PDMX: A Large-Scale Public Domain MusicXML Dataset for Symbolic Music Processing

Phillip Long, Zachary Novack, Taylor Berg-Kirkpatrick, Julian McAuley

Background and Motivations

Data-Driven Artificial Intelligence

AI models increasingly require more training data, and AI-Music models are no exception.

Gemini

DALL·E

AI-Music Datasets

Current AI-Music datasets are usually either small, lack quality metrics, or in MIDI format.

Dataset	Format	Hours	Size
Lakh MIDI	MIDI	>9,000	174,533
SymphonyNet	MIDI	>3,200	46,359
MAESTRO	MIDI	201.21	1,282
Wikifonia Lead Sheet Dataset	MusicXML	198.40	6,405
POP909	MIDI	60.0	909
EMOPIA	MIDI	11.0	1,078
MMD	MIDI	-	1,524,557
MetaMidi	MIDI	-	612,088

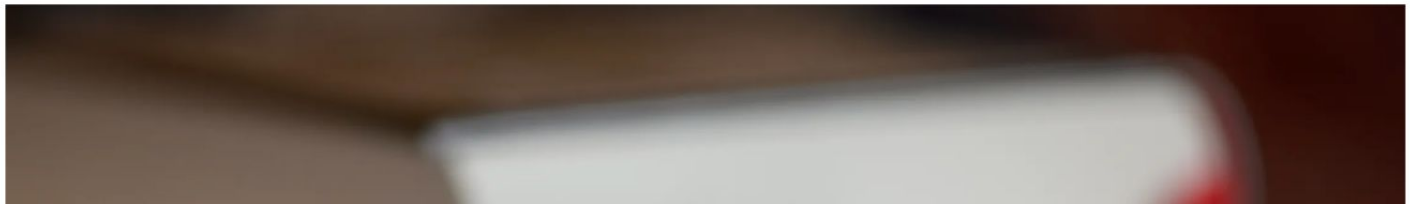
Additionally, these datasets often contain copyrighted data, an issue magnified by recent lawsuits against leading AI-Music companies Suno and Udio just this past year.

Music Industry Groups Sue AI Companies for Stealing Artists' Work to Generate Music

Two lawsuits claim that AI companies Suno and Udio infringed copyright by training AI on the likes of Drake, Bruce Springsteen, and Green Day

By Nina Corcoran and Jazz Monroe

June 25, 2024



To alleviate these issues, we present…

Public Domain Music XML

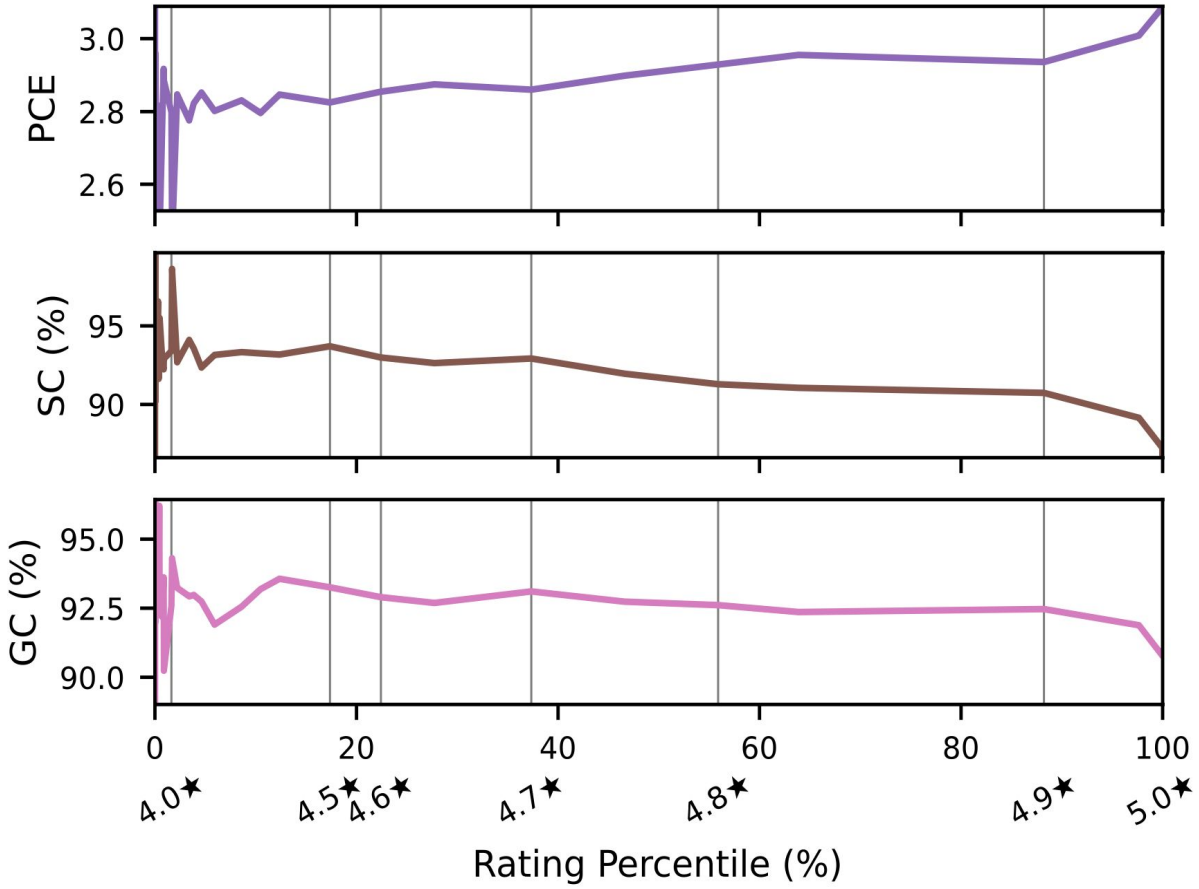
Basic Statistics

- >250K songs, comprising over 6,250 hours
- All songs in the public domain
- Genre, tag, description, and popularity metadata for every song
- Scraped from MuseScore, an online sheet-music sharing forum
- >12M performance directive tokens, >10M lyric tokens
- Largest publicly-available, copyright-free MusicXML dataset (CC-BY License)!

Data Quality

- The MuseScore metadata provides a crowd-sourced “rating” attribute, which represents the five-star rating of a song, averaged over many MuseScore users.
- Rather than use some proxy for data quality, we use these ratings instead!
- ~14k (6%) public domain songs have ratings.

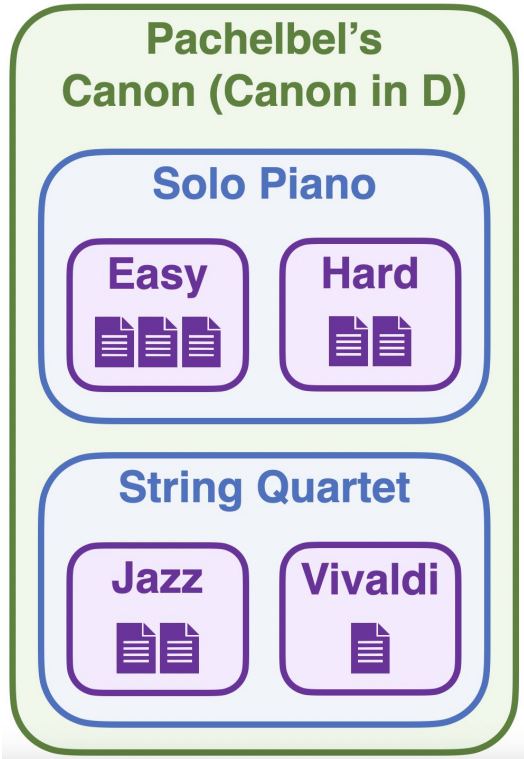
We report the musical statistics of Pitch Class Entropy, Scale Consistency, and Groove Consistency, broken down by rating percentile. Higher-rated songs tend to be more harmonically diverse, while rating seems to have little effect on rhythm.



Deduplication

Heavy duplication within a dataset can cause bias towards particular high frequency data points and degrade downstream modeling tasks. Direct string matching alone is insufficient for PDMX, since:

- the same song may have different names (e.g. “Pachelbel’s Canon” and “Canon in D”), and
- a song may have multiple different instrumentations and arrangements, which are all valuable for modeling symbolic music.



To remove duplicates:

- We use Sentence-BERT (Reimers and Gurevych 2019) to generate vector embeddings for song “descriptors” (a combination of their title, composer, and subtitle, if available).
    - Using cosine similarity, songs more than 80% similar are grouped together.
- Within each song-title grouping, we next group songs together by instrumentation (e.g. solo piano vs. string quartet).
- Within each song-title and instrumentation grouping, we finally group together by arrangement (e.g. easy vs. hard). We define similar arrangements as those whose total note count are less than 5% different.
- Within each song-title, instrumentation, and arrangement grouping, we select the song with the highest rating as the “best unique arrangement”, or the most notes if no ratings are available.

~152k duplicates removed (60% of all songs), and ~102k songs remain

Modeling

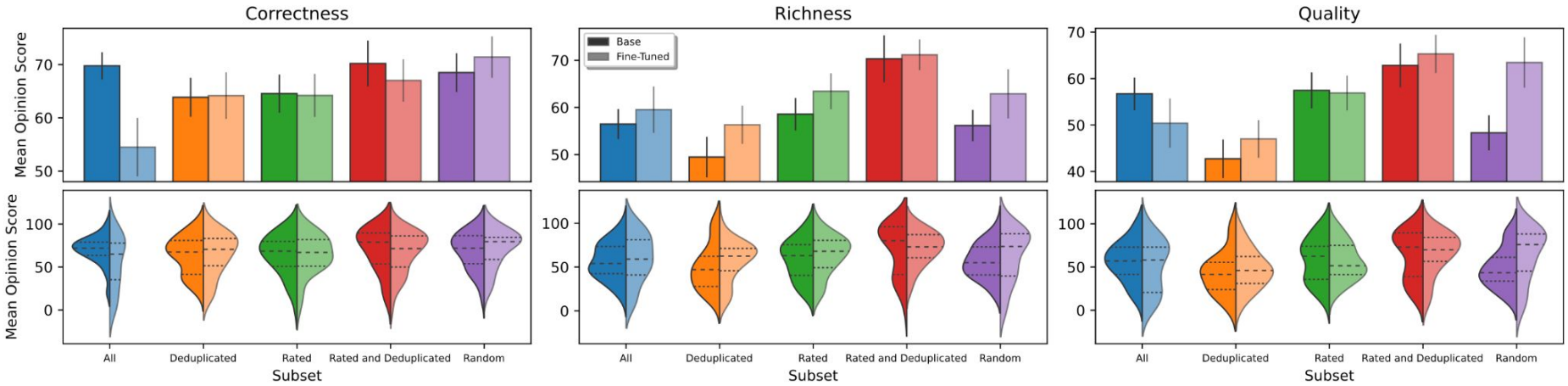
- Do data quality filtering and deduplication affect later symbolic music modeling?
- Does fine-tuning on a small but high-quality subset meaningfully change behavior?

We train autoregressive decoder-only transformers to generate symbolic music on five subsets of our data: all songs (All), unique arrangements (Deduplicated), rated songs (Rated), unique-rated arrangements (Rated and Deduplicated), and a random subset at the size of the Rated and Deduplicated subset (Random). Additionally, we fine-tune each model on the top 50% of rated songs.

Dataset Subsets

Subset	Hours	Size
All	6,250	254,077
Deduplicated	3,756	102,635
Rated	1,001	14,182
Rated and Deduplicated	941	13,187
Random	941	13,187

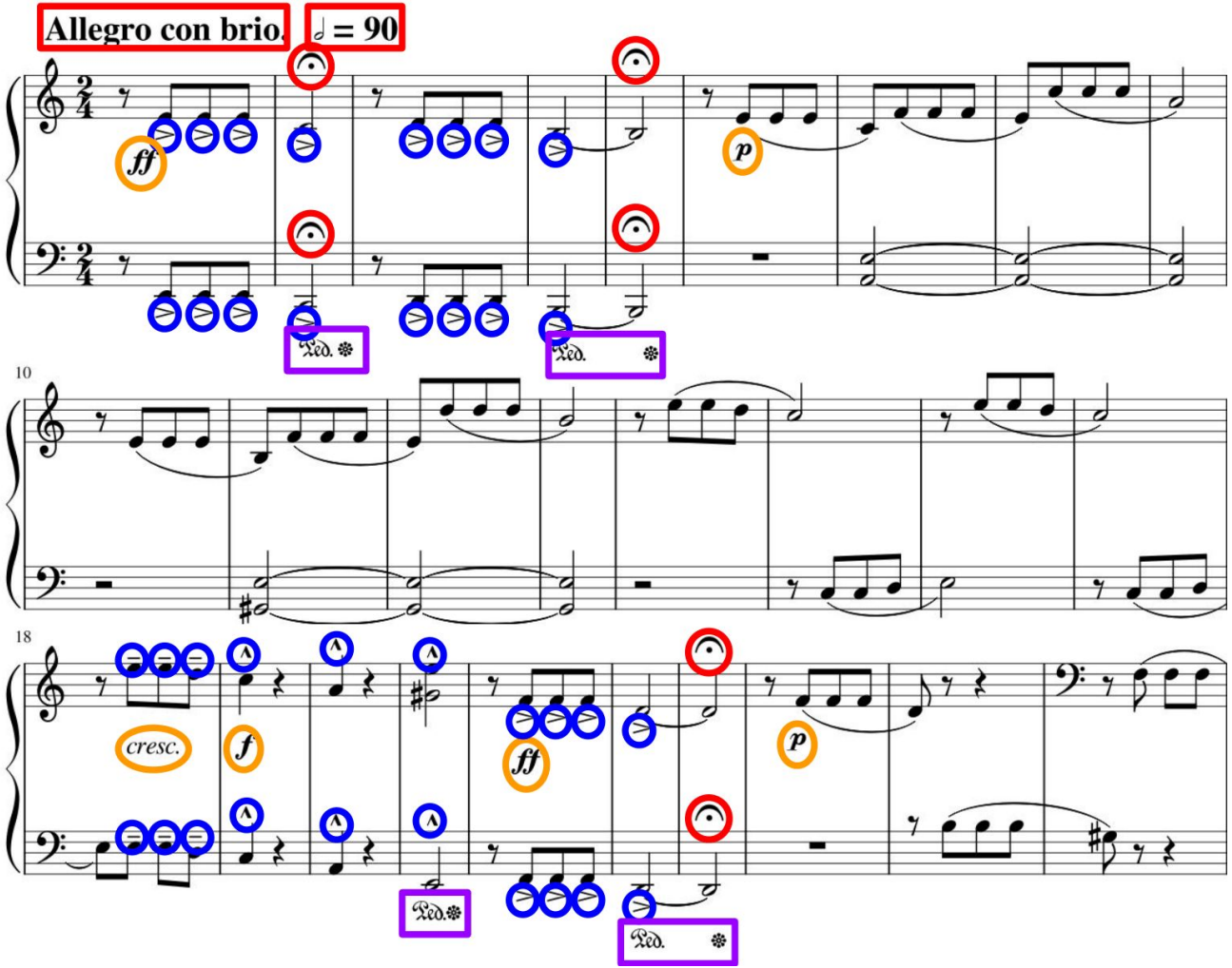
Generated Samples



After conducting a listening test on the samples generated by our models, we found that…

- data quality filtering lead to gains in richness and quality in the base models.
- fine-tuning on a high-quality subset increased richness in all models and improved quality in three.

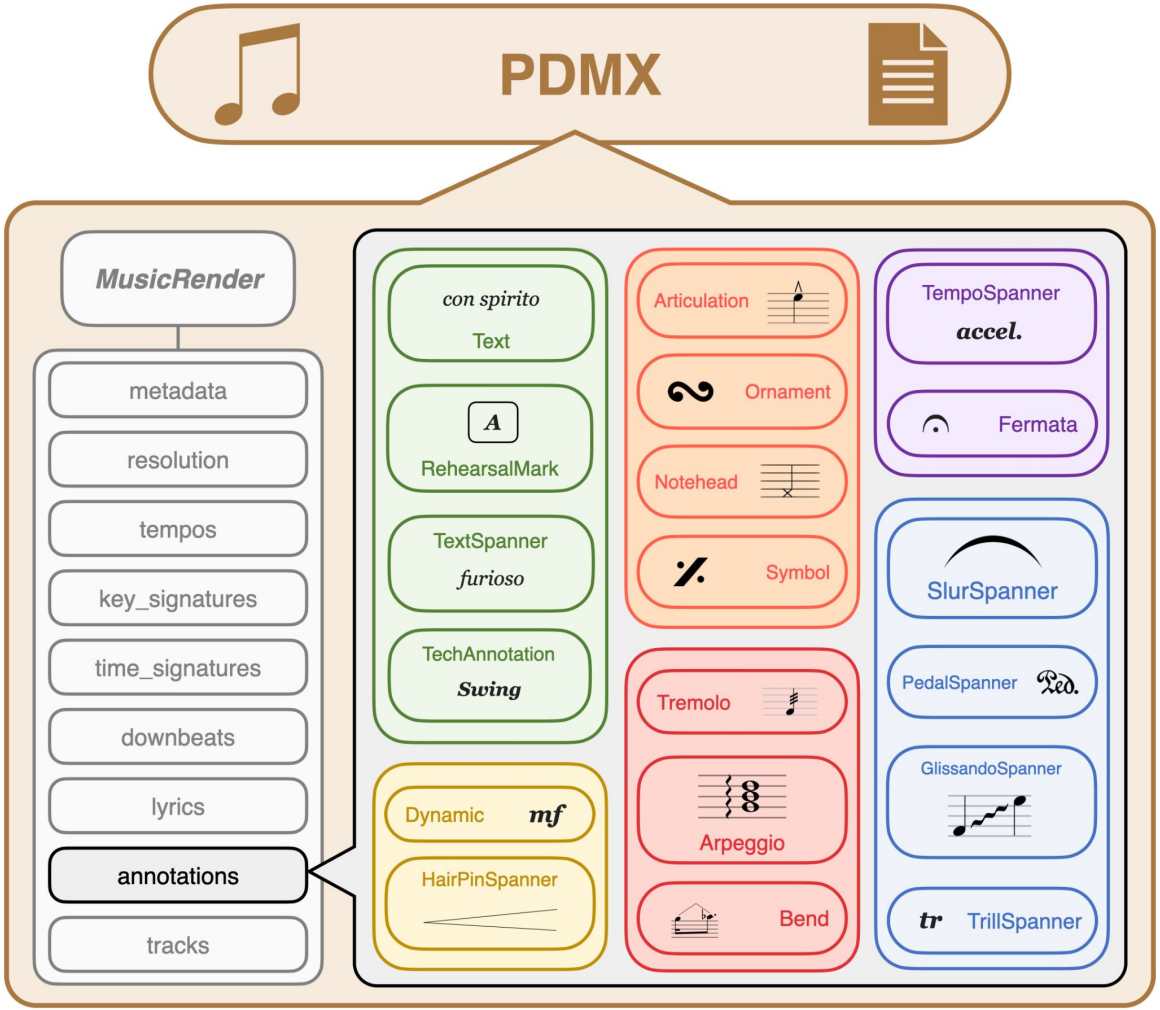
Additional Features



Unlike MIDI, MusicXML contains a plethora of extra non-note information like tempo markings and crescendos, which we call performance directives. While some performance directives are available in MIDI (such as non-text tempo markings), other directives are either implied (such as dynamics through velocity) or not present entirely.

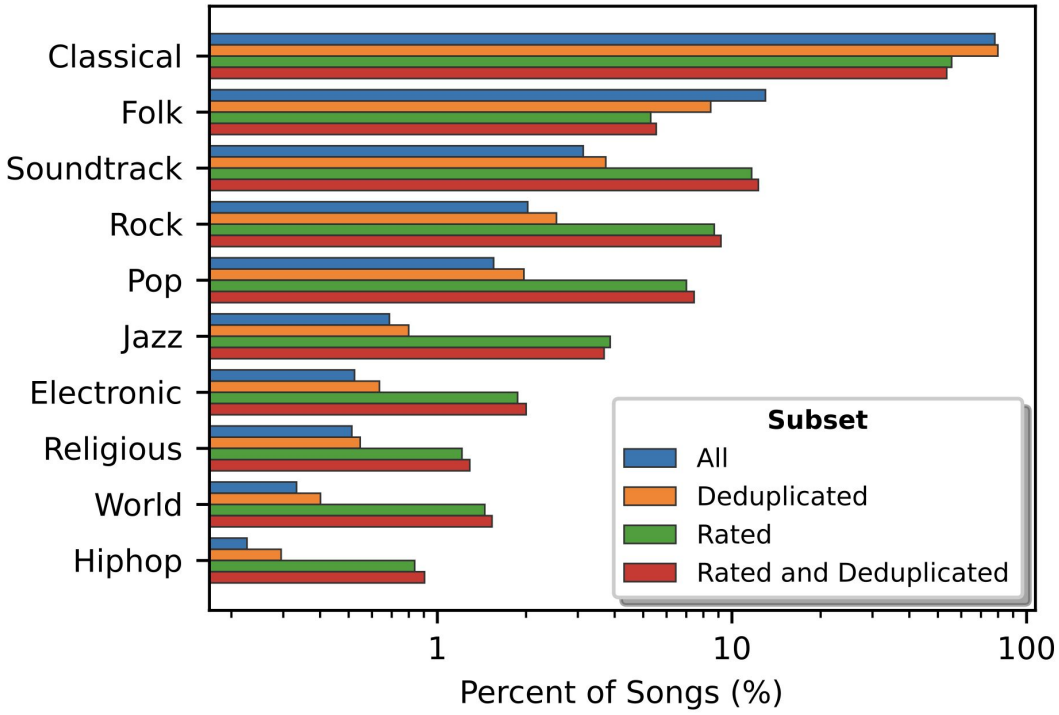
MusicRender

We present ***MusicRender***, an extension of MusPy's *Music* object (Dong et al 2020), that stores the performance directives available in MusicXML data inside a universal Python object. Each type of performance directive is represented with a different Python class. *MusicRender* supports directive-aware audio (WAV) and symbolic (MIDI, MusicXML) output.

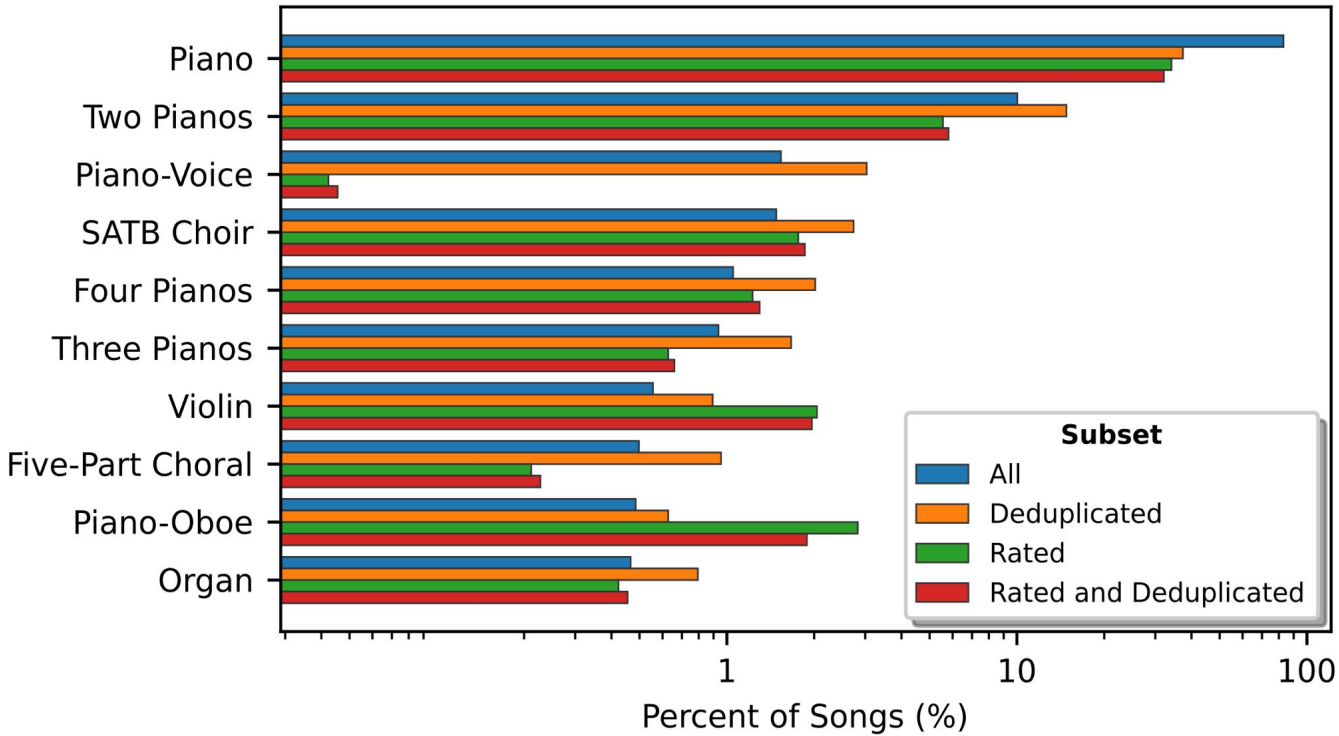


Genres

PDMX encompasses 20 different genres, the most common being classical and folk music. More modern genres, like hip-hop and electronic music, are comparatively much fewer in number, likely due to limited public domain content for more recent works. We find that the rated subsets of PDMX contain a significantly longer “tail” of genres than the full dataset, denoting a large amount of unrated classical music present in PDMX.



Instrumentations



While a significant portion (>70.1%) of PDMX is solo piano music, the dataset also contains nearly 76K songs (29.9%) arranged for other instrumentations, which is in and of itself larger than some existing multitrack datasets.

Public Domain Music XML