CHORDONOMICON: A Dataset of 666,000 Songs and their Chord Progressions

Spyridon Kantarelis[*], **Konstantinos Thomas**[†], **Vassilis Lyberatos**[†], **Edmund Dervakos**[†], **Giorgos Stamou**

Artificial Intelligence and Learning Systems Laboratory, National Technical University of Athens, Greece

Abstract

Chord progressions encapsulate important information about music, pertaining to its structure and conveyed emotions. They serve as the backbone of musical composition, and in many cases, they are the sole information required for a musician to play along and follow the music. Despite their importance, chord progressions as a data domain remain underexplored. There is a lack of large-scale datasets suitable for deep learning applications, and limited research exploring chord progressions as an input modality. In this work, we present *Chordonomicon*, a dataset of over 666,000 songs and their chord progressions, annotated with structural parts, genre, and release date - created by scraping various sources of user-generated progressions and associated metadata. We demonstrate the practical utility of the Chordonomicon dataset for classification and generation tasks, and discuss its potential to provide valuable insights to the research community. Chord progressions are unique in their ability to be represented in multiple formats (e.g. text, graph) and the wealth of information chords convey in given contexts, such as their harmonic function . These characteristics make the Chordonomicon an ideal testbed for exploring advanced machine learning techniques, including transformers, graph machine learning, and hybrid systems that combine knowledge representation and machine learning.

Introduction

Large, publicly available, well structured datasets have been crucial for technological progress in recent decades. They allow for wide-spread experimentation and serve as a common basis for comparison, reproducibility, and evaluation. In the domain of music, the development of such datasets has been extremely slow compared to other domains, hindered by issues such as copyrights, unavailability of audio, distribution shift, and bias. Datasets with significant influence on Music Information Retrieval (MIR) studies, featuring contemporary music data instead of public domain content, have addressed copyright concerns by offering signal-processing attributes rather than copyrighted audio (Bertin-Mahieux et al. 2011) or opting for royalty-free music suitable for commercial purposes (Bogdanov et al. 2019b). Our proposed Chordonomicon dataset utilizes abundant, non-

copyrightable (Booth 2016) user-generated chord progressions, which are robust to distribution shift as chords and their usage have remained consistent over centuries.

Chord recognition is one of the most widely explored topics involving chord progressions within the computer science community (Takuya 1999; Boulanger-Lewandowski, Bengio, and Vincent 2013; Park et al. 2019). Researchers have approached this task through a variety of methodologies, ranging from knowledge-driven, to data-driven deep learning approaches in more recent years. Despite decades of active research in this area, open problems and questions remain (Pauwels et al. 2019). We anticipate that our Chordonomicon dataset, containing more than twenty times the amount of chord progression than the largest dataset to date (de Berardinis et al. 2023), and incorporating structural annotations absent from prior works, can facilitate further advancements in this domain.

Chord progressions offer valuable opportunities for computer science research due to the availability of formal representations of music theory knowledge in the form of knowledge graphs, ontologies, and linked data (Fazekas et al. 2010). These knowledge-based resources have emerged as a promising complement to machine learning systems, improving performance and model explainability across various domains (Futia and Vetrò 2020; Ji et al. 2021; Tiddi and Schlobach 2022; Lyberatos et al. 2023a). The standardized Harte syntax (Harte et al. 2005) used to represent chords in our dataset enables easy linking with these knowledge graphs, allowing for the enrichment of the data and the exploration of hybrid systems that combine machine learning and knowledge-based approaches. Furthermore, chord progressions can be represented as graphs (Louboutin and Bimbot 2016), enabling the comparison of graph-based methods with sequence-based techniques, such as the evaluation of graph machine learning and transformer models for different tasks. We believe the wealth of representations, structure, and semantics available for chord progressions provides a unique opportunity for impactful AI research.

Our contribution can be summarized as follows: firstly, we introduce the largest chords progression dataset to date. Secondly, through the implementation of semantic labeling techniques, we enhance the depth of understanding of musical structures. Thirdly, by providing the dataset in graph format, we facilitate versatile analytical approaches. Lastly,

---

[*]corresponding author, mailto:spyroskanta@ails.ece.ntua.gr

[†]These authors contributed equally.

our establishment of interlinkability with domain ontologies ensures integration with existing knowledge frameworks.

Related Work

Musical compositions can be encoded and expressed in diverse formats, with the most prevalent being audio, MIDI (Moog 1986), and symbolic forms like MusicXML (Good 2001), tabs and text. The associated metadata, furnishing additional details about the musical pieces within a dataset, varies and encompasses aspects such as music annotations (Speck et al. 2011), genre (Bogdanov et al. 2019a), mood (Bogdanov et al. 2019c), artists (Oramas et al. 2015), emotions (Lyberatos et al. 2023b), instruments (Gong et al. 2022), and other musical attributes.

Numerous music datasets have been introduced incorporating a mix of various formats and metadata. They are utilized to address diverse tasks, including but not limited to genre classification (Dervakos, Kotsani, and Stamou 2021), music generation (Wang et al. 2020) and analysis (De Haas et al. 2011), chord recognition (Nadar, Abeßer, and Grollmisch 2019), key estimation (George, Mary, and George 2022), and more [1]. In our work, we present a very large scale dataset featuring the symbolic representation of more than 666,000 contemporary music compositions through the use of music chords and chord progressions. We offer metadata for details such as genre, sub-genre, and release date. Additionally, we include structural information related to different parts of the music piece and Spotify IDs for tracks and artists.

More specifically, in relation to Music Structural Analysis (MSA), there is a scarcity of datasets containing semantic labels (Wang, Hung, and Smith 2022) necessary for accurately defining distinct music components such as *Verse* and *Chorus*. It is notable that efforts in this field (Paulus and Klapuri 2009; Smith et al. 2011; Nieto et al. 2019) primarily emerged more than a decade ago, and the quantity of semantically labeled music tracks remains notably limited (around 900-1500). In our research, we provide 2,670,457 structural labels for 397,580 music tracks, adhering to the terminology proposed in these earlier studies.

Within the domain of graph classification, the existing landscape is often characterized by a dearth of extensive and diverse datasets, particularly in comparison to the well-established datasets in MIR tasks. While widely recognized for their utility, prevalent graph-level datasets, such as those found in the TU collection (Morris et al. 2020), exhibit limitations attributable to their relatively modest sizes, typically consisting of fewer than 1,000 graphs. Although datasets from the Open Graph Benchmark (Hu et al. 2020) provide a sizable resource, they fall short in terms of relevance to the MIR domain. In response to this identified gap, we introduce a novel dataset comprising a substantial 666,000 graphs, poised to make a noteworthy contribution by providing a more expansive and diversified resource.

Another significant aspect of music datasets pertains to the structure of music compositions and their associated metadata. The presence of considerable ambiguity (Weiß, Schreiber, and Müller 2020) and diverse representations of identical musical elements, such as chords (Koops et al. 2019; Hentschel et al.), necessitates curation for the dataset to achieve semantic integration. For example, the ChoCo dataset (de Berardinis et al. 2023) semantically integrates harmonic data from 18 different sources using heterogeneous representations and formats, while Midi2Vec (Lisena, Meroño-Peñuela, and Troncy 2022) offers a framework for converting MIDI files to sequences of embeddings. Consequently, several ontologies have been introduced, providing not only a vocabulary for defining concepts of music aspects and annotation (Jones et al. 2017; Cherfi et al. 2017) but also axioms that can be applied to enhance the dataset by incorporating notions from music theory and harmony (El Achkar and Atéchian 2020).

More notably the Music Ontology (Raimond et al. 2007) proposes a formal framework for dealing with music-related information on the Semantic Web, including editorial, cultural and acoustic information. The Music Theory Ontology (MTO) (Rashid, De Roure, and McGuinness 2018) and the Functional Harmony Ontology (FHO) (Kantarelis et al. 2023) introduce concepts from the western music theory and harmony respectively, while the Chord Ontology (Sutton et al. 2007) provides a vocabulary for describing chords. Our methodology aligns with this objective as it involves representing chords in a manner that ensures interlinkability of our dataset with the Chord Ontology and the FHO.

Dataset

This section delves into insights related to data collection and processing. Specifically, we provide detailed descriptions of the methodology used for data collection and curation, while at the same time explaining the process of data selection and exclusion. Table 1 summarizes the dataset's key statistics, offering a quick overview as we get ready to explore the upcoming discussions. On average, each progression consists of 76.48 chords, with a median of 71 chords and a standard deviation of 54.59.

Data Gathering and Curation

We leveraged web scraping techniques to extract song chord data from the Ultimate Guitar (UG) platform. By adhering to UG's robots.txt, Terms of Service, and established web accessibility protocols (see Ethical Statement), we obtained the site's XML sitemap and filtered the URLs containing the "*-chords-*" substring, which denotes song chord pages. This yielded a dataset of 1.4 million song chord pages, encompassing multiple interpretations (versions) per song. We then scraped the raw content, user ratings, vote counts, artist IDs, song IDs, artist names, and song titles from each of these URLs. Utilizing the song and artist IDs, along with fuzzy string matching on the song and artist names, we consolidated this data into 793,362 unique songs from 93,872 unique artists. As many popular songs had multiple chord versions, we retained only the "best" version per song. The "best" version was determined by a weighted average of the version's user rating and its percentage of the total vote count for that song.

---

[1]https://www.music-ir.org/mirex/wiki/MIREX_HOME

The next step was to format the raw HTML files into the full chord progression of each song, collapsing repeating identical chords into a single chord ('A G G A' became 'A G A'), removing songs that had less than four chords in total and removing any songs that had odd chord encoding (e.g. Cyrillic chord characters) and used UTF-8 encoding to make sure all chords are same format. Finally, this remaining dataset was curated by music experts to retain the chord progressions where they were confident about the terminology of the chords (removing songs that contained bizarre or musically incoherent chord symbolisms), which gave us the ability to accurately convert them into the Harte syntax and the syntax proposed by the FHO. The selection and conversion were performed manually and were supervised by music experts: we compiled a document comprising all encountered chords, which was then reviewed by music experts who aligned them with the Harte syntax only when they were confident about the accuracy of the chord (e.g., "SOL5" to "G:(*3)", "hmin" to "B:min", etc.). Eventually, we conclude into the final dataset of 679,807 unique, chordified, musical tracks.

By utilizing both the Harte syntax and the syntax suggested by the FHO, our dataset can be integrated with the FHO and Chord Ontology, allowing it to be further enriched with concepts from music harmony and theory. Specifically, the chords in our dataset can serve as entities within the FHO and Chord Ontology framework.

Regarding the segments of the tracks, we ultimately employed eight distinct part categories: *Intro, Verse, Chorus, Bridge, Interlude, Solo, Instrumental*, and *Outro*, following an extensive analysis of our data, ensuring alignment with terminology used in prior works (see section). We manually associated alternative and misspelled part names with the aforementioned categories (e.g., mapping "Refrain" to "Chorus"). If any part of a track couldn't be classified, all parts of that track were excluded, resulting in the retention of only the chord progression. Their distribution is illustrated in Figure 2.

Spotify Web API[2] was utilized for metadata collection. Specifically, we gathered information about the release date of the tracks and the musical genres associated with their respective artists. To verify their correctness, we made certain that the names of both the music track and the artist in our data match the corresponding information in Spotify data. We also provide the Spotify IDs of tracks and artists where available.

Alongside the meticulously curated dataset, we offer three Python scripts: one for transposing chords into all tonalities (for data augmentation purposes), another for converting chords into their corresponding notes (e.g., A:7 → ['la','do#','mi','sol']), and a third script that generates a binary 12-semitone list representation for each chord, commencing with the note C (e.g., C:maj7 → [1,0,0,0,1,0,0,1,0,0,0,1])[3],[4].
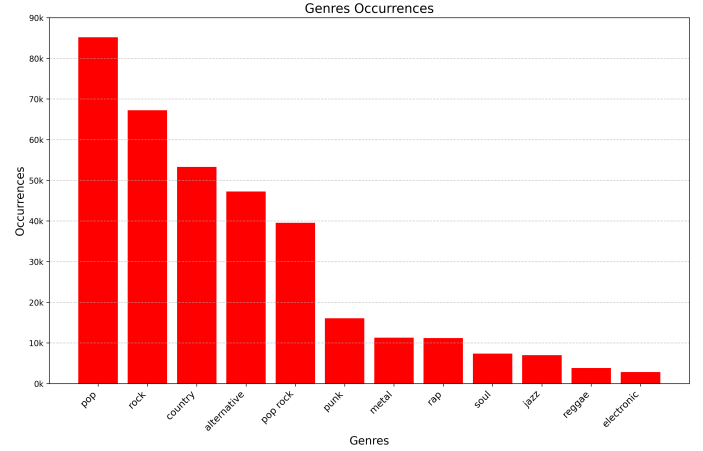

Figure 1: Dataset genres' distribution

Genres Spotify encompasses genres that can be quite specific and may pose challenges for users to easily comprehend (e.g., *lilith*, *otacore*). To address this complexity, we categorized the genres into the 12 most prevalent ones (see Figure 1) following an analysis of our data and referring to the genre genealogy provided by Musicmap[5]. This assisted us to keep the most dominant genre of each artist in order to characterize the tracks' *main genre*, leading to a multiclass classification problem. Tracks whose artists do not fall into any of the 12 most common genres were not assigned a *main genre* and, as a result, are excluded from the genre classification task.

Additionally, we established a classification task specific to rock music using tracks whose artists are categorized within the rock genre. In this context, we retained the least prevalent rock genre associated with each track, focusing on genres with at least 100 occurrences. We've identified 179 distinct rock sub-genres. This task is considered a multiclass classification problem but is quite difficult due to the nuanced similarities between rock sub-genres.

We encoded release dates by decade, resulting in a multiclass classification problem with 15 different decades The majority of tracks were dated from 2010, with the earliest from 1890. This arises from the predominant composition of our dataset, which mainly consists of contemporary music tracks.

Graph Representation In our data analysis, we integrated graph representations. Music tracks often feature intricate harmonies, leading to a multitude of chord connections. Graphs are well-known for their ability to model complex structures and relationships in various systems. For this reason, we used graphs to represent music tracks, as they effectively capture the complex interconnections between chords and are able to detect patterns more easily. Our approach involved transforming each music track into a weighted directed graph (Rousseau and Vazirgiannis 2013), where nodes corresponded to chords, and edges represented
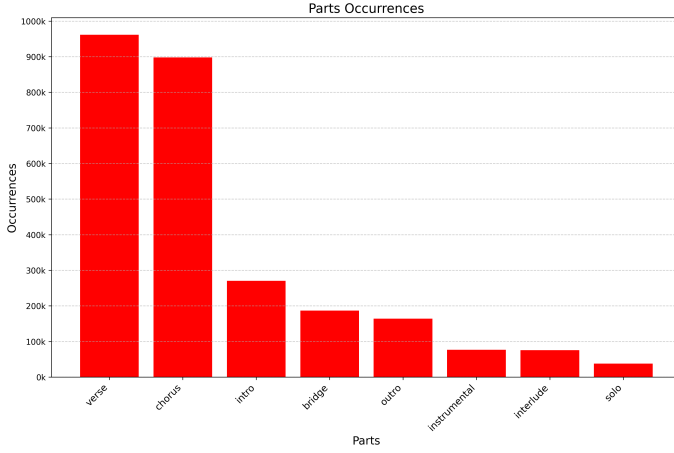
---

[2]https://developer.spotify.com/documentation/web-api

[3]*Dataset link*: https://huggingface.co/datasets/ailsntua/Chordonomicon

[4]*Github link*: https://github.com/spyroskantarelis/chordonomicon

[5]https://musicmap.info

Figure 2: Dataset parts' distribution

No. of Tracks	679,807
No. of Chords	51,994,634
No. of Unique Chords	749
No. of Unique Inverted Chords	3,577
No. of Tracks with Parts	397,580
No. of Parts	2,670,457
No. of Tracks with Release Date	422,181
No. of Tracks with Spotify Genres	429,753
No. of Tracks with Main Genre	352,111
No. of Tracks with Rock Sub-Genre	145,218
No. of Tracks with Spotify Track ID	440,284
No. of Tracks with Spotify Artist ID	510,986

Table 1: Dataset Key Statistics

Exploratory Data Analysis

In this subsection, we conduct an Exploratory Data Analysis (EDA) to systematically examine the dataset's patterns and characteristics. We offer insights on the distribution of chords and explore genre similarities through the use of chord n-grams.

Chords A music chord is a group of notes played together. In western music, chords are usually built on interval of thirds (tertian chords). Different types of chords arise from the quality of their intervals. Chords typically comprise three notes (triads), and those with four or up to seven notes are termed "extended" chords. An exception is the "power" chord, featuring two notes with a fifth interval, commonly employed in metal music. When its notes are rearranged, a chord is called an "inverted" chord.

Our dataset comprises 51,994,634 chords, with 1,464,392 being inverted. Their types are shown in Table 2, while the number of their notes in Table 3. Figure 4 illustrates the chords' distribution of our dataset. It's evident that natural triads, especially the "G" chord, stand out as the most frequently occurring chords. This prevalence is likely due to the simplicity favored in modern music, which dominates our dataset, and the widespread use of capos. Additionally, since these chords are user-generated, they often reflect simplified interpretations of tracks, as non-professional transcribers may reduce complex chords to triads, even if the original music features more intricate harmonies.

It is noteworthy that the data in our dataset are highly similar to the proposed ground-truth reference sets used in chord recognition and music analysis (Burgoyne, Wild, and Fujinaga 2011).

Chord n-grams To enhance our comprehension of the music genres present in our dataset, we employed chord n-grams to compute cosine similarity between genres. Specifically, we focused on quadrigrams, representing chord progressions consisting of four chords. This choice provides insight into the dominant harmonic structures within each genre. It is noteworthy that we excluded quadrigrams formed by repeating two identical bigrams (e.g., "C G C G," which
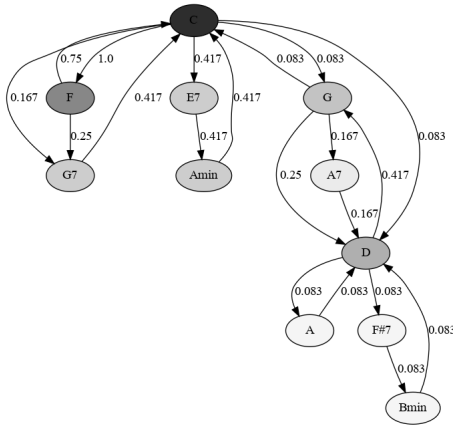

Figure 3: Graph representation of music track with ID 1

chord progressions. The weight associated with each edge reflected the frequency of the corresponding progression. We systematically determined weights by dividing the frequency of a specific progression by the frequency of the most prevalent progression in the dataset.

The graph is defined as $G = (V, E, w)$, with sets V and E representing chords and chord progressions, respectively, and w denoting the weight function assigning weights to edges. Each chord progression e is denoted as an ordered pair (u, v) from the set V, signifying a directed edge from chord u to chord v. The weight function w assigns a numerical value $(0, 1]$ to each directed edge, representing the weight associated with that edge.

In Figure 3, the visual representation of music track ID 1 demonstrates a recurring chord progression from note C to note F, evident in the graph with a weight of 1. The coloration of nodes reflects the frequency of this chord, with darker shades indicating higher occurrence. The dataset comprises 667,858 graphs, each with an average of 6.82 nodes and 13.71 edges, providing insights into the structural characteristics of musical compositions.

Figure 4: Dataset chord distribution

Chord Type	Occurrences
Major	37,031,150
Minor	13,368,240
Suspended	731,520
Power	730,443
Half-diminished	105,281
Augmented	27,627
Diminished	373

Table 2: Dataset Chord Types

Number of Notes	Occurrences
Three	46,529,077
Four	4,466,488
Two	730,443
Five	197,732
Six	41,738
Seven	29,156

Table 3: Dataset Chord Note Count

duplicates "C G") from our analysis. The outcomes are depicted in Figure 5.

Analyzing the cosine similarity among genres yields several significant observations. First, the high similarity values, all close to one, indicate that classification tasks in the MIR domain are challenging when relying solely on the chords of music tracks. Second, there exists a notably higher
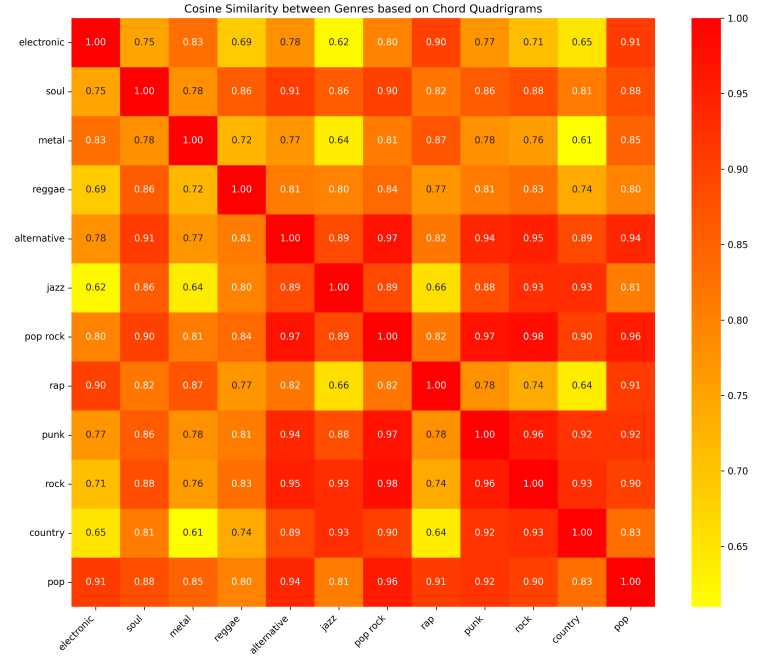

Figure 5: Cosine similarity between genres

similarity among music genres such as *pop*, *rock*, *pop-rock*, *punk* and *alternative*. This suggests that these genres share substantial harmonic structures, given that their melodies are built on similar chords. Consequently, classifying them accurately poses a very challenging task, demanding a more sophisticated harmonic analysis for improved results. This also underscores the inherent ambiguity within music genres, where, for instance, a pop artist might release a rock song or vice versa. Third, the analysis reveals substantial differences between certain genres. For instance, there is a markedly lower similarity observed between *jazz* and *electronic*, as well as between *country* and *rap*. This discrepancy underscores the distinctiveness of these genres, highlighting the diverse harmonic elements that set them apart, that can be exploited in classification tasks.

Learning and Tasks

We focused our experiments on the chord prediction task, as, unlike the classification tasks which are novel to our dataset, we can compare our results with related work (Korzeniowski, Sears, and Widmer 2018). We also present some baseline results for decade and genre classification tasks, serving as initial benchmarks that can be refined and enhanced through further improvements.

Chord Prediction

Chord prediction is a task where given an incomplete chord progression, we have to predict the next chord in the sequence. Formally, given a sequence of chords $(y_1, y_2, \ldots, y_k)$ we predict $P(y_k|(y_1, \ldots, y_{k-1}))$, which means that this task is essentially language modelling. It
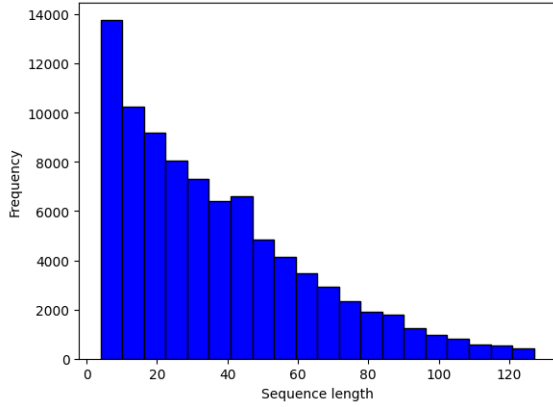
Figure 6: Sequence length distribution

Sequence length	Accuracy (%)
(0, 25.4)	61.23
(25.4, 50.8)	61.15
(50.8, 76.2)	58.44
(76.2, 101.6)	55.65
(101.6, 127)	53.46

Table 4: Chord prediction accuracy

Sequence length	Accuracy (%)
(0, 25.4)	75.85
(25.4, 50.8)	76.03
(50.8, 76.2)	74.80
(76.2, 101.6)	73.34
(101.6, 127)	72.28

Table 5: Note prediction accuracy

has been approached in the past using n-grams and recurrent neural networks (Korzeniowski, Sears, and Widmer 2018). In our experiments, we wanted to explore the efficacy of transformers (Vaswani et al. 2017) for chord prediction, given their performance in natural language processing tasks.

Setting For our implementation we used the Huggingface transformers python package[6], and trained, from scratch a modified version of a GPT-2 architecture (Radford et al. 2019). Specifically, we used the causal language model pre-training on unlabeled chord progressions, and fine-tuned the model for classification on the "decades" task, on the relevant training set. Then we used this fine-tuned model for predicting the next chord on 87,590 chord progressions from the test set of the"decades task" for different sequence's lengths. For each chord progression, a random end was chosen, ranging from the fourth chord up to the second to last chord in the progression. We ended up with the input sequence lengths distribution that is shown in Figure 6.

Results The chord prediction task achieved an accuracy of 60.13%. However, it is noteworthy that the outcomes are contingent on the length of the sequence, as depicted in Table 4. It is observed that the model performs more effectively for sequences with lengths less than 51 and its performance declines as the length increases. In a more in-depth exploration of the model's performance and its musical sensibility, we transformed the chords into their respective notes. The accuracy of note prediction was then calculated, where, for instance, predicting "C" instead of "Amin" would result in a 66.67% accuracy, considering they share two of their three notes. Impressively, the model attains a 75.45% accuracy, indicating that some incorrectly predicted chords still align with musical meaning. Notably, the model consistently achieves higher accuracy for sequences with lengths less than 51, as illustrated in Table 5.

Finally, we also computed the average cumulative probability for chord progressions that had at least 100 chords, so that we could to an extent compare our results with those

---

[6]https://huggingface.co

reported in (Korzeniowski, Sears, and Widmer 2018). The results cannot be directly compared, since we use a much larger chord vocabulary (all chords as opposed to only major and minor triads), and we run our experiment on a much larger dataset (a test set of 87,000 progressions, opposed to 136), however we were interested to see if the results were on par, in addition to observing in more detail the model's behaviour across different sequence lengths. Specifically, we computed:

$$\mathcal{L}(k; M, \mathcal{Y}) = \frac{1}{k|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sum_{i=1}^{k} \log[P(y_k|(y_1, \ldots y_{k-1}))]$$

where M is the GPT-2 based language model, $\mathcal{Y}$ is the set of all chord progressions in the dataset, and k is the length of the progression up to which the cumulative probabilities are computed. This led to the result shown in figure 7. Even though a direct comparison is not apt, we can observe that in general the probabilities assigned to the ground truth chord are mostly higher for the GPT-2 language model, than what related work reports for n-gram and RNN based models, even with a much larger vocabulary. The exception to this seems to be the beginning of the curve (up to the 20 first chords), where the transformer performance seems on par with the RNNs, while later in the sequence the performance of the transformer is significantly improved, then slightly deteriorates for very large sequences, while still being more performant than the RNNs and the n-grams.

Genre and Decade classification

In our methodology, we leverage the graph structure of our dataset to address challenges in the decade and genre classification problem. To accomplish this, we employ a baseline approach based on kernel matrices derived from graph kernels, including the Weisfeiler-Lehman (WL) Kernel (Leman and Weisfeiler 1968), the Shortest Path Graph Ker-
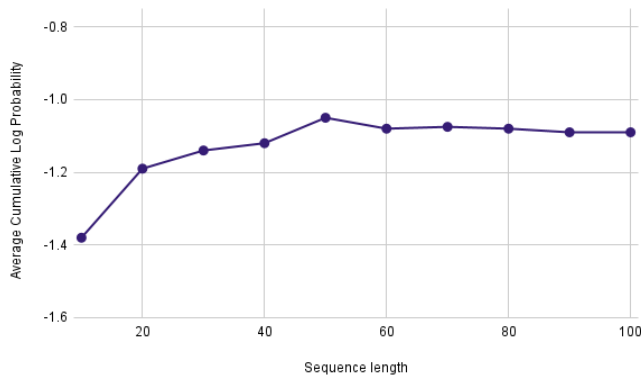
Figure 7: Cumulative log probability per input sequence length

nel (Borgwardt and Kriegel 2005), and the Graph Sampling Kernel (Pržulj 2007).

The methodology employed for the classification tasks establishes a foundational framework for our dataset. The outcomes achieved for the decade classification task and genre classification task were 40.3% and 26.6%, respectively. These figures highlight specific areas that warrant further refinement, suggesting potential avenues for improvement, including the exploration of balanced training, the utilization of Graph Neural Networks (GNNs), and the semantic enrichment of our graph through its interlinkabilty with diverse music domain ontologies.

Conclusion

We present the *Chordonomicon*, an extensive dataset comprising over 666,000 user-generated chord progressions, along with structural information and additional metadata. Accompanying the dataset is a graph representation of it in which each track forms a weighted directed graph, adding a multi-modal dimension. Additionally, we offer insights into its corpus and demonstrate baseline performance on both generative and classification tasks.

For future work, we aim to conduct additional experiments in both generative and classification tasks. Specifically, we plan to experiment on a large scale for determining good tokenization schemes for chords and chord progressions, in addition to appropriate neural architectures and hyperparameters, and we are certain that the contributions of the wider community will be imperative for this endeavour. Another important aspect of this work is exploring how our dataset can augment model performance in chord recognition tasks. Additionally, we intend to enrich our dataset in the realm of music theory and harmony by incorporating notions of music ontologies. This approach seeks to examine how the infusion of musically meaningful information can improve the performance of difficult tasks related to genre and decade classification.

Finally, as Chordonomicon offers a unique blend of information and representations on a very large scale, we hope that it might be useful as a tool for AI researchers from other domains, besides music, such as graph machine learning, and natural language processing.

Ethical Statement

Data Harvesting Legality and Ethicality   1. *Web Crawler Permissions*: Ultimate Guitar (hereafter "UG") does not flag "tabs.ultimate-guitar.com/tab/*" as "Disallowed" in their /robots.txt file. This indicates that UG permits automated crawlers to access and index their tablature and chord pages, aligning with established web protocols for content accessibility. 2. *Terms of Service Compliance*: A thorough examination of UG's Terms of Service (TOS) reveals no obvious prohibition against automated content crawling or scraping. This absence of such restrictions further supports the permissibility of data collection for research purposes. 3. *Legal Precedents*: U.S. courts have consistently ruled in favor of web-scraping publicly accessible data, particularly for research purposes. Notably, in Sandvig v. Barr (D.D.C. 2020), the court held that such activities are protected under the First Amendment and do not violate the Computer Fraud and Abuse Act (CFAA). Additionally, the hiQ Labs, Inc. v. LinkedIn Corp. (9th Circuit, 2019) decision further reinforced the legality of scraping publicly available data.

Intellectual Property Considerations   1. *User-Generated Content Ownership*: UG's TOS explicitly states that "Ultimate Guitar does not claim any ownership rights in User Generated Content that you transmit, submit, display or publish on, through or in connection with the Service". And further clarifies that "User Generated Content includes, without limitation, tablatures (text or electronic)...". This clause clearly delineates the ownership status of the content we're collecting. 2. *Non-Copyrightable Elements*: Legal precedents consistently support the notion that chord progressions are not copyrightable. Cases such as Granite Music Corp. v. United Artists Corp. (1977) and Swirsky v. Carey (2004) have established that chord progressions are fundamental building blocks of music, akin to "common musical property", since many songs share the same or similar chord progressions. This legal stance corroborates that our dataset solely comprises of non-copyrightable elements. 3. *Data Processing*: To further ensure ethical and legal compliance, our data collection process excludes copyrightable elements such as lyrics, song titles, and artist names. We retain only the chord progressions, the associated song structure information (e.g., intro, verse, chorus), release year, the genres of the relevant artist and Spotify IDs of track and artist (non-copyrightable alphanumeric strings).

Context and Licensing   1. *Precedent in Existing Datasets*: The Common Crawl dataset, widely used for training most top-tier large language models, consistently crawls and already includes numerous, if not all, chord pages from ultimate-guitar.com in its archives. A relevant search on https://index.commoncrawl.org/ will show all the harvested UG URLs, and the relevant Common Crawl web archive (.warc) files will show the raw scrapped pages, along with the lyrics, song names, bands, and all information visible on each tab's webpage. This precedent demonstrates the accepted practice of including such data in large-scale research

datasets. 2. *Licensing Considerations*: The Common Crawl dataset's license is a limited license that allows users to access and utilize the data while agreeing to respect the copyrights and other applicable rights of third parties in and to the material contained therein. Despite the non-copyrightable nature of our data, we still adopt the same limited licensing approach for peace of mind.

In conclusion, by adhering to web crawling conventions, respecting Terms of Service, focusing on non-copyrightable elements, and following well-established precedents in data collection and licensing, our approach to creating this dataset stands on solid legal and ethical ground. The potential scientific and cultural value of this research further justifies its creation and use within the bounds of fair use and academic freedom.

References

Bertin-Mahieux, T.; Ellis, D.; Whitman, B.; and Lamere, P. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 591–596.

Bogdanov, D.; Porter, A.; Schreiber, H.; Urbano, J.; and Oramas, S. 2019a. The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale. In *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019): 2019 Nov 4-8; Delft, The Netherlands.[Canada]: ISMIR; 2019.* International Society for Music Information Retrieval (ISMIR).

Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019b. The mtg-jamendo dataset for automatic music tagging. ICML.

Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019c. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, ICML.*

Booth, N. 2016. Backing Down: Blurred Lines in the Standards for Analysis of Substantial Similarity in Copyright Infringement for Musical Works. *J. Intell. Prop. L.*, 24: 99.

Borgwardt, K. M.; and Kriegel, H.-P. 2005. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, 8–pp. IEEE.

Boulanger-Lewandowski, N.; Bengio, Y.; and Vincent, P. 2013. Audio Chord Recognition with Recurrent Neural Networks. In *ISMIR*, 335–340. Curitiba.

Burgoyne, J. A.; Wild, J.; and Fujinaga, I. 2011. An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis. In *ISMIR*, volume 11, 633–638.

Cherfi, S. S.-s.; Guillotel, C.; Hamdi, F.; Rigaux, P.; and Travers, N. 2017. Ontology-Based Annotation of Music Scores. In *Proceedings of the 9th Knowledge Capture Conference*, K-CAP '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355537.

de Berardinis, J.; Meroño-Peñuela, A.; Poltronieri, A.; and Presutti, V. 2023. Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs. *Scientific Data*, 10(1): 641.

De Haas, W. B.; Rodrigues Magalhães, J.; Veltkamp, R. C.; Wiering, F.; et al. 2011. Harmtrace: Improving harmonic similarity estimation using functional harmony analysis. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR).*

Dervakos, E.; Kotsani, N.; and Stamou, G. 2021. Genre recognition from symbolic music with cnns. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*, 98–114. Springer.

El Achkar, C.; and Atéchian, T. 2020. Supporting Music Pattern Retrieval and Analysis: An Ontology-Based Approach. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, WIMS 2020, 17–20. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375429.

Fazekas, G.; Raimond, Y.; Jacobson, K.; and Sandler, M. 2010. An overview of semantic web activities in the OMRAS2 project. *Journal of New Music Research*, 39(4): 295–311.

Futia, G.; and Vetrò, A. 2020. On the integration of knowledge graphs into deep learning models for a more comprehensible AI—Three challenges for future research. *Information*, 11(2): 122.

George, A.; Mary, X. A.; and George, S. T. 2022. Development of an intelligent model for musical key estimation using machine learning techniques. *Multimedia Tools and Applications*, 81(14): 19945–19964.

Gong, X.; Zhu, Y.; Zhu, H.; and Wei, H. 2022. ChMusic: A Traditional Chinese Music Dataset for Evaluation of Instrument Recognition. In *Proceedings of the 4th International Conference on Big Data Technologies*, ICBDT '21, 184–189. New York, NY, USA: Association for Computing Machinery. ISBN 9781450385091.

Good, M. 2001. MusicXML for notation and analysis. *The virtual score: representation, retrieval, restoration*, 12(113–124): 160.

Harte, C.; Sandler, M. B.; Abdallah, S. A.; and Gómez, E. 2005. Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations. In *ISMIR*, volume 5, 66–71.

Hentschel, J.; Moss, F. C.; McLeod, A.; Neuwirth, M.; and Rohrmeier, M. ???? Towards a Unified Model of Chords in Western Harmony. In *Music Encoding Conference 2021.*

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Philip, S. Y. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2): 494–514.

Jones, J.; de Siqueira Braga, D.; Tertuliano, K.; and Kauppinen, T. 2017. MusicOWL: The Music Score Ontology. In

Proceedings of the International Conference on Web Intelligence, WI '17, 1222–1229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349512.

Kantarelis, S.; Dervakos, E.; Kotsani, N.; and Stamou, G. 2023. Functional harmony ontology: Musical harmony analysis with Description Logics. *Journal of Web Semantics*, 75: 100754.

Koops, H.; de Haas, W.; Burgoyne, J.; Bransen, J.; Kent-Muller, A.; and Volk, A. 2019. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, 48: 1–21.

Korzeniowski, F.; Sears, D. R.; and Widmer, G. 2018. A large-scale study of language models for chord prediction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 91–95. IEEE.

Leman, A.; and Weisfeiler, B. 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 2(9): 12–16.

Lisena, P.; Meroño-Peñuela, A.; and Troncy, R. 2022. MIDI2vec: Learning MIDI embeddings for reliable prediction of symbolic music metadata. *Semantic Web*, 13(3): 357–377.

Louboutin, C.; and Bimbot, F. 2016. Description of chord progressions by minimal transport graphs using the System & Contrast model. In *ICMC 2016-42nd International Computer Music Conference*.

Lyberatos, V.; Kantarelis, S.; Dervakos, E.; and Stamou, G. 2023a. Perceptual Musical Features for Interpretable Audio Tagging. *arXiv preprint arXiv:2312.11234*.

Lyberatos, V.; Kantarelis, S.; Kaldeli, E.; Bekiaris, S.; Tzortzis, P.; Menis-Mastromichalakis, O.; and Stamou, G. 2023b. Employing Crowdsourcing for Enriching a Music Knowledge Base in Higher Education. In *International Conference on Artificial Intelligence in Education Technology*, 224–240. Springer.

Moog, R. A. 1986. Midi: Musical instrument digital interface. *Journal of the Audio Engineering Society*, 34(5): 394–404.

Morris, C.; Kriege, N. M.; Bause, F.; Kersting, K.; Mutzel, P.; and Neumann, M. 2020. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*.

Nadar, C.-R.; Abeßer, J.; and Grollmisch, S. 2019. Towards CNN-based acoustic modeling of seventh chords for automatic chord recognition. In *International Conference on Sound and Music Computing. Málaga, Spain*.

Nieto, O.; McCallum, M. C.; Davies, M. E.; Robertson, A.; Stark, A. M.; and Egozy, E. 2019. The Harmonix Set: Beats, Downbeats, and Functional Segment Annotations of Western Popular Music. In *ISMIR*, 565–572.

Oramas, S.; Sordo, M.; Espinosa-Anke, L.; and Serra, X. 2015. A semantic-based approach for artist similarity. In *Müller M, Wiering F, editors. Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference; 2015 Oct 26-Oct 30; Malaga, Spain.[Sl]: International Society for Music Information Retrieval; 2015*.

p. 100-6. International Society for Music Information Retrieval (ISMIR).

Park, J.; Choi, K.; Jeon, S.; Kim, D.; and Park, J. 2019. A bidirectional transformer for musical chord recognition. *arXiv preprint arXiv:1907.02698*.

Paulus, J.; and Klapuri, A. 2009. Labelling the structural parts of a music piece with Markov models. In *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music: 5th International Symposium, CMMR 2008 Copenhagen, Denmark, May 19-23, 2008 Revised Papers 5*, 166–176. Springer.

Pauwels, J.; O'Hanlon, K.; Gómez, E.; Sandler, M.; et al. 2019. 20 years of automatic chord recognition from audio.

Pržulj, N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2): e177–e183.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raimond, Y.; Abdallah, S.; Sandler, M.; and Giasson, F. 2007. The Music Ontology. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*.

Rashid, S. M.; De Roure, D.; and McGuinness, D. L. 2018. A music theory ontology. In *Proceedings of the 1st International Workshop on Semantic Applications for Audio and Music*, 6–14.

Rousseau, F.; and Vazirgiannis, M. 2013. Graph-of-word and TW-IDF: new approach to ad hoc IR. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 59–68.

Smith, J. B. L.; Burgoyne, J. A.; Fujinaga, I.; De Roure, D.; and Downie, J. S. 2011. Design and creation of a large-scale database of structural annotations. In *ISMIR*, volume 11, 555–560. Miami, FL.

Speck, J. A.; Schmidt, E. M.; Morton, B. G.; and Kim, Y. E. 2011. A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. In *ISMIR*, volume 104, 549–554.

Sutton, C.; Raimond, Y.; Mauch, M.; and Harte, C. 2007. The chord ontology. *URL http://purl. org/ontology/chord*.

Takuya, F. 1999. Realtime chord recognition of musical sound: Asystem using common lisp music. In *Proceedings of the International Computer Music Conference 1999, Beijing*.

Tiddi, I.; and Schlobach, S. 2022. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302: 103627.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.-C.; Hung, Y.-N.; and Smith, J. B. L. 2022. To Catch A Chorus, Verse, Intro, or Anything Else: Analyzing a Song with Structural Functions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 416–420.

Wang, Z.; Chen, K.; Jiang, J.; Zhang, Y.; Xu, M.; Dai, S.; Gu, X.; and Xia, G. 2020. Pop909: A pop-song dataset for music arrangement generation. *arXiv preprint arXiv:2008.07142*.

Weiß, C.; Schreiber, H.; and Müller, M. 2020. Local Key Estimation in Music Recordings: A Case Study Across Songs, Versions, and Annotators. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2919–2932.