Musicological Interpretability in Generative Transformers

Nicole Cosme-Clifford
Dept of Music
Yale University
New Haven, CT, USA
nicole.cosme@yale.edu

James Symons
Louiville, KY, USA
WordWood Collective
symons.james@gmail.com

Kavi Kapoor
Department of Music
Harvard University
Cambridge, MA
kmkapoor@college.harvard.edu

Christopher Wm. White
Department of Music and Dance
University of Massachusetts Amherst
Amherst, MA, USA
ORCHID: 0000-0002-6435-6423

2023 4th International Symposium on the Internet of Sounds | 979-8-3503-8254-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59510.2023.10335202

***Abstract*—What might an unsupervised autoregressive transformer learn about chord syntax in chorale-style Western-European-style chord progressions? In this paper, we implement a novel chord representation to explore the behavior and constitution of such a model with a proprietary corpus of chorale-style church hymns. We then conduct an explainability study on our model. First, we track updates to the model's token embeddings over time, and we visualize the resulting embedding spaces to which the model converges. This tells us what kind of syntactic relationships are learned between chords in this corpus. We then study chord progressions generated by two such models, with short and long respective training periods, and identify what musicological and pedagogical concepts are being learned at each stage. We find the model learns basic syntactic categories of chords, and that these categories resonate with some musicological discourse surrounding chord behavior in this style. The model also learns several components harmonic behavior in this repertoire, including smooth voiceleading, ending formulas, and treatment of chromatic pitches, and learns these concepts in an order that approximates undergraduate harmony textbooks. However, the model shows no evidence of learning broader organizational principles, like phrase structure, repetition, and meter.**

***Index Terms*—music, transformer, symbolic music, interpretability, alignment, harmonic function**

I. Introduction

Automatic music generation has been a longstanding pursuit within the domain of Music Information Retrieval (MIR). The advent of deep learning revolutionized the field in the early 2010s, particularly through the use of Deep Neural Networks (DNNs), which exhibited the capacity to capture intricate musical relationships that were previously unattainable [1, 2, 3]. Subsequently, Recurrent Neural Networks (RNNs) and their Long Short-Term Memory (LSTM) variants enhanced neural networks' ability to model sequential patterns in music, surpassing the capabilities of Hidden Markov Models (HMMs) [4, 5, 6]. Furthermore, Convolutional Neural Networks (CNNs) proved valuable in modeling relationships between musical phenomena derived from spectral features [7, 8, 9]. However, despite these advancements, both RNN-based and CNN-based algorithms still encounter challenges in effectively modeling and generating musical relationships across broader time scales [10, 11].

Transformers emerged several years ago as a promising solution to broader dependency problems in music generation, particularly in how the musical events are contingent on other events separated by large amounts of time [12]. Since their inception, Transformers have found extensive applications in various generative music tasks, such as conditional melody generation [13, 14], chord progression generation [15], rhythm generation [16], and even full-audio generation [17]. Moreover, the integration of reinforcement learning and human-tagged data raises the potential for these models to grasp increasingly sophisticated aspects of music composition. Despite their widespread adoption and success, however, Transformers' inner workings often remain opaque, especially concerning the limits of their unsupervised learning capacity and their alignment with cognitive and pedagogical understandings of music. Furthermore, because transformers are frequently complemented with a supervised procedure or train on human-tagged data, the actual potential of (and limits to) purely unsupervised machine-learning using raw, un-tagged data remains unclear. To address this, our study limits our corpus to unanalyzed pitch events, and uses a simple Transformer with no supervised supplementations.

Recent studies have taken strides towards addressing the opacity of Transformers in music generation by exploring the concept of musical self-attention [18, 19, 20, 21]. However, these investigations have primarily focused on classification models for automatic chord recognition (ACR). In such models, the notion of *harmonic function*, which involves categorizing chords based on their syntactic roles in chord progressions, is treated as an *a priori* concept. Chords are strictly assigned to single categorical labels, often derived explicitly from music theory pedagogy, such as Roman numerals like I or V, which signify the chord's position within a key and scale. This categorical approach clashes with musicological understandings of harmonic function, which propose that a single chord may fulfill multiple roles within a musical grammar [22, 23]. To overcome this limitation, we propose an alternative approach

based on generative models that embrace the broader and more nuanced definition of harmonic function in music.

In this study, we employ a generative autoregressive transformer to model harmonic progressions from the Western-European tonal tradition, represented in symbolic notation. Our training data exclusively utilizes the pitch parameter, and no human tagging or reinforcement is involved in the training process. The model undergoes analysis to assess its musicality and the quality of its generated results. We begin by examining chord progressions generated by the model after small-scale training and then proceed to assess the outcomes after more extensive training. By doing so, we compare the resulting progressions against pedagogical concepts and norms associated with this particular musical repertoire. Moreover, to gain insights into the model's understanding of harmonic grammar and function, we visualize chord embeddings at different layers of the transformer architecture. This analysis enables us to investigate how the model converges towards an understanding of harmonic grammar that aligns with musicological pedagogy. Overall, our aim is to identify which aspects of this musical style the model has learned and, conversely, which aspects it may be lacking.

II. Related Work

A. Representation of musical events

In the realm of music research, several approaches have been employed to organize symbolic musical data in formats that are both machine-learnable and suitable for music generation. One prevalent method is to utilize MIDI (Musical Instrument Digital Interface) or an analogous representation. In the MIDI format, pitches are represented as numerical values, where 0 corresponds to C-1 (8.18 Hz), and each subsequent integer represents a half step higher. For instance, middle C is represented by 60, and the C-sharp above it is represented by 61.

The usage of integer-based representations proves highly effective in capturing sequential "harmonies" within a musical piece. Through the segmentation of MIDI pitch content at each sequential pitch change, we create a series of ordered moments representing the harmonic progression [24]. Transposing each piece in a corpus to the key of C can limit the corpus's vocabulary size by mapping all different key instantiations of a chord progressions to one key [25]. Moreover, by discarding octave designations, set ordering, and doubling, we can further decrease vocabulary size (for instance, chords containing the notes C, E, and G are deemed identical, irrespective of their octaves or note occurrences). However, for these simplified representations to be effectively used in generative models, it becomes essential to devise rules that can accurately map these simplified tokens back to specific pitches in specific octaves [26].

Alternatively, some researchers have employed voiceleading between chords as the central object for musical machine learning [27, 28]. In this approach, the focus shifts from the individual sounding events to the changes in each constituent instrument, voice, or musical line between consecutive moments. The emergent behavior in this context arises not from the sounding events themselves but rather from the connections and transitions between these events.

B. Harmonic Function

Harmonic function has been an interest of music researchers for centuries [29, 30]. The concept groups the wide diversity of chords, keys, and pitch deployments available to a composer and reduces them to a handful of *functions*, a concept roughly akin to grammatical parts of speech [31]. Unlike grammar, however, the concept also relies on the pitch constituency of a chord to define its functional role [32]: because particular scale degrees have particular syntactic behaviors and attractions to other scale degrees, a chord's constituency will influence its own syntactic behavior. While the constituency and number of these categories can be contentious [33], three central functions are often included in these models: 1) a *tonic* function that contains primarily chords build on a key's first scale degree (i.e., a C major triad in the key of C would be "tonic"), 2) a *dominant* function that contains chords that primarily progress to tonic chords and who use a key's fifth scale degree (i.e., a G major triad in C major is usually a "dominant" chord, and frequently progresses to tonic chords), and 3) *subdominant* chords that succeed tonic chords and precede either dominant or tonic chords, and which usually contain a scale's fourth scale degree (D minor triads in the key of C major would be a subdominant chords, insomuch as they usually follow tonic chords and either returns there or progresses to dominant harmonies). Regardless of its complexity and fragility, this concept has exhibited cognitive validity [34]. Additionally, because it relies on the behaviors and tendencies of a harmony's constituent notes/scale degrees, *functional harmony* goes beyond merely specifying which chords can follow one another; it also encompasses the concept of *voiceleading*. Voiceleading pertains to how the individual notes (or voices) of the chords connect to one another [23], with certain pitches within a key tending to progress to other pitches in very specific ways [30]. In this style, "smooth" voiceleading — or, using the smallest amount of distance between notes in sequential harmonies — is favored ([35]).

Efforts to capture musical grammars and harmonic function have taken various forms in the realm of machine learning. One approach involves using the states of a Hidden Markov Model (HMM) to study this concept [33, 36, 37]. Additionally, clustering techniques and information bottlenecks have been employed to extract relevant patterns and structures related to harmonic function [38, 39].

C. Music Transformers

Music-based transformers emerged soon after text-based transformers [40]. However, in their early stages, these models were mostly limited to generating short musical passages, often lacking musical coherence. Music Transformer was the first mainstream model to successfully generate music snippets with long-term structure and internal consistency [12].

Transformer models have found extensive applications in various facets of music research. Music generation tasks have been addressed by models like Museformer [41], Pop Music Transformer [42], and EnsMuse [43]. The domain of automatic chord recognition (ACR) has seen the contributions of Harmony Transformer (HT) [18] and Bi-Directional Transformer for Chord Recognition (BTC) [19].

Beyond these tasks, transformer models have demonstrated their utility in diverse areas. Music tagging has been explored using vision transformers [44] and semi-supervised transformers [45]. Novel frequency information has been captured by transformers processing spectral representations of audio recordings [46] [47]. Transformers have also shown their capabilities in style identification research [48] [49]. More recently, transformer models have become assistants to music composers [50] [51] and have begun to model human expressiveness in music [52].

Expanding beyond conventional music tasks, transformers have exhibited their prowess in text-to-audio generation with ERNIE [53] and MusicLM [54]. Emotion-conditioned music generation has also been explored [55] [56].

D. Harmonic Function and Emergent Behaviors in ACR Transformers

Automatic chord recognition (ACR) is a classification problem that assigns static labels (chord identity labels, chord quality labels, functional labels) to chords. Bi-Directional Transformer (BTC) was the first transformer model applied to ACR [19]. Similar to text-based transformers, it improves upon the shortcomings of CNNs and RNNs in capturing long-term dependency information between musical features. Taking a slightly different approach, Harmony Transformer (HT) is an autoregressive transformer that predicts chord labels for a given sequence according to a learned segmentation scheme [18]. Most relevant to our purposes here, these studies find that each attention head in an ACR transformer pays attention to a different section of the input depending on the syntactical nature of the chord whose label is being predicted [20]. This suggests that harmony-focused transformers have some implicit sense of harmonic function.

However, the above methods consistently rely on *a priori* categories and human labeling, raising concerns about the potential influence of preconceived notions on the learning process. Ebrahimzadeh et al. [57] have identified this issue and propose an unsupervised model that learns a chord embedding matrix. This matrix is then incorporated into a transformer-based ACR model, although it remains static during training, unlike the updating process in BTC or HT. Another approach to ACR involves the use of pre-trained chord embeddings based on pitch overlaps between chords [35]. These methods have shown promising results, yet they once again introduce predetermined notions of chord relationships into the learning process.

Rather than focusing on performance gains or pre-training, we ask if a well-trained end-to-end baseline transformer will learn to draw clear and sufficient boundaries around grammatical harmonic categories in its chord embedding space. Furthermore, we ask if such boundaries, and the clusters of chords they define, align with traditional understandings of harmonic function in common practice Western chorales. If not, what new things do we learn about harmony and harmonic function in our corpus?

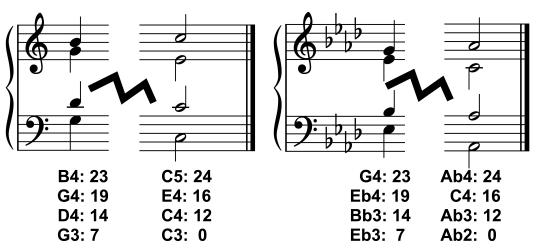

Fig. 1. Normalizing MIDI numbers by indexing from a hymn's last lowest pitch

III. METHODS

A. Corpus Methods

The corpus comprises hymns sourced from the Western European Christian tradition, presented in MIDI format. These hymns were gathered from various free online collections available at openhymnal.org, hymntime.com, lutheranmusic.com, opc.org, mountainretreatorg.net, carolynshymns.com, gbod3.org, smallchurchmusic.com, and emp.byui.edu. Specifically, representations were chosen based on their chorale-style homophonic texture, featuring keyboard and/or choir arrangements. The resulting collection encompassed a total of 2,932 files. As some files captured subtle microtiming from human performances, the events were bundled into the nearest eighth note using music21's quantize function [58]. To ensure a consistency throughout our training data, files with more than 20% of their chord events unique to that particular file were excluded. This filtration process resulted in a refined corpus containing 2,225 files, consistent with the broader harmonic style of the collection.

In this study, each moment wherein a pitch is added or subtracted from the texture is considered a *chord* [25]. The pitch content of these chords underwent a key-based normalization using a novel method. As depicted in Figure 2, the last lowest pitch in each hymn is designated as zero, and all other notes in the hymn are named according to their half-step distance from that pitch. For instance, a pitch a perfect fifth above the last lowest pitch is labeled as 7, while a minor third lower is designated as -3. This normalization approach ensures consistency across keys. While the chords shown in Figure 2 might represent different pitches in two keys, they maintain the same interval relationships with respect to the hypothetical hymn's last, lowest note.

These integers were organized as ordered sets, subsequently represented as sequential cells in a comma-separated values

(csv) file, accessible upon request. To signify both the beginning of a hymn's sequence and the conclusion of the previous one, a consistent *Start/End* token was inserted between hymns. For processing, a unicode character was assigned to each unique set, enabling the representation of the entire corpus as a unicode string. Concurrently, a decoding key was constructed, and this string was used as the training corpus. Ultimately, the final training corpus encompassed 350,528 characters, featuring a vocabulary of 5,185 unique characters.

This corpus offers a valuable advantage by enabling easy comparisons with an extensive range of music-theory pedagogical validation. Christian hymnody often employs functional harmony, a grammar which was both prevalent in European and American art- and sacred-music composition from the 17th to the 20th centuries, and continues to influence contemporary traditional-sounding music [59]. Additionally, analysts of this chord grammar and of this style explicitly reference harmonic function [60]. This grammar also boasts a long and intricate pedagogical heritage [23, 22]. Consequently, this well-defined harmonic grammar serves as a robust yardstick to evaluate the output of a model trained on this specific musical style.

B. Computational Methods

Architecture. Our model, shown in Figure 1, is a decoder-only autoregressive transformer, an altered version of Andrej Karpathy's NanoGPT [61]. It is deliberately simple and narrow in scope, as we wanted to investigate how effectively a baseline transformer with minimal training could learn difficult concepts like chord syntax and harmonic function. As explained in the preceding subsection, our model's input is a sequence of chords, each of which is represented by a unicode character. The model otherwise follows the standard architecture of a decoder-only transformer: it has L layers, each of which does multi-head self-attention followed by a feedforward pass and a normalization layer. The output of the final layer is then passed through a fully-connected layer and a softmax activation layer. At time step t, this draws a probability distribution over all possible unicode chord tokens, from which the model samples to generate a token at time step $t+1$.

For the purposes of this paper our hyperparameters were set as follows: a block size of 6, a learning rate of $1e^{-5}$, 64-dim embedding vectors, 4 attention heads, 4 layers, and a 20% dropout rate. In addition to optimizing our model in mathematical terms (i.e. *cross-entropy* loss), we also optimize in musicological terms by selecting models that produced desirable musical features for further fine-tuning.

Evaluation. Two primary methods were used to evaluate and analyze the model's output.

1) We extract the model's embedding vectors (the result of its encoder) and the linear transformations of these that result from each self-attention layer, l. In each case, we project these vectors to a 2D space to study how the model's understanding of a chord's grammatical (or, *functional*) similarity changes as it trains and as the model's layers become deeper. Figure 3 shows the resulting projections of (a) the embedding space and (b) the logit vector space from the final self-attention layer. This enables us to evaluate the full scope of our model's grasp on harmonic function.

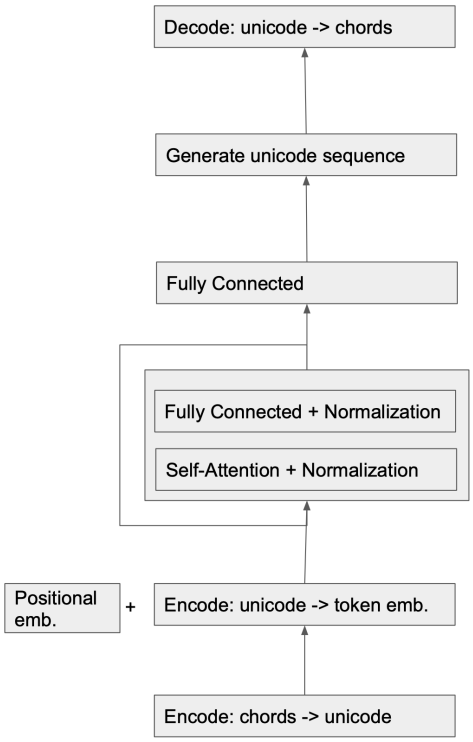

Fig. 2. Model architecture. Our model has two encoding steps: chord vectors to single unicode characters, and unicode characters to embedding vectors. Those vectors then pass through the body of the model: a series of attention layers, then linear layer, then softmax activation. Last, the model generates a unicode sequence (its predicted chord progression) and we decode that sequence back into chord vectors.

2) We undertook expert analysis of our model's outputs at 10,000 and 1,000,000 training iterations (about 10 minutes and 12 hours of training, respectively). After each of these training sessions, the model generated 5 "compositions" by producing unicode tokens until a *Start/End* token was generated. These tokens were translated back into chord structures, which were themselves translated into MIDI representations, such that 0 aligned with C3 (the octave below middle C). While future studies would use external evaluators to produce this feedback, in the current pilot study the authors evaluated the results. Each of the authors is an expert in Western functional harmony, with several having extensive experience in in European Hymnody. We evaluate these compositions for their alignment to music-theoretical benchmark concepts, such as the ability to model cadential syntax, assert a tonic, and modulate.

IV. Results

n-dimensional vector embedding Figure 3 illustrates two representative graphs depicting the chord syntax learned by our model after 1 million iterations. The chords are presented using a simplified representation based on octave equivalence modulo 12 (e.g., all C's represented as 0, all D's as 2, etc., as shown in the legend). We have color-coded the chords to indicate their degree constituency, particularly whether they are rooted on the *tonic* scale degree (blue), rooted on the *dominant* degree (orange), or prominently feature the *subdominant* scale degree (green and purple). This was done to make the graph's clusters' interaction with traditional notions of harmonic function easier to analyze. Although the clustering is not perfect from a musicological or pedagogical perspective, several notable groupings occur. For one, the dominant triad [2, 7, 11] often clusters with the dominant harmony with an added seventh [2, 5, 7, 11]. Additionally, the subdominant triad [0, 5, 9] clusters with other triads that share two of its scale degrees (like [2, 5, 9] and [0, 4, 9], the supertonic and submediant triads).

Conversely, some clusters appear atypical based on chord spelling but correctly represent chords with similar syntactic functions. For instance, the orange cluster near the bottom of Figure 3(a) (left of center), contains chords that are spelled like dominant-functioning chords but serve as a hymn's initial chord, thus sharing the phrase-initiating function of the nearby blue, tonic-functioning chords.

Moreover, the purple dots in various clusters of Figure 3(a) capture the versatile functions of the submediant (vi) triad [0, 4, 9]. Some clusters seem to express its function as a substitute for a tonic chords, while others denote its function as a subdominant or secondary dominant. (The respective clusters for these functions emerge near the top left corner of the graph, close to the green cluster at the bottom right corner, and adjacent to the orange clusters on the right side.) These grammatical roles are precisely those described in harmony textbooks ([23, 22, 33].

Interpretation of outputs. Figure 4 displays outputs from the model after 10,000 training iterations. (NB: the output does not have a designated meter, but for ease of reading, we have added simple quadruple bar lines.) Remarkably, even with this relatively small amount of training, the model exhibits emergent behaviors that reflect grammatical organization. Diatonic successions (or, "white-note" chords in C major), follow stylistically coherent chord orderings and demonstrate predictable voiceleading. Note, for instance, several instances of traditional dominant-to-tonic and tonic-to-subdominant progressions, along with the smooth, step-wise voiceleading between harmonies. For instance, in the first, left-hand "Successful Learning" passage, we observe a penultimate first-inversion C major triad leading into the final harmony with a voice exchange between the lower two voices. Furthermore, the initial two chords of that measure feature a rising lowest, bass voice paired with descending upper voices; both of these techniques are commonly taught in undergraduate textbooks [23]. Additionally, sequences preceding "Start/End" tokens, marking the ends of phrases, exhibit recognizable stock patterns often used to conclude hymns. For instance, Figure 4 illustrates a phrase ending with a traditional "Amen" cadence (note the subdominant F major harmony moving to the final C major triads).

On the other hand, the "Unsuccessful Learning" row in the figure presents sequences that deviate from the stylistic norms. In the left-hand phrase, harmonies from the parallel minor (C minor in this case) appear in an unusual (seemingly arbitrary) grammatical context, and the voiceleading is disjointed, featuring numerous large leaps from one chord to the next. Additionally, the model has not grasped the behavior of certain chords in this style that follow strict grammatical constraints. Specifically, chords with raised notes should lead upward, and chords with lowered notes should lead downward —— a characteristic of *chromatic* chords in this style. However, the right-hand "Unsuccessful Learning" passage flouts these strictures (Note: the example is a composite of segments from various outputs, with double bar lines denoting the boundaries between these segments).

Continued training, however, adds sophistication to these models, as shown in Figure 5. After a million iterations, the outputs now both feature stylistically coherent passages in the relative minor and expected resolutions of chromatic harmonies (although there are intermittent mistakes as shown in the figure), and even feature some extended passages of sophisticated chromatic harmony. Here, not only are the orderings of the harmonies stylistically grammatical, but the voiceleading between the chords follows the expected smooth, step-wise norms.

V. Discussion

Pedagogical Resonance and Learning. The model's improvements during both shorter and longer training sessions demonstrate alignment with the ordering of comparable pedagogical material. Tonal harmony textbooks often introduce fundamental concepts like "Tonic, Dominant, and Voice Leading" or "Expanding Tonic and Dominant" before covering more advanced topics such as "Applied Chord," "Modulation," and "Modal Mixture" [22, 23]. In other words, when it first learns diatonic chords progressions and subsequently adds knowledge of chromatic and minor-mode harmony, the model's learning progression echoes a pedagogical sequence.

Moreover, behavioral experiments have revealed that individuals with greater musical training develop heightened sensitivity to the behaviors of chromatic chords [62, 34] and the distinct voiceleading tendencies found in specific musical styles [63]. The model's learning order also parallel this difference between less and more experienced musicians.

The model's learning order is likely influenced by the statistical frequency and predictability of different musical events. Frequent and highly predictable events are learned first, while less common and more variable sequences require additional iterations for the model to master. This learning process mirrors how humans gain confidence in predicting

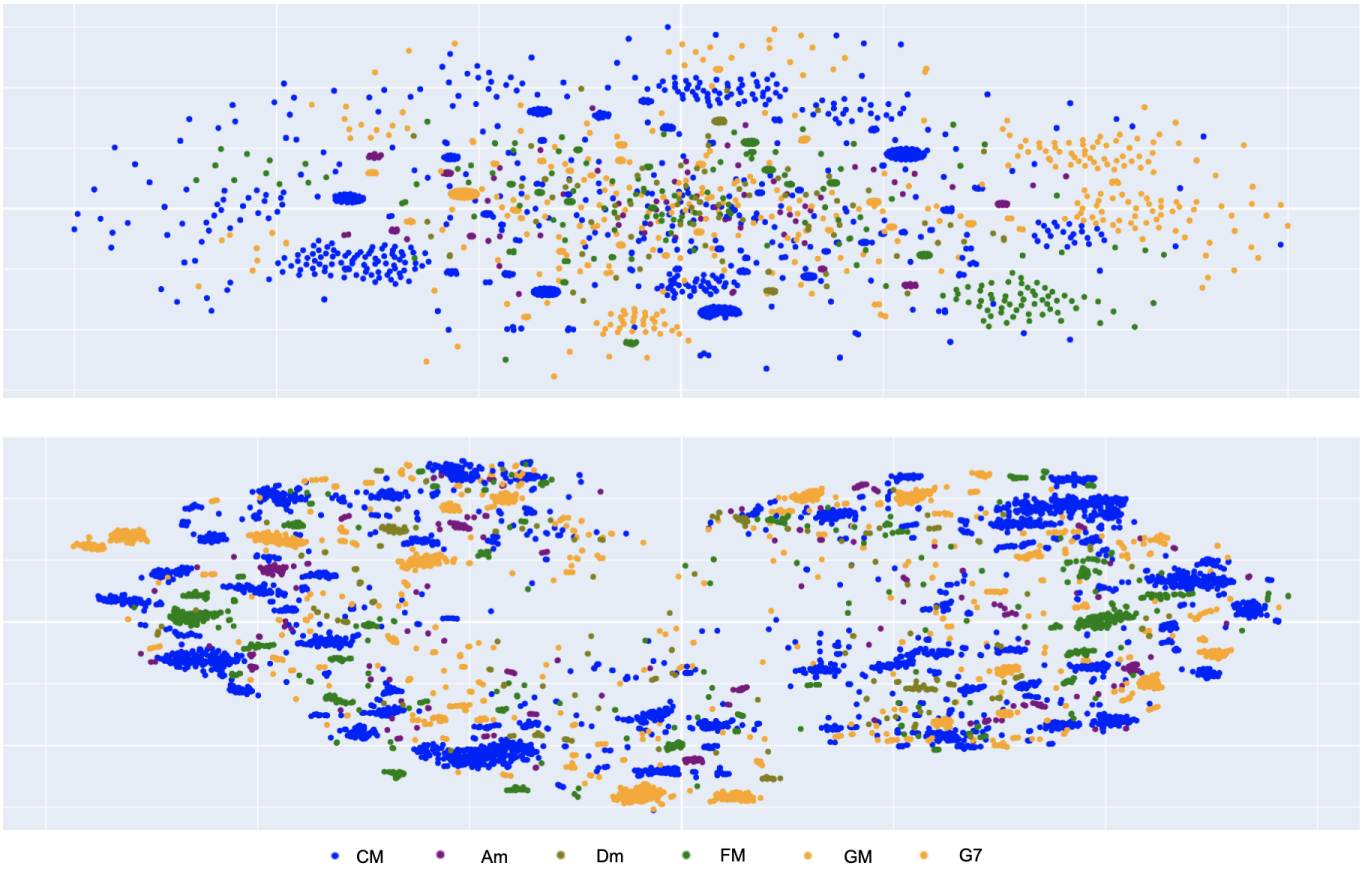

Fig. 3. 2-D representations of our model's embedding vector space and final logit vector space, respectively. Both graphs result from t-SNE reduction. That is, both are visual proxies that enable us to make inferences about the model's clustering tendencies in high-dimensional embedding spaces. Each dot represents a chord and is shaded according to that chord's pitch-class constituents, as shown in the legend.

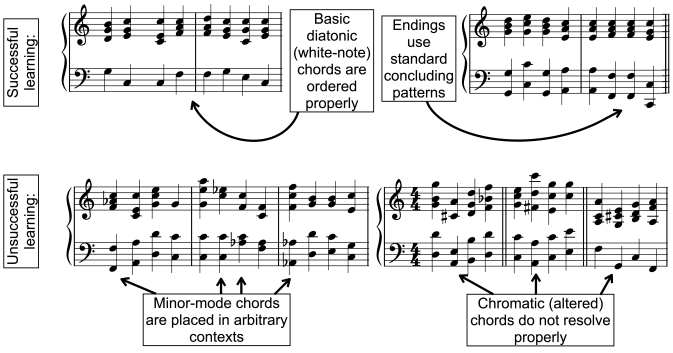

Fig. 4. a) A brief summary of the emergent behavior of the GPT when trained briefly. Diatonic ("white-note") harmony appears to be learned along with some concluding gestures; usage of chromatic and minor-mode ("black-note") harmony is less successful

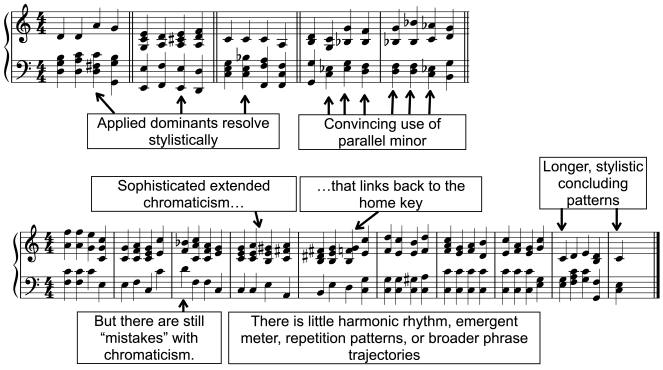

Fig. 5. A brief summary of the emergent behavior of the GPT when trained for 12 hours. While some subtle aspects of harmonic syntax are not absent, many chromatic and minor-mode gestures are learned, as are more sophisticated ending structures.

harmonic successions—growing more certain when encountering frequent and predictable events within a familiar musical corpus [64].

Grammar, Voiceleading, and Harmonic Function. Both the graphing and output analyses of the model reveal its ability to capture fundamental aspects of the chord syntax within this musical style, and Figures 4 and 5's outputs demonstrate the model's proficiency in expressing step-wise voiceleading characteristic of this style. Notably, the model appears to have internalized the various chord types that commonly appear

at different points in a musical phrase, particularly evident in its production of convincing and stylistic ending patterns. This suggests the model's broader awareness of grammatical categorization.

The graphical analyses further indicate that the model creates similar encoding vectors for chords with the same grammatical roles— the same harmonic function. This consistent encoding highlights the model's capacity to identify and differentiate chords based on their shared roles within the musical contexts.

What fails to be learned with no supervision. Despite the transformer's ability to learn aspects of voiceleading and chord grammar without supervision or reinforcement, certain crucial components of this musical style remain absent in the model's embeddings and outputs. Notably, both meter and phrase structure, which serve as significant organizing principles in this style, are not expressed. Musical phrases in this style typically exhibit internal repetitions and concluding gestures in predictable locations, often every 4 or 8 measures [65, 66]. However, the model's outputs show no evidence of this type of organizational structure. Furthermore, even in the absence of rhythm, chord progressions generally convey some aspect of a passage's meter. This may involve the repetition of harmonic patterns or the alternation of stable and unstable harmonies [67, 32]. Again, the outputs exhibited no indication of such patterning.

Our model's limitations may stem from factors such as the relatively small corpus size and the minimal length of training. In contrast to the vast availability of written texts for training transformers in natural language processing, symbolic musical corpora of hymns are significantly smaller and less diverse. Moreover, music presents unique challenges for unsupervised learning. While spoken language can often be learned through passive exposure, producing convincing musical expressions typically requires explicit tutelage [32]. The gap between music listeners and music producers may similarly apply to the limits of our model's outputs, reflecting the need for expert-guided musical learning. For instance, learning concepts like meter and phrasing may rely on bodily engagement with music [68, 69] or be facilitated by the particular biological constraints of the human brain [70]. These factors might make learning these concepts solely through unsupervised machine learning exceedingly challenging, if not impossible.

VI. Conclusion and Future Directions

This paper has demonstrated the potential of a GPT-style transformer to learn musicologically significant behaviors through unsupervised training on a corpus of chorale-style hymns. The model successfully acquired basic chord grammar, harmonic function, and the accompanying voiceleading behaviors characteristic of this musical style. However, it also highlighted limitations in the learning process, as the model's outputs lacked metrical or phrase structure.

Future studies should aim to address these limitations by expanding the size of the corpus and experimenting with longer training periods. Introducing semi-supervised components to the model could provide insights into the relative efficacy of unsupervised versus supervised machine learning for musical generative modeling using Transformers.

Additionally, connecting such models to human behavioral and cognitive theory holds promise. Designing tests to directly compare similarities and differences between human and machine learning of musical material could offer valuable insights into the learning processes involved in music generation.

Acknowledgment

ChatGPT was used in the editing of this document. All ideas, analyses, and the first and final draft were created and crafted by the authors.

References

[1] Jean-Pierre Briot. "From artificial neural networks to deep learning for music generation: history, concepts and trends". In: *Neural Computing and Applications* 33.1 (2021), pp. 39–65.

[2] Vasanth Kalingeri and Srikanth Grandhe. "Music generation with deep learning". In: *arXiv preprint arXiv:1612.04928* (2016).

[3] Allen Huang and Raymond Wu. "Deep learning for music". In: *arXiv preprint arXiv:1606.04930* (2016).

[4] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. "Audio Chord Recognition with Recurrent Neural Networks." In: *ISMIR*. Curitiba. 2013, pp. 335–340.

[5] Tsung-Ping Chen, Li Su, et al. "Functional Harmony Recognition of Symbolic Music Data with Multi-task Recurrent Neural Networks." In: *ISMIR*. 2018, pp. 90–97.

[6] David Sears, Filip Korzeniowski, and Gerhard Widmer. "Evaluating language models of tonal harmony". In: (2018).

[7] Safaa Allamy and Alessandro Lameiras Koerich. "1D CNN architectures for music genre classification". In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021, pp. 01–07.

[8] TL Li, Antoni B Chan, and AH Chun. "Automatic musical pattern feature extraction using convolutional neural network". In: *Genre* 10.2010 (2010), p. 1x1.

[9] Jordi Pons et al. "Timbre analysis of music audio signals with convolutional neural networks". In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 2744–2748.

[10] Keunwoo Choi, George Fazekas, and Mark Sandler. "Explaining deep convolutional neural networks on music classification". In: *arXiv preprint arXiv:1607.02444* (2016).

[11] Adam Roberts et al. "A hierarchical latent vector model for learning long-term structure in music". In: *International conference on machine learning*. PMLR. 2018, pp. 4364–4373.

[12] Cheng-Zhi Anna Huang et al. "Music transformer". In: *arXiv preprint arXiv:1809.04281* (2018).

[13] Kyoyun Choi et al. “Chord conditioned melody generation with transformer based decoders”. In: *IEEE Access* 9 (2021), pp. 42071–42080.

[14] Shuyu Li and Yunsick Sung. “MelodyDiffusion: Chord-Conditioned Melody Generation Using a Transformer-Based Diffusion Model”. In: *Mathematics* 11.8 (2023), p. 1915.

[15] Shuyu Li and Yunsick Sung. “Transformer-Based Seq2Seq Model for Chord Progression Generation”. In: *Mathematics* 11.5 (2023), p. 1111.

[16] Thomas Nuttall, Behzad Haki, and Sergi Jorda. “Transformer Neural Networks for Automated Rhythm Generation”. In: *NIME 2021*. https://nime.pubpub.org/pub/8947fhly. Apr. 2021.

[17] Prateek Verma and Chris Chafe. *A Generative Model for Raw Audio Using Transformer Architectures*. 2021. arXiv: 2106.16036 [cs.SD].

[18] Tsung-Ping Chen and Li Su. “Harmony Transformer: Incorporating chord segmentation into harmony recognition”. In: *Neural Netw* 12 (2019), p. 15.

[19] Jonggwon Park et al. “A bi-directional transformer for musical chord recognition”. In: *arXiv preprint arXiv:1907.02698* (2019).

[20] Tsung-Ping Chen and Li Su. “Attend to chords: Improving harmonic analysis of symbolic music using transformer-based models”. In: *Transactions of the International Society for Music Information Retrieval* 4.1 (2021).

[21] Mikaela Keller, Gabriel Loiseau, and Louis Bigo. “What musical knowledge does self-attention learn?” In: *Workshop on NLP for Music and Spoken Audio (NLP4MuSA 2021)*. 2021.

[22] Steven G. Laitz and Michael R. Callahan. *The Complete Musician: An Integrated Approach to Theory, Analysis, and Listening, Fifth Edition*. Oxford, 2023.

[23] Edward Aldwell, Carl Schachter, and Allen Cadwallader. *Harmony and Voice Leading, Fifth Edition*. Cengage, 2019.

[24] David Cope. “Experiments in Music Intelligence”. In: *Proceedings of the International Computer Music Conference* (1987), pp. 170–173.

[25] Ian Quinn. “Are Pitch-Class Profiles Really ‘Key for Key’?.” In: *Zeitschrift der Gesellschaft für Musiktheorie* 2 (7 2010), pp. 151–163. DOI: https://doi.org/10.31751/513.

[26] Martin Rohrmeier. “Towards a generative syntax of tonal harmony”. In: *Journal of Mathematics and Music* 5.1 (2011), pp. 35–53. DOI: 10.1080/17459737.2011.573676. eprint: https://doi.org/10.1080/17459737.2011.573676. URL: https://doi.org/10.1080/17459737.2011.573676.

[27] Ian Quinn and Panayotis Mavromatis. “Voice-Leading Prototypes and Harmonic Function in Two Chorale Corpora”. In: *Mathematics and Computation in Music*. Ed. by Carlos Agon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 230–240. ISBN: 978-3-642-21590-2.

[28] Anna Huang et al. “Bach Doodle: Approachable music composition with machine learning at scale”. In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*. 2019. URL: https://arxiv.org/abs/1907.06637.

[29] Hugo Riemann. *Vereinfachte Harmonielehre, oder die Lehre von den tonalen Funktionen der Akkorde*. London: Augener, 1893.

[30] Daniel Harrison. *Harmonic Function in Chromatic Music: A Renewed Dualist Theory and an Account of Its Precedents*. Chicago: University of Chicago Press, 1994.

[31] Fred Lerdahl and Ray Jackendoff. *A generative theory of tonal music*. Cambridge. MA: The MIT Press, 1983. ISBN: 0262120941.

[32] Christopher Wm. White. *The Music in the Data: Music Analysis, Corpus Analysis, and Tonal Traditions*. Routledge, 2022.

[33] Christopher WM White and Ian Quinn. “Chord Context and Harmonic Function in Tonal Music”. In: *Music Theory Spectrum* 40.2 (Nov. 2018), 314–335O. ISSN: 0195-6167. DOI: 10.1093/mts/mty021. eprint: https://academic.oup.com/mts/article-pdf/40/2/314/26446308/mty021.pdf. URL: https://doi.org/10.1093/mts/mty021.

[34] Jenine Brown, Daphne Tan, and David John Baker. “The Perceptual Attraction of Pre-Dominant Chords”. In: *Music Perception* 39.1 (Sept. 2021), pp. 21–40.

[35] Anton Ayzenberg et al. “Chordal Embeddings Based on Topology of the Tonal Space”. In: *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer. 2023, pp. 20–33.

[36] Panayotis Mavromatis. “HMM Analysis of Musical Structure: Identification of Latent Variables Through Topology-Sensitive Model Selection”. In: *Mathematics and Computation in Music*. Ed. by Elaine Chew, Adrian Childs, and Ching-Hua Chuan. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 205–217.

[37] Hiroaki Tsushima et al. “Generative Statistical Models with Self-Emergent Grammar of Chord Sequences”. In: *CoRR* abs/1708.02255 (2017). arXiv: 1708.02255. URL: http://arxiv.org/abs/1708.02255.

[38] Nori Jacoby, Naftali Tishby, and Dmitri Tymoczko. “An Information Theoretic Approach to Chord Categorization and Functional Harmony”. In: *Journal of New Music Research* 44.3 (2015), pp. 219–244. DOI: 10.1080/09298215.2015.1036888.

[39] Jason Yust, Jaeseong Lee, and Eugene Pinsky. “A Clustering-Based Approach to Automatic Harmonic Analysis: An Exploratory Study of Harmony and Form in Mozart’s Piano Sonatas”. In: *Transactions of the International Society for Music Information Retrieval* 5 (Oct. 2022), pp. 113–128. DOI: 10.5334/tismir.114.

[40] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[41] Botao Yu et al. “Museformer: Transformer with fine- and coarse-grained attention for music generation”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1376–1388.

[42] Yu-Siang Huang and Yi-Hsuan Yang. “Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions”. In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 1180–1188.

[43] Jeff Ens and Philippe Pasquier. “Mmm: Exploring conditional multi-track music generation with the transformer”. In: *arXiv preprint arXiv:2008.06048* (2020).

[44] Yash Khasgiwala and Jash Tailor. “Vision transformer for music genre classification using mel-frequency cepstrum coefficient”. In: *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE. 2021, pp. 1–5.

[45] Minz Won, Keunwoo Choi, and Xavier Serra. “Semi-supervised music tagging transformer”. In: *arXiv preprint arXiv:2111.13457* (2021).

[46] Wei-Tsung Lu et al. “SpecTNT: A time-frequency transformer for music audio”. In: *arXiv preprint arXiv:2110.09127* (2021).

[47] Yixiao Zhang et al. “Spectrogram Transformers for Audio Classification”. In: *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE. 2022, pp. 1–6.

[48] Shih-Lun Wu and Yi-Hsuan Yang. “MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 1953–1967.

[49] Kristy Choi et al. “Encoding musical style with transformer autoencoders”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1899–1908.

[50] Halley Young et al. “Compositional Steering of Music Transformers”. In: *Proc. 3rd IUI Workshop on Human-AI Co-Creation with Generative Models*. 2022.

[51] Pedro Ferreira, Ricardo Limongi, and Luiz Paulo Fávero. “Generating Music with Data: Application of Deep Learning Models for Symbolic Music Composition”. In: *Applied Sciences* 13.7 (2023), p. 4543.

[52] Jingjing Tang, Geraint Wiggins, and George Fazekas. “Reconstructing Human Expressiveness in Piano Performances with a Transformer Network”. In: *arXiv preprint arXiv:2306.06040* (2023).

[53] Pengfei Zhu et al. “ERNIE-Music: Text-to-Waveform Music Generation with Diffusion Models”. In: *arXiv preprint arXiv:2302.04456* (2023).

[54] Andrea Agostinelli et al. “Musiclm: Generating music from text”. In: *arXiv preprint arXiv:2301.11325* (2023).

[55] Shulei Ji and Xinyu Yang. “EmoMusicTV: Emotion-conditioned Symbolic Music Generation with Hierarchical Transformer VAE”. In: *IEEE Transactions on Multimedia* (2023).

[56] Naomi Imasato et al. “Using a Language Model to Generate Music in its Symbolic Domain while Controlling its Perceived Emotion”. In: *IEEE Access* (2023).

[57] Maral Ebrahimzadeh, Valerie Krug, and Sebastian Stober. “Transformer-Based Chord Recognition with Unsupervised Pre-training of Input Embeddings”. In: ().

[58] Michael Scott Cuthbert and Christopher Ariza. “Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data.” In: *ISMIR*. Ed. by J. Stephen Downie and Remco C. Veltkamp. International Society for Music Information Retrieval, 2010, pp. 637–642. ISBN: 978-90-393-53813. URL: http://dblp.uni-trier.de/db/conf/ismir/ismir2010.html#CuthbertA10.

[59] Daniel Harrison. *Pieces of Tradition: An Analysis of Contemporary Tonal Music*. Oxford, 2016.

[60] Liam Hynes-Tawa. “How the Phrygian Final Lost Its Finality”. PhD thesis. Yale University, 2020.

[61] Karpathy. *GitHub - karpathy/nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs.* en. URL: https://github.com/karpathy/nanoGPT.

[62] David R. W. Sears, J. E. Verbeten, and H. M. Percival. “Does order matter? Harmonic priming effects for scrambled tonal chord sequences”. In: *Journal of Experimental Psychology: Human Perception and Performance* 49 (7 2023), pp. 999–1015.

[63] Dominique T. Vuvan and Bryn Hughes. “Probe Tone Paradigm Reveals Less Differentiated Tonal Hierarchy in Rock Music”. In: *Music Perception* 38.5 (2021), pp. 425–434.

[64] Emily Schwitzgebel and Christopher Wm. White. “Effects of Chord Inversion and Bass Patterns on Harmonic Expectancy in Musicians”. In: *Music Perception* 39.1 (Sept. 2021), pp. 41–62.

[65] Robert Gjerdingen. *Music in the Galant Style*. Oxford, 2007.

[66] William Caplin. *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven*. Oxford, 1998.

[67] Richard Cohn. “Meter”. In: *Oxford Handbook of Critical Concepts in Music Theory*. Oxford, 2015.

[68] Mariusz Kozak. *Enacting Musical Time: The Bodily Experience of New Music*. Oxford, 2019.

[69] Justin London. *Hearing in Time: Psychological Aspects of Musical Meter 2nd Edition*. Oxford, 2012.

[70] Edward Large. “Modeling beat perception with a nonlinear oscillator”. In: *Proceedings of the eighteenth annual conference of the cognitive science society*. Routledge, 2019, pp. 420–425.