and batch normalization. Following each convolutional block, we apply 2-by-2 average-pooling. The number of filters used in each block is $[64, 128, 256, 512, 1024]$. The final state of the GRU is used as the embedding to pass into a fully-connected layer with output dimension equal to the number of artists to predict (30). The total number of parameters is 27.1M.

5.1.2 Transcription Features CRNN

The second class of model uses a piano transcription model as a frozen backbone for providing features to a learnable CRNN head. The model used is also a CRNN, specifically the High-Resolution Piano Transcription model by Kong et al. (2021). We take the pretrained network and feed in the mel spectrogram as above. Kong's transcription network has four similar acoustic models that predict at each time frame and for each piano note: onsets, offsets, frame activity, and velocity. These four feature maps are layered to create a four channel "image" of size $(4, 88, T)$. From this feature representation, we train a CRNN. Two convolutional layers are followed by a max pooling layer, followed by one last convolutional layer with average pooling. The result is processed by a two-layer bidirectional GRU with hidden dimension of 256, yielding a 1024 dimensional feature vector. This is fed into a fully connected layer to predict the artist. The total number

Model Condition	Split	Test Accuracy	Album Effect
Spectrogram CRNN	Track	0.914	0.647
	Album	0.267	
Spectrogram CRNN (Data Augmentation)	Track	0.782	0.383
	Album	0.399	
Transcription Feature CRNN	Track	0.632	0.176
	Album	0.457	
Transcription Feature CRNN (Data Augmentation)	Track	0.629	0.085
	Album	0.545	
Piano Roll CRNN	Track	0.556	0.055
	Album	0.502	

Table 4: Accuracy of artist prediction models. Two test scores are presented for each model condition: the accuracy on the track-split (all tracks of the dataset shuffled into an 80-10-10 split) and the average accuracy across three album-splits (one random album held out for each artist, yielding roughly an 80-10-10 split). The Album Effect column is the difference between accuracies on the track-split and average album-split.

of learnable parameters is 27.6M.

5.1.3 Piano Roll CRNN

The last model operates on piano rolls using MIDI data from the PiJAMA dataset. Here we use MIDI from the Onsets and Frames model. Each piano roll is a one channel image, $(1, 88, T)$, where the channel dimension is the velocity of the note onset.

5.1.4 Overfitting and "The Album Effect"

Early experiments for the spectrogram model condition were suspiciously accurate on test data. Further analysis revealed the neural network was learning to recognize the acoustic properties of the piano and recording conditions, and *not* the qualities of the pianist. Previous research has termed this phenomenon the "album effect" (see Flexer and Schnitzer, 2010; Rodríguez-Algarra et al., 2019). The original train-validation-test split was created by shuffling all tracks in the dataset, such that every sample in the test set is from a track not contained in the training set. However, tracks from the same album did appear across splits. New splits were created to prevent this. For each artist, we randomly select one album to hold out as test data, while still ensuring an approximate 80-10-10 split. We perform this random process three times. We call the new train-test splits *album splits* and the original split the *track split*.

It is worth mentioning that the use of album splits has its own complications. In some cases, multiple album releases correspond to the same recording conditions. For instance, the artist in the dataset with the second most albums is Art Tatum. However, according to jazz critic Scott Yanow on AllMusic.com, all the albums were captured in very similar conditions: "During four marathon recording sessions in 1953-55, Norman Granz recorded Art Tatum playing 119 standards, enough music for a dozen LPs. The results have been recently reissued separately on eight CDs [...]". Furthermore, artists may perform in a meaningfully different style between albums, which could make the album split inherently more difficult even with identical acoustic conditions.

5.1.5 Results

Table 4 contains the experimental results. Each model condition is trained four times: once on the track split and once on each of the three album splits. Test accuracy is computed across all 15-second segments in the test set. A true positive is only achieved when the highest predicted pianist matches the ground truth. For the album splits, the test accuracy is averaged across the three sets. In the rightmost column we report the delta between the accuracy on the track split and the average accuracy on the album splits.

When the Spectrogram CRNN approach is trained on the track split, it has a test accuracy of 0.914. However, when trained and evaluated on the album splits, its average test accuracy falls to 0.267. For the Spectrogram and Transcription Feature models, we show how data augmentation can reduce the acoustic overfitting. The transformations applied include pitch shifting, gain adjustment, high- and low-pass filtering, and adding coloured noise.[10] In both cases, the album effect is reduced. The Piano Roll CRNN suffers the least from the album effect, as expected.

Our main takeaways from the experiments are as follows:

- The Spectrogram CRNN memorizes the acoustic condition of each recording session, and uses that to predict the artist with a very high degree of accuracy. Informally, it simply learns how each piano sounds.
- The piano rolls remove almost all acoustic information from the signal, and thus the album effect is minimized.
- Some acoustic information is leaking into the Transcription Feature network. This is not surprising, since no thresholding is performed on the activations from the transcription model.
- Data augmentation can reduce but not eliminate

reliance on non-musical features in the predictive task.

We invite future researchers to explore ways of learning distinguishing musical features from raw audio without overfitting to acoustic conditions, and we hope the PiJAMA dataset will be a useful benchmark for progress.

5.2 Generative Modeling

As a final demonstration with the dataset, we conduct a brief generative modeling experiment. Our experiments leverage the Music Transformer of Huang et al. (2019), working with the `Score2Perf` implementation, a codebase by Google Magenta based on the `Tensor2Tensor` project.

First we train a system from scratch. Initially we use the default settings in the `Score2Perf` project: 6 hidden layers, 8 attention heads, an initial learning rate of 0.2 with a linear warmup for 8000 steps followed by exponential decay. Training runs for 1M steps. The generated samples attain some superficial similarity to the ground truth data, but have sporadic and disconnected quality. In quantitative terms, the best negative log perplexity score reached is -1.953.

In a follow-up blog post to the Music Transformer publication (Simon et al., 2019), the authors provide a pretrained model that has trained on over 10,000 hours of automatically transcribed piano performances. The architecture is deepened to 16 hidden layers. We finetune the model on a subset of the PiJAMA dataset for 400K steps with a variety of learning rates $(0.001, 0.005, 0.01, 0.05)$. Quantitative performance was similar across runs, with an optimal negative log perplexity score of -1.884 computed on a held-out test set. For comparison, we also train this architecture from scratch for 1M steps, and the quantitative results are not significantly different from the first experiment of training from scratch. The optimal negative log perplexity score is -1.936.

Rendered audio for many samples is included in our supplemental materials. The finetuned model produced the most compelling musical samples. Subjectively, we found the performances more coherent and realistic than the results of training from scratch, while still sounding much more like the PiJAMA data compared to samples taken from the pre-trained model prior to the fine-tuning. However, the samples lack the basic structure of jazz performances. Most solo jazz piano performances contain a clear statement of the melody followed by improvisations based upon a repeated harmonic structure. Nearly all of the samples sound more like random sections of improvisation. There is often local harmonic stability and familiar melodic language, but on the whole they do not convey a pianist performing a tune. Informally, it sounds a bit like a professional pianist "noodling" or "riffing" at the keyboard. Nevertheless, we are encouraged and excited by the potential for future research in modeling jazz piano performances in the symbolic domain.

6. Conclusion

We present PiJAMA, a curated dataset of over 200 hours of jazz piano performances with automatically transcribed MIDI. The data collection methodology relies on jazz piano pedagogy and uses open-source software to find YouTube performances. Modern techniques for filtering and audio tagging are employed to enhance the collection's metadata. Multiple experiments are conducted to demonstrate the utility of the dataset in the context of music information research. For future directions, we encourage using PiJAMA to pursue tasks such as cover song detection, melody extraction, unsupervised pre-training, and music language modeling.

From a critical perspective, there are some limitations of our approach. The need to manually compile artists and albums is a bottleneck to greater scalability. Whereas classical music has sources like IMSLP, jazz music lacks such a well-curated catalogue. Methods for automatic data collection, such as genre detection ("solo jazz piano"), could possibly permit a larger scale data effort. Another thing lacking from our dataset is annotations beyond MIDI. Having labels for beats, bars, chords, phrases, sections, and other structural aspects would enrich the dataset (see Eremenko et al., 2018; Balke et al., 2022). From a more musical perspective, our methodology focuses entirely on solo jazz piano. However, there exists much more jazz piano playing in the ensemble format, and it is likely the case that piano performances from group playing have been hugely influential on the development of jazz piano, including solo playing.

We have released the source code to generate the full PiJAMA dataset and provide the MIDI and metadata for direct download at `REDACTED_FOR_ANONYMOUS_REVIEW`.

Notes

[1] The dataset states there are five solo piano performances, but one of the performances has two piano parts (one accompaniment and one lead) with significant intersection of pitches. This was excluded from the evaluation.

[2] `http://mir.audiolabs.uni-erlangen.de/jazz-piano/`

[3] `https://www.patreon.com/Joe_Bagg`

[4] `https://daanschreuder.gumroad.com/`

[5] `https://hancockinstitute.org/competition/`

[6] `https://www.americanpianists.org/jazz`

[7] `https://www.discogs.com/`, a crowdsourced online database and marketplace of music releases.

[8] `https://github.com/spotDL/spotify-downloader`, by Kah, J., Kot, J., and Malho-

tra, R.

[9] Dick Hyman is an educator, writer, and accomplished pianist. He is well represented in the PiJAMA dataset. The piano lesson is available at https://www.youtube.com/watch?v=R_B8nHqsGsI

[10] Pitch shifting is not used for the Transcription Feature condition.

Acknowledgements

We would like to acknowledge the open source software projects that made this research possible, namely spotDL and MusicBrainz Picard. We also want to acknowledge the musicians in the dataset whose artistic output makes this research possible. ACKNOWLEDGEMENT_REDACTED_FOR_ANONYMOUS_REVIEW.

References

Balke, S., Reck, J., Weiß, C., Abeßer, J., and Müller, M. (2022). JSD: A dataset for structure analysis in jazz music. *Transactions of the International Society for Music Information Retrieval*, 5(1):156–172.

Bayley, L. R. (2023). Granz' Art Tatum recordings: A piano method for jazz. https://artmusiclounge.wordpress.com/2023/01/26/granz-art-tatum-recordings-a-piano-method-for-jazz/.

Elsdon, P. S. (2001). *Keith Jarrett's solo concerts and the aesthetics of free improvisation 1960-1973*. PhD thesis, University of Southampton.

Emiya, V., Bertin, N., David, B., and Badeau, R. (2010). MAPS - A piano database for multipitch estimation and automatic transcription of music. Research report inria-00544155, Telecom ParisTech.

Eremenko, V., Demirel, E., Bozkurt, B., and Serra, X. (2018). Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*.

Flexer, A. and Schnitzer, D. (2010). Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music Journal*, 34(3):20–28.

Forte, A. (1962). *Tonal Harmony in Concept and Practice*. Holt, Rinehart and Winston.

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

Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. In *Proceedings of Interspeech 2021*, pages 571–575.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2002). RWC Music Database: Popular, Classical and Jazz Music Databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, Paris, France. ISMIR.

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., and Eck, D. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*.

Hawthorne, C., Simon, I., Swavely, R., Manilow, E., and Engel, J. (2021). Sequence-to-sequence piano transcription with transformers. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J., and Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. (2019). Music transformer. In *International Conference on Learning Representations*.

Höger, F., Frieler, K., Pfleiderer, M., and Dixon, S. (2019). Dig that lick: Exploring melodic patterns in jazz improvisation. In *20th International Society for Music Information Retrieval Conference: Late Breaking Demo*.

Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., and Widmer, G. (2016). On the potential of simple framewise approaches to piano transcription. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 475–481.

Kong, Q., Li, B., Chen, J., and Wang, Y. (2022). GiantMIDI-piano: A large-scale MIDI dataset for classical piano music. *Transactions of the International Society for Music Information Retrieval*, 5(1):87–98.

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 on Audio, Speech and Language Processing*, 29:3707–3717.

Levine, M. (1989). *The Jazz Piano Book*. Sher Music Co.

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.-G., and Burkhart, B., editors (2017). *Inside the Jazzomat - New Perspectives for Jazz Research*. Schott Campus.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. *CoRR*. http://arxiv.org/abs/1506.02640.

Rodríguez-Algarra, F., Sturm, B. L., and Dixon, S. (2019). Characterising confounding effects in music classification experiments through interventions. *Transactions of the International Society for Music Information Retrieval*, 2(1):52–66.

Sigtia, S., Benetos, E., and Dixon, S. (2016). An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(5):927–939.

Simon, I., Huang, C.-Z. A., Engel, J., Hawthorne, C., and Dinculescu, M. (2019). Generating piano music with transformer. `https://magenta.tensorflow.org/piano-transformer`.

Siskind, J. (2020). *Playing Solo Jazz Piano: A New Approach for Creative Pianists*. Jeremy Siskind Music Publishing.

Swartz, A. (2002). MusicBrainz: A semantic web service. *IEEE Intelligent Systems*, 17(1):76–77.

Taylor, B. (1975). *The History and Development of Jazz Piano: A New Perspective for Educators*. University of Massachusetts Amherst.

Zhang, H., Tang, J., Rafee, S., Dixon, S., and Fazekas, G. (2022). ATEPP: A dataset of automatically transcribed expressive piano performance. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, pages 446–453.