Learning, probability and logic : towards a unified approach for content-based Music Information Retrieval

Helene-Camille Crayencour, Carmine Emanuele Cella

frontiers in Digital Humanities

METHODS
published: 16 April 2019
doi: 10.3389/fdigh.2019.00006

Learning, Probability and Logic: Toward a Unified Approach for Content-Based Music Information Retrieval

Helene-Camille Crayencour[1*†] *and Carmine-Emanuele Cella*[2*†]

[1] *Centre National de la Recherche Scientifique, Paris, France,* [2] *Center for New Music and Audio Technology, University of California, Berkeley, Berkeley, CA, United States*

Within the last 15 years, the field of Music Information Retrieval (MIR) has made tremendous progress in the development of algorithms for organizing and analyzing the ever-increasing large and varied amount of music and music-related data available digitally. However, the development of content-based methods to enable or ameliorate multimedia retrieval still remains a central challenge. In this perspective paper, we critically look at the problem of automatic chord estimation from audio recordings as a case study of content-based algorithms, and point out several bottlenecks in current approaches: expressiveness and flexibility are obtained to the expense of robustness and vice versa; available multimodal sources of information are little exploited; modeling multi-faceted and strongly interrelated musical information is limited with current architectures; models are typically restricted to short-term analysis that does not account for the hierarchical temporal structure of musical signals. Dealing with music data requires the ability to tackle both uncertainty and complex relational structure at multiple levels of representation. Traditional approaches have generally treated these two aspects separately, probability and learning being the usual way to represent uncertainty in knowledge, while logical representation being the usual way to represent knowledge and complex relational information. We advocate that the identified hurdles of current approaches could be overcome by recent developments in the area of Statistical Relational Artificial Intelligence (StarAI) that unifies probability, logic and (deep) learning. We show that existing approaches used in MIR find powerful extensions and unifications in StarAI, and we explain why we think it is time to consider the new perspectives offered by this promising research field.

Keywords: music information retrieval (MIR), content-based, chord recognition, statistical relational artificial intelligence, audio

OPEN ACCESS

Edited by:
Mark Brian Sandler,
Queen Mary University of London,
United Kingdom

Reviewed by:
Alberto Pinto,
Centro Europeo per Gli Studi in Musica e Acustica (CESMA),
Switzerland
Emmanouil Benetos,
Queen Mary University of London,
United Kingdom

***Correspondence:**
Helene-Camille Crayencour
helene.camille.crayencour@gmail.com
Carmine-Emanuele Cella
carmine.cella@berkeley.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Digital Musicology, a section of the journal Frontiers in Digital Humanities

Received: *12 June 2018*
Accepted: *14 March 2019*
Published: *16 April 2019*

Citation:
Crayencour H-C and Cella C-E (2019) Learning, Probability and Logic: Toward a Unified Approach for Content-Based Music Information Retrieval. Front. Digit. Humanit. 6:6. doi: 10.3389/fdigh.2019.00006

1. INTRODUCTION

Understanding music has been a long-standing problem for very diverse communities. Trying to formalize musical knowledge, and to understand how human beings create and listen to music has proven to be very challenging given the huge amount of levels involved, ranging from hearing to perception, from acoustics to music theory. In the past 30 years, an impressive amount of research

work in different fields related to music has been done in the aim of clarifying the relations between these levels and in order to find good representations for musical knowledge in different forms such as scores, intermediate graphic representations and so on.

1.1. A Brief History of MIR

The development of computer hardware technology and the advancements in machine learning techniques have fostered the development of Artificial Intelligence (AI) techniques for musical research in a number of directions such as musicology, digital sound processing, composition, and performance. In the meantime, considerable advances in compression storage and dissemination technologies for digital signals have favored the emergence of huge online music collections, as well as a growing users request of listening to music in a personalized way, creating the need to develop advanced techniques to organize and analyze the ever-increasing large and varied amounts of music and music-related data available digitally. Different music research communities have pursued efforts toward the same goal of modeling human analysis of music and getting insight into the intellectual process of music under various names such as Machine Listening (Malkin, 2006), Music Artificial Intelligence (Dobrian, 1993), Intelligent Audio Analysis (Schuller, 2013), or Music Informatics (Humphrey et al., 2013).

As such, music research communities have converged to the creation of the emerging field of *Music Information Retrieval* (MIR). In a few years, with the improvement of computers and the advancements in machine learning techniques, the MIR field has brought impressive results in many directions and has been able to address problems that appeared to be unsolvable only 20 years ago, such as cover song identification (Serrà, 2011), structure analysis (Paulus et al., 2010), or automatic orchestration (Maresz, 2013; Pachet, 2016). It has been possible to develop technologies that allow users to understand, access, and explore music in all its different dimensions, from browsing personal collections, to managing the rights of music creators or to answering new musicological questions, at a level of abstraction and a scale that were previously not possible without the help of AI. For instance, large-scale exploration of music-related data using MIR techniques has made possible to bring new insight to musicological questions that were previously studied only at small scale, such as exploring and visualizing the harmonic structure of Richard Wagners *Ring des Nibelungen* (Zalkow et al., 2017).

One of the most important streams of research in the MIR field, and of particular interest to this paper, is the development of content-based methods to enable or improve multimedia retrieval, in particular from audio signals (see Foote, 1997; Lew, 2006; Casey et al., 2008; Muller et al., 2011; Orio et al., 2011; Grosche et al., 2012; Schedl et al., 2014 for comprehensive reviews of this aspect). This subfield of MIR research concerned with the automatic extraction of relevant content information from music data, especially from audio signals. It involves the development of algorithms for solving tasks at various levels of abstraction, ranging from low-level signal processing and signal-centered feature extraction (such as pitch onset detection Bello et al., 2005, or chroma extraction Bartsch and Wakefield, 2001), mid-level information extraction and music content description and indexing (such as automatic beat tracking Ellis, 2007a, melody Salamon et al., 2013 or chord progression estimation McVicar et al., 2014), to high-level user-centered and contextual level of abstraction (such as music recommendation Schedl et al., 2017 or music visualization Herremans and Chuan, 2017).

1.2. Progress in the MIR Field

After a first decade of constant progress, using techniques that commonly adopt a two-stage process of feature extraction followed by semantic interpretation of audio and scores, the MIR field experienced a first ceiling in most of its research topics (Aucouturier and Pachet, 2004). The community then turned massively to deep learning techniques, with the hope to overcome several obstacles, such as the sub-optimal use of handcrafted features, the limited power of shallow architectures, and the limitation to short-time analysis that cannot encode musically meaningful structure (Humphrey et al., 2013). The introduction of deep learning in the MIR community, hence, enforced the switch to large-scale problems and promoted the development of algorithms that are able to address more general problems for acoustic signals. State-of-the-art results in most MIR tasks are obtained with neural networks.

Nonetheless today it feels like the progress in many if not all research areas have reached a *plateau* again. Despite the positive results, these complex programmable machines brought us to a very difficult and partially unknown mathematical world. Lot of research, today, is in the difficult position of providing empirical results without being able to explain the theoretical motivations behind them. It is not uncommon to see papers about new classification methods that outperform previous research but does not explain why a specific architecture has been chosen. Often, indeed, deep learning architectures are based on previous architectures and are just augmented with more layers, more samples or more complex units without a real explanation of the scientific intuitions that support these changes. It is our belief that despite the important results obtained, entering a research pattern that is only motivated by empirical results can be a problem for understanding music.

Today, many tasks that are relatively easy for humans (such as following a speakers voice in a crowded and noisy place or detecting harmonic changes in a cover song) are still hard to tackle. In this perspective article, we critically look at the problem of automatic chord estimation from audio recordings as a case study of content-based algorithms, and point out several bottlenecks in current approaches: expressiveness and flexibility are obtained to the expense of robustness and vice versa; available multimodal sources of information are little exploited; modeling multi-faceted and strongly interrelated musical information is limited with current architectures; models are typically restricted to short-term analysis that does not account for the hierarchical temporal structure of musical signals; simplified versions of MIR problems cannot be generalized to real problems. Starting from these observations, the rest of this paper is an attempt to propose possible directions for music processing, looking at

recent advances in other fields that have been confronted to similar problems.

1.3. The Need to Integrate Diverse Directions Into a Common Perspective

In its efforts of designing intelligent systems that are able to answer complex queries, the music community has faced the need of dealing with uncertain reasoning over complex, multiple, relational objects. Current approaches are unable to tackle all aspects together and generally focus on either one of them. We believe that it is necessary to define a general framework able to integrate different approaches for the representation of musical and acoustical signals into a common perspective.

Dealing with real music data requires the ability to handle both uncertainty and complex relational structure at multiple levels of representation. Established approaches have generally treated these two aspects separately, probability and learning being the usual way to represent uncertainty in knowledge, while logical representation being the usual way to represent knowledge and complex relational information. In our opinion, the multiplicity of levels that musical knowledge exhibits can only be captured and characterized by a mixture of approaches that takes into consideration learning, statistical and logical standpoints of the problem. This view is shared by an emerging field in artificial intelligence and machine learning called *Statistical Relational Artificial Intelligence* (StarAI) (De Raedt et al., 2016), also referred to as *statistical relational learning* (SRL) (see Getoor and Taskar, 2007 for original definition).

1.4. StarAI: Unifying Logic, Learning and Probabilities

We advocate that the identified hurdles of current approaches could be overcome by recent developments in the area of Statistical Relational Artificial Intelligence that unifies probability, logic and (deep) learning.

Most real world applications of artificial intelligence algorithms deal with data that have both inherent uncertainty and are characterized by complex relational structure. The rich relational structure of the data of interest appears either at the internal level (e.g., complex relationships of notes composing a chord in the case of a chord estimation algorithm, or at the external level (e.g., relationships between chords in a piece of music). As recalled by Russell in a recent paper (Russell, 2014), classical AI adopted first-order logic as a formal language to describe the real world and "*things in it*," and to allow reasoning over explicitly represented knowledge. Modern AI addressed the problem of inherent uncertainty of real world data. Uncertainty arises from noise and incomplete information in the data (e.g., occlusions and misspellings in text, percussive sounds blurring the harmonic content of music signals). But also from many other aspects such as the data type, the uncertainty about the existence and the number of objects of interest (e.g., how many instruments are playing together?). The language of probability theory adopted by modern AI allows dealing with uncertain reasoning. Recent years have seen a massive regain of attention for neural networks (Minsky and Papert, 1969), especially a tremendous interest for deep learning (Bengio, 2016), enabling learning complex abstract representations from the input data through the combination of hierarchical simpler parts.

Logic is able to handle the complexity of the real world, and is capable of reasoning with large numbers of interacting heterogeneous objects, but it cannot deal with its uncertainty. Probabilistic graphical models are a powerful framework for dealing with uncertainty, but they cannot handle real-world complexity. Deep learning is able to create complex abstract representations from large-scale raw data but its mechanism for structure learning remains to be understood.

In recent years, there has been a considerable body of research in combining between knowledge representation, learning and reasoning, with new impulse coming from the area of deep learning, evolving into the new research field Statistical Relational Artificial Intelligence (StarAI). They combine first order logic, relational representations, and logical inference, with concepts of probability theory and machine learning. Ideas from probability theory and statistics are used to address uncertainty while tools from logic, databases, and programming languages are introduced to represent structure. Relational and logical abstraction allows one to specify and reason about regularities across different situations using rules and templates rather than having to specify them for each single entity separately. It becomes possible to make abstractions that apply to all individuals that have some common properties. Knowledge can be incorporated in a declarative, expressive, and compact way.

StarAI has raised promising and exciting new perspectives in many fields of science, humanities, and technology, and approaches developed in this field have been successfully applied in various domains and used for many tasks in artificial intelligence, such as natural language processing (Riedel and Meza-Ruiz, 2008), event extraction (Venugopal, 2015), bioinformatics (Mallory et al., 2016), collective classification (Crane and McDowell, 2012), activity recognition (Sztyler et al., 2018), entity resolution (Singla and Domingos, 2006; Pawar et al., 2017), machine reading (Poon, 2011), semantic image interpretation (Donadello et al., 2017b) or language modeling (Jernite et al., 2015) to name a few. In this paper, we explain that existing approaches used in the MIR field find powerful extensions and unifications in StarAI, and provide arguments to support why we think it is time to consider the new perspectives offered by this area.

1.5. Paper Organization

The rest of the paper is organized as follows: section 2 critically reviews existing content-based MIR approaches, focusing on the task of automatic chord estimation as a case study, and identifies four major deficiencies of computational analysis models: the inability to handle both uncertainty and rich relational structure; the incapacity to handle multiple abstraction levels and the incapability to act on multiple time scales; the unemployment of available multimodal information, and the ineptitude to generalize simplified problems to complex tasks. section 3 discusses the need of an integrated research framework and presents the perspectives offered by statistical relational

AI models for music processing. In section 4, our conclusion encourages an exploration of this promising research field.

This transversal paper covers concepts from many area, in particular probability theory, logic and deep learning. We have provided, when it seemed necessary, some background knowledge and vocabulary in the form of footnotes, in order not to disturb the reading of the article.

2. MIR: LIMITATIONS, CHALLENGES, AND FUTURE DIRECTIONS

In this section, we critically review common approaches to content-based analysis and we point out four major limitations of existing methods. To illustrate our exploration on the state-of-the-art of content-based MIR, we mainly focus on the emblematic task of automatic chord estimation (ACE), but our conclusions can generalize to other content-based MIR tasks. Among the various subtopics in content-based MIR, automatic chord estimation has received a considerable and sustained attention since its inception (Fujishima, 1999). Not only it is a very interesting and challenging problem, but also it has many applications in high-level tasks such as cover song identification (Marsík et al., 2017), genre classification (Pereira and Silla, 2017) or automatic accompaniment (Ojima et al., 2017). Also, there is a high demand on chord-based representations by users in the music community, as illustrated by the popularity of websites such as *Ultimate Guitar*[1]. This task constitutes a good example to understand the potential shortcomings of current MIR methodologies for content-based retrieval. Moreover, a clear understanding of the shortcomings of ACE systems prepares the reader to the proof-of-concept case study on chord recognition using concepts from StarAI that we will present in section 3.3.

In the literature, different problems are referred to under the same name of automatic chord estimation (Humphrey and Bello, 2015). In this article, chord estimation is understood as chord recognition, where the task is to label each frame of an audio signal with a chord from a given dictionary. It is to be distinguished from more abstract tasks, such as analyzing structural functions in harmony (Schoenberg, 1969), that integrate high-level musical concepts such as key or musical structure.

As most music signal processing systems, the majority of computational models for chord estimation from audio share a common two-step architecture consisting in a feature extraction step (traditionally handcrafted low-to-mid level features such as chroma, now deep-learned features) followed by a classification step (Cho and Bello, 2014). The features extraction step outputs a sequence of temporally ordered features that are often related to the metrical structure, such as beats-synchronous features, and predicts chord label distributions for each time frame. The classification step can consist in a simple point-wise prediction (Papadopoulos and Peeters, 2007; Korzeniowski and Widmer, 2016a), but also in a more a sophisticated dynamics model [either a probabilistic model such as a hidden Markov model (HMM) (Sheh and Ellis, 2003) or a dynamic Bayesian network (DBN) (Mauch and Dixon, 2010), or a recurrent neural network (RNN) (Boulanger-Lewandowski et al., 2013; Sigtia et al., 2015)] that will provide temporal smoothing and the possibility to add some context. We refer the reader to Cho and Bello (2014), McVicar et al. (2014) for comprehensive reviews of traditional handcrafted features and probabilistic graphical models approaches, and to Deng and Kwok (2016) and McFee and Bello (2017) for reviews of more recent approaches using deep-learning.

2.1. Designing Robust, Flexible and Expressive Models Simultaneously

Real data such as music signals exhibit both uncertainty and rich relational structure at multiple levels of representation. In general, signal observations are incomplete and noisy due to the great variability of audio signals, background interferences, presence of transient sounds, etc. Moreover, musical entities such as chords, keys, melody, present in general complex relations among them. Standard approaches to music signal processing are generally not able to cope at once with both aspects. Approaches that handle uncertainty generally have insufficient expressive power to tackle complex domains and vice versa, robustness is obtained at the expense of flexibility and expressiveness. The literature offers two main different perspectives. Uncertainty is well handled by probabilistic graphical models and complexity is generally tackled by logical approaches, which is the standard distinction between "modern" and "traditional" AI. Both perspectives are not in conflict with each other but on the contrary synergistic.

2.1.1. Probabilistic Graphical Models: Handling Noise, Uncertainty, and Incomplete Information

Probabilistic models have the general ability to handle noise, uncertainty, and incomplete information. Graphical models, in particular Bayesian networks (Pearl, 1988) and Markov networks (Murphy, 2012) are efficient and elegant frameworks for representing and reasoning with probabilistic models.

Probabilistic graphical models (PGM) allow compactly representing the manner in which the variables depend on each other. In a graphical model, each node represents a random variable, and edges represent probabilistic influence between nodes. A graphical model is a family of probability distributions over the variables that factorize according to an underlying graph. Probabilistic graphical models exploit probabilistic independencies in the data. The absence of an edge between two variables means that they are *conditionally independent* given all other random variables in the mode. This allows decomposing complex probability distributions into a product of independent factors, and to represent a distribution over a large number of random variables by a product of *potential functions* that each depend on only a smaller subset of variables (see the example of **Figure 2**, right). Graphical models include many model families. The way the original probability distribution is factorized allows distinguishing between *directed* and *undirected* graphical models. Since many concepts of the theory of graphical models have been developed independently in different areas, they may thus have different names. Directed

[1]http://www.ultimate-guitar.com

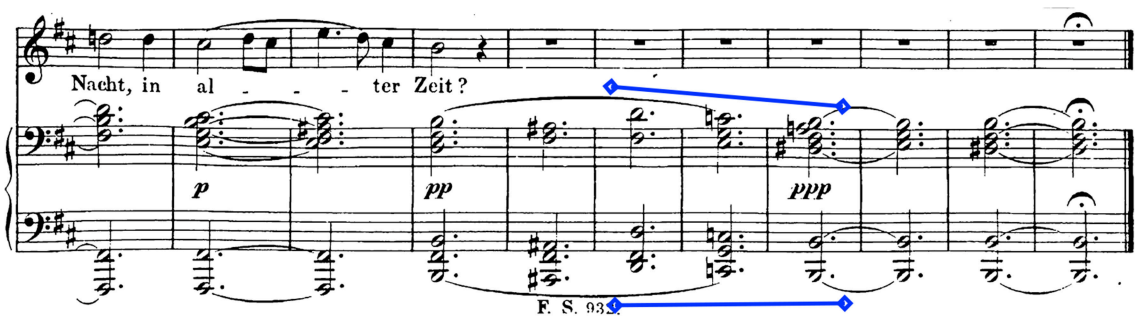

FIGURE 1 | Extract of Schubert's lied "Der Doppelgänger" (Henrich Heine). Chords in a chord progression are not independent from each other but are linked according to musical rules or musical ideas. An example of such musical idea is shown in the work of Mishkin (1978), who analyzes that Schubert employs in some lieder of his last year harmonic parallelism between triads a half step apart when the poetic images evoke the supernatural. Mishkin provides an example with "Der Doppelgänger" where *"the vocal line is supported by hollow sonorities in the piano accompaniment to suggest the dreamlike horror of beholding one's own double weeping before the house of the departed beloved. The hallucinatory impression is intensified, in the concluding passage for piano alone (Indicated by the blue line in this figure), through an organum-like progression that moves in parallel motion through a triad on the lowered supertonic degree"*.

graphical models are also commonly known a *Bayesian networks* and undirected models are also referred to as *Markov random fields* or *Markov networks* (Kindermann and Snell, 1980).

In music processing, probabilistic graphical models (PGM) (Pearl, 1988; Murphy, 2012) have been widely used to model tasks where objects can be represented as sequential phenomena, such as in the case of chord estimation (Papadopoulos and Peeters, 2011). The extraction of music content information can be often seen as a classification problem. For instance the ACE problem can be seen as predicting a succession of chord labels $y \in Y$ given some observations over time $x \in X$ (e.g., chroma vectors). In general, the set of variables $X \cup Y$ has a complex structure. For example, chords in a chord progression are not independent from each other: they are linked according to complex musical rules or musical ideas (see the example in **Figure 1**). Observations are also linked according to the underlying states they represent (see **Figure 2**, left).

One important aspect in music is *time*: events appear at particular times and those times (that can be points or intervals) are ordered. The MIR community has thus looked into probabilistic techniques that take into account the temporal nature of music. Hidden Markov models (HMM) have been particularly popular for the task of ACE and proved to be an elegant way to address this difficult problem (Sheh and Ellis, 2003). However in most cases, these temporal models are used in a way that make them unable to cope with the rich relational structure of real chord progressions. For instance, a common choice is to use first-order HMMs in a frame-by-frame model, where each state corresponds to a chord, and successive observations are conditionally independent from previous observations given the current state. Such model are not adapted to the actual complexity of music and cannot learn higher-level knowledge about chord progressions. For instance they cannot adequately model the prominent 12-bar blues jazz chord sequence (Kernfeld, 2007). As noted in Cho and Bello (2014), in the away they are designed, such temporal models generally enforce temporal continuity of individual chords rather than providing information about chord transitions. Other formalisms that allow considering more complex dependencies between data in the model have scarcely been explored such as tree structures (Paiement et al., 2005) or dynamic Bayesian network (Mauch and Dixon, 2010). Conditional random fields (CRF) (Burgoyne et al., 2007; Korzeniowski and Widmer, 2016b; Wu and Li, 2018) have started receiving attention of the MIR community the last few years. Even when using deep learned features that are expected to perform better than handcrafted features, probabilistic graphical models remain an important post processing step to be applied on top of high-level learned features (Zhou and Lerch, 2015; Korzeniowski and Widmer, 2017; Wu and Li, 2018).

However, the use of probabilistic graphical models that allow taking into account complex dependencies between data remains limited in the MIR field. For instance, most work considering CRFs for music processing focus on linear-chain CRFs, modeling only dependency between consecutive chord labels. If more flexible CRFs language models have already been explored in other fields, such as the skip-chain CRFs (Sutton and McCallum, 2007) that allow Sutton and McCallum to model complex distant structure between words in a Natural Language Processing application, they remain scarcely studied in the MIR area.

Probabilistic models can handle the inherent uncertainty of audio. However, they fail to capture important aspects of higher-level musical relational structure and context. This aspect has been more specifically explored within the framework of logic.

2.1.2. Logic: Dealing With Complex Relational Objects

The logic framework offers a major advantage because its expressiveness allows modeling music rules in a compact and human-readable way. This allows an intuitive description of music. Knowledge, such as music theory, can be introduced to construct rules that reflect the human understanding of music (Ramirez et al., 2007). For a particular content-based MIR task such as labeling each frame of an audio signal with a chord, a typical approach is to construct low-level features

(such as a chromagram) and solve the problem using a first-order hidden Markov model. Although providing interesting results, this approach hides the rich logical structure of the underlying data that is essential for modeling more complex problems. The question that we ask is whether a given chroma observation is connected to a given chord. In a probabilistic graphical model, the importance of the connection is measured as the probability that a link exists between the two given nodes, under the assumption that edges are mutually independent and that each edge is true with the specified probability (De Raedt et al., 2007). Such queries can be easily expressed in logic by defining a predicate[2] of the type `link(chroma,chord)`. A more complex question is for instance not just labeling frames as isolated chords but also considering consistency of harmony with the global structure of a piece. We may like to label a frame not only with a chord, but also perform functional analysis (Schoenberg, 1969) by considering for instance the function of the chord inside the global structure of the piece ("G major chord on the fifth degree in the key of C major," etc.). Such relationships would be difficult to express using a graphical model, but, as we shall see, predicate logic[3] can easily be used to express such more complex possible relations.

The expressive power of logic-based approaches is usually higher than those of probabilistic models. Graphical models such as Bayesian (directed) and Markov (undirected) networks can represent many probability distributions compactly, but their expressiveness is only at the level of *propositional* logic: they assume a fixed set of variables, each of which can take on a value from a fixed range. For instance in the hidden Markov model proposed in Bello and Pickens (2005) (see also the illustration in **Figure 2**), the authors estimate the chord progression from the observation of a succession of chroma vectors (Chroma_0, . . ., Chroma_N). They consider a chord lexicon of 24 possible output labels, the 24 major and minor triads (CM, C#M, . . ., BM, Cm, . . ., Bm). The model represents a probability distribution over variables of the domain it considers only: ChordType (Chroma_0, CM), ChordType (Chroma_0, C#M), . . ., ChordType (Chroma_N, Bm), . . ., ChordType (Chroma_0, Bm), and encodes dependencies between pairs of chords of the dictionary: ChordTransition (CM, CM), . . ., ChordTransition (CM, C#M), . . ., ChordTransition (Bm, BM). However, it is not possible to formulate general probabilistic regularities among the probabilities of chord types and the rules of transitions that could be applied to arbitrary chords. Moreover, such approach is not scalable since it has a complexity in $O(n^2)$.

Moreover, during learning of probabilistic models, the examples are generally assumed to be independent of each other. This approximation does not reflect real-world data, in which instances are not truly independent. For instance, in a classical chord HMM (see **Figure 2**), atomic chroma observations are treated as independent random variables given the current state. However, neighboring chroma observation are not independent, since in general chords do not change on a frame-by-frame basis. As a result, it is likely that several consecutive chroma features will be similar and correspond to the same underlying state (see **Figure 2**). This property is not taken into account in the HMM: the model computes observation probabilities independently for each frame, and thus makes a useless effort, since we know that consecutive frames are likely to correspond to the same underlying harmony[4].

Another advantage is that logical inference of rules allows taking into account all events including those which are rare (Anglade and Dixon, 2008b). This property is particularly important when analyzing music, where a high probability of an event does not necessarily reflect his musical importance (see **Figure 3**). On the other side, a drawback of logic-based representations is that they are rigid in the sense that a single counter example is enough to make a formula false. It is thus difficult to use logic only to model empirical data, which is often noisy or uncertain, and that is difficult to describe with facts that hold universally. A paradigmatic example is given by the rules of harmony. The rules given by Rameau in his Treatise on Harmony (Rameau and Gossett, 1971) codify the principles of tonality that had governed Western music for almost two centuries and using these rules during MIR tasks have been proven very important. However, these rules are not *always* true in a logical sense and models that imply only logical truth can be mistaken by this ambiguity.

Among logic-based approaches, Inductive Logic Programming (ILP) (Muggleton, 1991) refers to logical inference techniques that are subset of First-Order Logic (FOL). These approaches combine logic programming with machine learning. They have been used to model and learn music rules, especially in the context of harmony characterization (Morales and Morales, 1995; Morales, 1997; Ramirez and Palamidessi, 2003; Anglade and Dixon, 2008a), and in the context of expressive music performance (Dovey et al., 1995; Van Baelen et al., 1997; Widmer, 2003; Ramirez et al., 2007). They have a high expressive power. For example, the authors in Anglade et al. (2009) use ILP to learn logical descriptions of harmonic sequences which characterize particular styles or genres. They are able to express complex musical rules such as `genre(genre1,A,B,Key) :- gap(A,C),degreeAndCategory(5 ,7,C,D,Key), degreeAndCategory(1 ,maj,D,E,Key),gap(E,B)` which

[2]In logic, *predicates* represent properties of objects [e.g., `lsMajor(chord)`] and relations between them [e.g., `AreHarmonicallyRelated(chord`$_1$`, chord`$_2$`)`]. *Grounding* is the process of replacing variables with constants in logical formulas [e.g., `lsMajor(CMinorChord)`]. A predicate takes as outputs either True (synonymous with 1) or False (synonymous with 0). A *world* is an assignment of a truth value (0 or 1) to each possible ground predicate. A *ground predicate* is called an *atomic formula* or an *atom*. A *positive literal* is an atomic formula and a *negative literal* is the negation of an atomic formula.

[3]*First-order* logic is also known as *predicate* logic because it uses predicates and quantifiers, as opposed to *propositional* logic that deals with simple declarative propositions and is less expressive. The adjective "first-order" distinguishes first-order logic, in which quantification is applied only to variables, from higher-order logic in which quantification can be applied to predicate symbols and function symbols. For more details, see e.g., Haack (1978) and Leivant (1994).

[4]The common choice of using beat-synchronous features instead of frame-based features allows reducing the complexity of the model. It has been lately advocated to use models at higher temporal levels than frame-wise models (Korzeniowski and Widmer, 2017; Korzeniowski et al., 2018), similarly to *language models* in speech recognition.

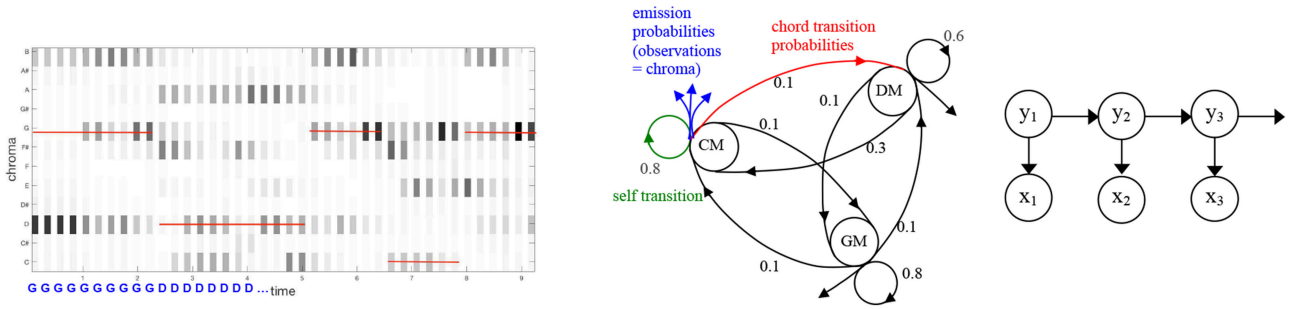

FIGURE 2 | (Left) Example of half-beat-synchronous chromagram computed in the beginning of Mozart piano sonata KV.283 and corresponding ground truth chord transcription. Chords changes on half measures, and chroma features belonging to the same half measures contain similar content. **(Middle)** In general, each chroma feature is used independently to train a HMM. **(Right)** Graphical model of a HMM describing the joint probability distribution $p(y, x)$ for a sequence of three input variables x_1, x_2, x_3 (the observations, e.g., chroma) and three output variables y_1, y_2, y_3 (the chords that are hidden states). Because of the conditional independence between variables, the models simplifies in: $p(x_1, x_2, x_3, y_1, y_2, y_3) = p(y_3|y_2) \cdot p(y_3|x_3) \cdot p(y_2|y_1) \cdot p(y_2|x_2) \cdot p(y_1) \cdot p(y_1|x_1)$.

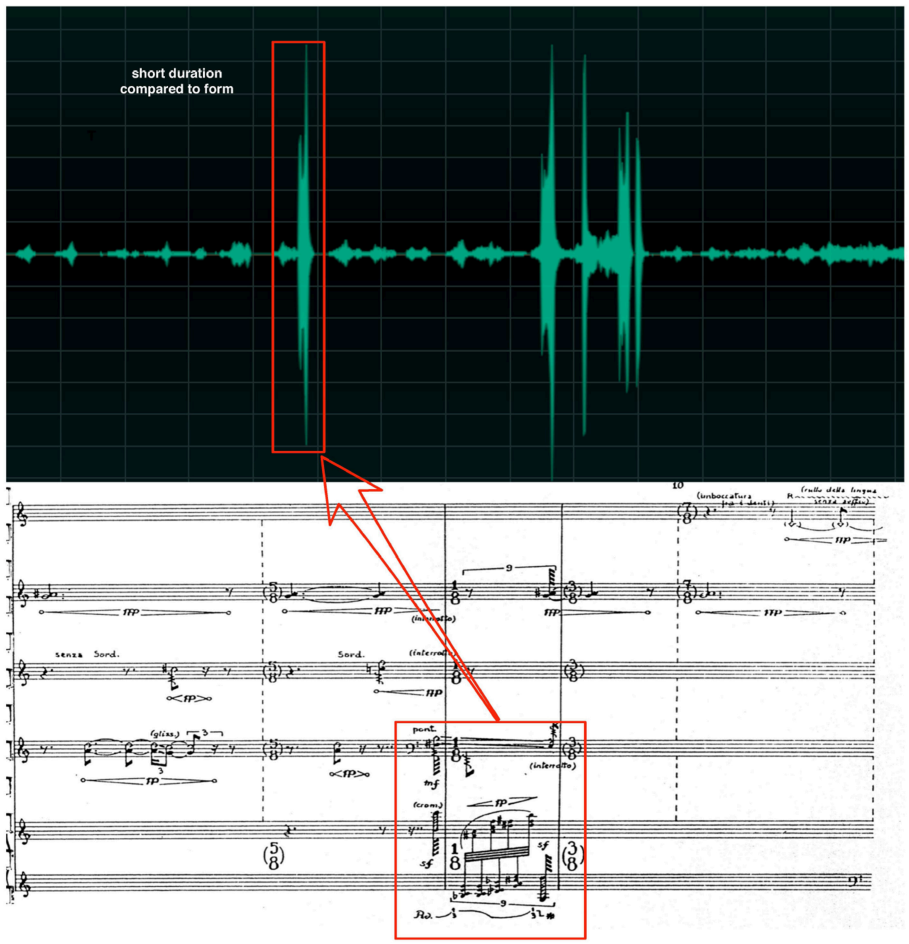

FIGURE 3 | The figure shows the beginning part of Salvatore Sciarrino's Lo spazio inverso. For the majority of the piece, the dynamics is very soft (from *ppp* to *p*) and the instrumentation does not include keyboards. All of a sudden, the keyboard is used in the piece with a dynamic of *fff* thus creating an immediate fracture in musical discourse. While these events are very scarse and sparse in time, they nonetheless represent the most important musical elements of the piece. In this context, a statistical approach aimed at finding most *probable* event, would classify the scarce phrases played by the keyboards as outliers while in reality they are actually the center of the musical logic.

can be translated as *"Some music pieces of genre1 contain a dominant 7th chord on the dominant followed by a major chord on the tonic (i.e., a perfect cadence),"* and these rules are used for automatic genre classification. The authors in Arabi and Lu (2009) also use chord information for genre identification, by looking at the most frequent patterns of 3 or 4 triads that appear in a given genre. The approach proposed in Anglade et al. (2009) is more expressive, and captures some human-readable meaningful musicological concepts.

However, approaches based on logic have not been directly applied to audio, but they have been used on symbolic representations such as MIDI files. Some ILP-based approaches for the automatic characterization of harmony have been extended to audio (Anglade et al., 2009, 2010), but they require an off-the-shelf transcription step: harmony characterization is induced from the output of a chord transcription algorithm and not directly from the audio signal.

2.1.3. Logic and Probability Theory: Synergistic Views

As reflected by previous work, both aspects of uncertainty and rich relational structure are important in music and should be fully considered. Traditional probabilistic approaches are not able to cope with rich relational structure, while logic-based approaches are not able to cope with the uncertainty of audio and need a transcription step to apply logical inference on a symbolic representation. Logic and probability theory offer complementary benefits to these two aspects:

- Probabilistic graphical models are elegant approaches to represent the manner in which variables statistically depend on each other and to deal with uncertainty. But they are propositional and thus insufficiently expressive to handle the relational complexity of the domain of interest.
- On the other hand logic has sufficient expressive power to model rich relational knowledge, but it cannot handle uncertainty of real data.

Music retrieval tasks would benefit from a unification of these two perspectives.

2.2. Complex Relational Structure at Multiple Abstraction Levels and Time Scales

2.2.1. Multiple Abstraction Levels

Music audio signals are complex from a semantic point of view and convey multi-faceted and strongly interrelated information (e.g., harmony, rhythm, structure, etc.). For instance, some chords are heard as more stable within an established tonal context (Krumhansl, 1990) and the chord progression cannot be analyzed without considering the tonal context (see **Figure 4**). They are also related to the metrical and semantic structures [e.g., in popular music chord changes are often related to downbeats (Papadopoulos and Peeters, 2011), semantically same segments (verse, chorus) often have similar chord progression (Papadopoulos and Tzanetakis, 2013)].

A number of work have shown that content-based MIR benefits from a unified musical analysis that estimates jointly interrelated musical information (Burgoyne and Saul, 2005; Paiement et al., 2005; Lee and Slaney, 2008; Mauch and Dixon, 2010; Papadopoulos, 2010; Böck et al., 2016). However, many of the existing computational models extracting content information tend to focus on a single music attribute, without exploiting prior context, which is contrary to the human understanding and perception of music that is known to process holistically the global musical context (Prince et al., 2009).

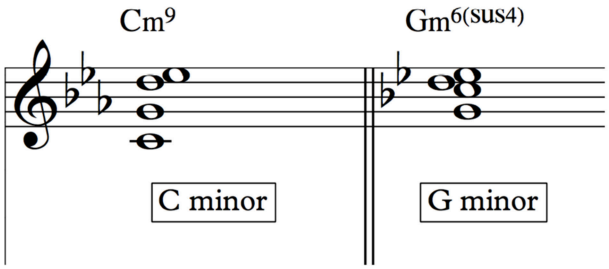

FIGURE 4 | Depending on the context, identical sets of notes can assume different functions and, as consequence, different names. In the figure we see the chord made of the notes [C, D, Eb, G]: this chord can be called C minor 9th in the context of the key of C minor and G minor sus4 6th in the context of the key of G minor. Note that this is not a simple change of function: in this case not only the function is different but also the very name of the chord. This ambiguity can be a problem in the context of data labeling, where different experts could assume different hypothesis on the context thus leading to different naming.

We believe that one of the reasons for this is that it is difficult to model these complex relationships at multiple levels of representations (multi-time scale, multi-features) using probabilistic graphical models, since they quickly become intractable. Current PGM approaches in MIR dealing with multiple interrelated features generally resort to some manual "tricks" to make inference tractable. For instance, the model presented in Pauwels and Peeters (2013) jointly estimates chords with keys and structural boundaries using a HMM in which each state represents a combination of a key, a chord, and a structural position. The authors manually incorporate some musicologically motivated constraints in the transition matrix, which allows reducing the computation time of the Viterbi decoding step. Another example can be found in the work (Mauch and Dixon, 2010) that presents a chord estimation model consisting in a DBN where states are a combination of chords, metric position, bass, and key. To be able to tractably infer the most likely state sequence, the authors perform a preprocessing step to remove the chords that appear least often among the locally best-fitting 10 chords at every beat.

2.2.2. Multiple Time Scales

In a piece of music, the above-mentioned various interrelated musical dimensions interact at multiple time scales. Musical information is structured at the temporal level in a hierarchical way. For instance, in many types of music, the chord progression is related to the semantic structure (here *semantic structure* is understood as the highest-level expression of the structure, e.g., segmentation into "ABA" form, verse/chorus, etc.), which itself

is dependent on musically lower-level organization such as beats and bars.

An important limitation of many existing models for music-content estimation is that, up to now, analysis is typically performed at a short time scale only, ignoring longer-term dependencies between music events, and resulting in outputs that are often incoherent and thus implausible. A fundamental question that remains open is how to use the long-term hierarchical semantic structure of musical pieces to improve the validity, and thus usability, of computational music analyses (Müller et al., 2016).

The MIR community has explored some approaches to encode longer-term structure with short-term analysis. At the feature representation level, the use of beat-synchronous features (Ellis, 2007a), or concatenated consecutive frame-wise features into a single feature vector (Ellis, 2007b; Gaudefroy et al., 2015) allows encoding some of the higher level structure.

More recently, the breakthrough of deep neural networks has made possible to incorporate some context into the feature representation and model the latent complexity of the data. In music, many events are not isolated but integrated into a global musical context. We have mentioned before for instance that in modern Western tonal music, it is necessary to take into account the tonal context to infer a chord (**Figure 4**). Deep architectures are capable to take into account some of the hierarchical nature of music, and allow building high-level representations from intermediate layers of non-linear transformations and representations, incorporating context.

For instance, in the chord recognition model presented in Humphrey and Bello (2012), a convolutional neural network (CNN) encodes local behavior five-second tiles of constant-Q pitch spectra in feature maps at higher levels. This fully data-driven approach achieves state-of-the-art results compared to traditional models using handcrafted short-time features combined with post-filtering temporal smoothing. However, the model is unable to classify chords sparsely represented in the data. Data scarcity is addressed in more recent works (McFee and Bello, 2017), where the proposed deep convolutional-recurrent model for automatic chord estimation allows exploiting structural relationships between chord classes and works with a large chord vocabulary of 170 classes. It is found that the structured representation facilitates modeling infrequently observed and complex chords, but that extending the approach to effectively support extended chords calls for the need of larger annotated corpora.

Efforts to incorporate longer-term temporal structure with have also been made at the statistical modeling level. In the work of Dannenberg (2005), music structure is used to constrain a beat tracking program based on the idea that similar segments of music should have corresponding beats and tempo variation. In the work of Mauch et al. (2009), the repetitive structure of songs is used to enhance chord extraction. However, in both piece of work, the modeling of the hierarchical structure is not flexible and dos not reflect the actual complexity of music. For instance, although commonplace, variations between several occurrences of a section cannot be taken into account. Until now, most models are limited to a short time scale analysis and longer-term structure is possibly encoded during post-processing step.

Recently, recurrent neural networks (RNN) have been explored for many other music tasks as an alternative to model longer-term temporal sequences, since they can in principle describe arbitrarily complex long-term temporal dependencies (see e.g., in the case of ACE Boulanger-Lewandowski et al., 2013; Sigtia et al., 2015; Deng and Kwok, 2016). However, they have proved to be difficult to optimize to indeed make the model learn long-term dependencies from data as expected (Korzeniowski and Widmer, 2017). Deep neural networks have allowed incorporating context into the representation, compared to handcrafted features mostly derived from short-time analysis frames. However they do not allow explicit modeling of music analysis at multiple time scales.

2.2.3. Capturing High-Level Information at Multiple Levels: A Challenge

The discussion above outlines that describing music at multiple levels of abstraction and detail remains a challenge for the MIR field. As shown above, traditional machine learning approaches and probabilistic graphical models for music processing are not able to cope with the rich, highly structured, relational structure of music. Purely data-driven deep learning approaches suffer from data insufficiency and are difficult to optimize to accurately describe complex long-term temporal dependencies. Existing probabilistic graphical models are limited in taking into account the interrelated music dimensions. Probabilistic inference is computationally expensive and standard inference techniques for very large probabilistic graphical models become quickly intractable.

The unification of logic, learning and probabilities offer several directions to overcome these shortcomings:

- **Describing music at multiple levels of abstraction and detail:** Music analysis would benefit from using statistical relational AI models that exploit the idea of leveraging long-term event analysis to short-term event analysis, as it has successfully been achieved in various other areas, such as in natural language processing (Lafferty, 2001) or bioinformatics (Liu et al., 2005). A recent chord estimation algorithm developed in the framework of Markov logic networks (Papadopoulos and Tzanetakis, 2013) has shown that StarAI approaches offer some perspectives to explicitly model multiple time scales with flexibility. **Figure 5** shows a comparison of chord progression estimated without and with taking into account the semantic structure to obtain a "structurally consistent" representation of music (as presented in Papadopoulos and Tzanetakis, 2013). It can be seen that incorporating long-term structural dependencies in the model results in a more coherent chord transcription.
- **Making abstractions that apply to individuals that share common properties:** Reasoning with music data means dealing with regularities and symmetries (Kempf, 1996) (see **Figures 6**, **7**). Logic is able to jointly handle these two important properties. Objects are put together into classes (such as "major" or "minor") because they share common

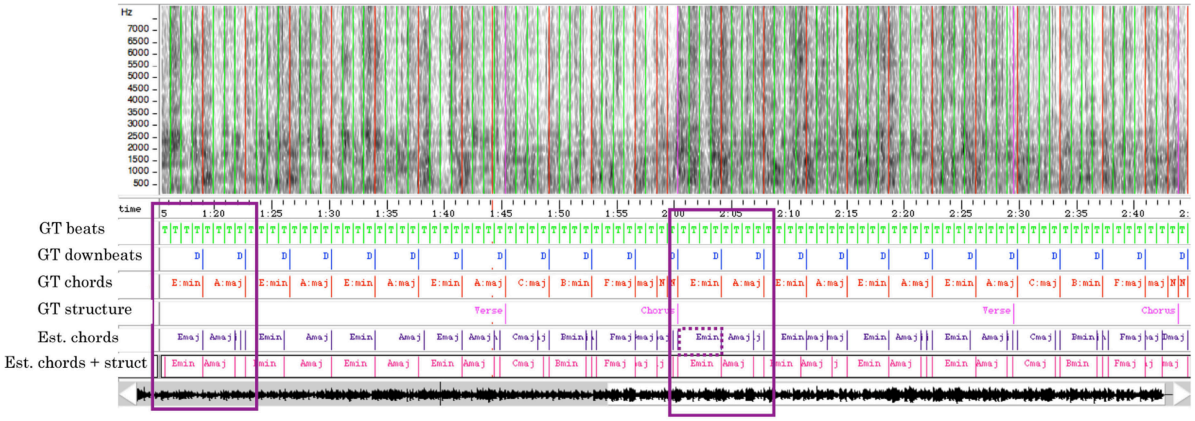

FIGURE 5 | Example of chord progression estimation enhanced with semantic structure information. Extract of the Pink Floyd song *Breathe*. The first 4 lines (beats, downbeats, chords, structure) correspond to the ground truth annotations. The next two lines show the results obtained with models proposed in Papadopoulos and Tzanetakis (2017), designed to obtain a "structurally consistent" chord transcription. The 5th line shows the results obtained with a baseline chord HMM, considering only chroma observation and transitions between successive chords. The last line shows the results obtained with a model that adds long-term chord dependencies: it favors same chord progressions for all instances of the same segment type (see the corresponding graphical model in **Figure 11** below). The ground-truth chord of the first bar of the verse is an Em chord. The baseline chord HMM correctly estimates the second instance of this chord, but makes an error for the first instance (EM instead of Em). This is corrected by the model that favors same chord progression in same segment types.

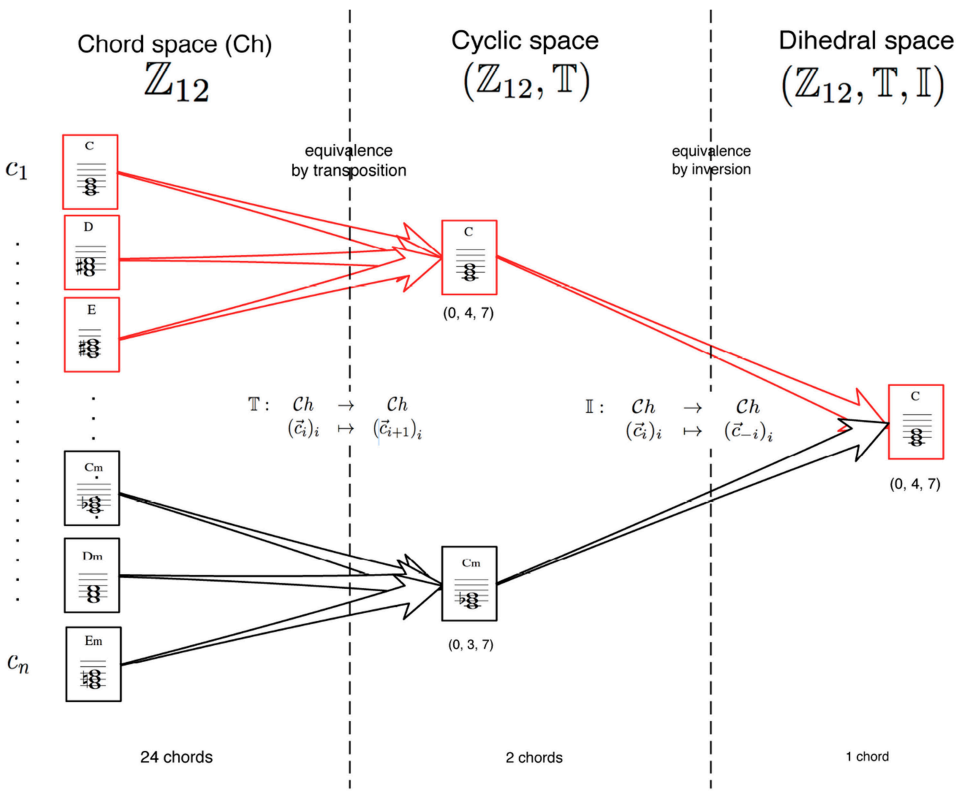

FIGURE 6 | The figure shows the so-called chord *factorization* in tempered space. This space is made by integers from 0 to 11 that correspond to specific pitch classes (C = 0, C# = 1 and so on). In the left column, all the possible major and minor chords are shown, for a total of 24 chords. Factorizing by transposition (that is removing the onset interval from the C) chords are mapped onto the cyclic space and they become 2 (C major and C minor); in other words, F major [5, 9, 0] is the same chord than C major [0, 4, 7] but with an onset interval of 5 semitones. The last column of the figure shows one more factorization applied, this time by inversion (that is changing sign to the pitch class and applying mod 12) that brings the chords in the dihedral space. After this process, C major and C minor appear to be the same chord: in tempered space C major is the pitch class set [0, 4, 7], while C minor is [0, 3, 7]. After inversion, the two sets are equivalent to [0, 4, 7] (Lewin, 1987).

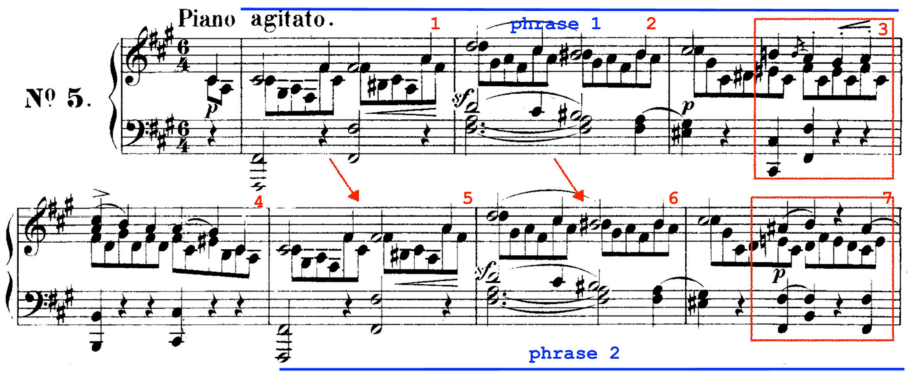

FIGURE 7 | The figure shows the beginning of Mendelssohn's Song without words Op. 19 No. 5. The excerpt contains two musical phrases (bars 1–4 and 5–7) that have highly regular structure; indeed, bars 1 and 5 and bars 2 and 6 are the, respectively, equal while the only differences are in the ending part of bars 4 and 7. This type of musical regularity is very common in classical and romantic music.

properties (e.g., all major chords share some common properties as for instance they all contain a major third above the root). Current MIR algorithms do not always take advantage of these symmetries. For instance many algorithms for cover song identification rely on the fact that cover versions of a song are likely to have a similar chord progression. However, the cover version will often present some variations, such as key transposition. This problem has been tackled in general by testing all possible feature transpositions (or related approaches that attempt to speed up this process Serrá et al., 2010), resulting in an additional computational effort. This effort could be leveraged taking advantage of logical representations that naturally exploit symmetries.

- **Scalability and efficient inference by reasoning about regularities across different situations:** Logic can handle the complexity of music by describing it with commonalities and regularities. New class of lifted inference algorithms (Poole, 2003) that take advantage of these regularities has been motivated by the advent of statistical relational languages. In these formalisms, first-order logic is used to define complex interactions between random variables in large-scale probabilistic graphical models. The model is declared over classes and hierarchies using variables and predicates, and these logic formulas are used as templates to construct graphical models, instead of having to stipulate them for each single entity separately (see **Figure 8**). With such a concise description, high-level structure and symmetries in the model can be exploited to restrict the search space for efficient inference, such as in the case of lifted inference (Kimmig et al., 2015).

2.3. Handling Multimodal Information

2.3.1. Increasing Amount of Various Heterogeneous Sources of Knowledge

Annotated corpora for research purpose are only a small part of available information and context for music signal processing. There is an explosive growth in the amount of available heterogeneous sources of music-related information, complementary to the audio (e.g., video, music sheets, user MIDI transcription, metadata such as social tags on forums, etc.). Moreover, because of the theoretical and cultural nature of music, there are many sources of expert knowledge that trained musicians use as cues to analyze music, and that can be incorporated into the models (e.g., theoretical music rules from music theory Riemann, 1896; Schoenberg, 1969; Rameau and Gossett, 1971). Music knowledge can also be derived from data, as for instance chord sequences learned using probabilistic context-free grammar model (Tsushima et al., 2018). Despite the sustained claim for the development of multimodal approaches for music analysis (Müller et al., 2012; Herremans and Chuan, 2017; Smith et al., 2017) for a few years, these multiple multimodal sources of knowledge are still little exploited in existing models for content-based retrieval.

2.3.2. Leaning-Based vs. Expert Knowledge-Based Approaches

Besides, computational music analysis approaches generally focus either on learning-based approaches or expert knowledge-based approaches. In early research, with few training data available, most parameters (such as the transition probabilities in a HMM) and features in early ACE approaches were initially set by hand (Bello and Pickens, 2005; Shenoy and Wang, 2005; Papadopoulos and Peeters, 2007; Oudre et al., 2009). Such purely expert systems have considered simplified versions of the task and performed chord recognition considering a very small vocabulary, typically 24 major and minor chords. The expansion of annotated databases in the MIR community has led to the development of purely data-driven approaches, and has allowed addressing the question of scalability. Data-driven approaches, in particular neural networks, are successful and efficient at large-scale modeling, and at obtaining insight from large data collections. They also allow discovering solutions not previously considered when using only expert knowledge. In

Knowledge	*First-order Logic formula*
A major chord implies a happy mood.	$\forall x\ \text{IsMajor}(x) \Rightarrow \text{IsHappyMood}(x)$
If 2 chords are neighbors, either the 2 are major chords or neither are.	$\forall x\ \forall y\ \text{AreNeighbors}(x, y) \Rightarrow (\text{IsMajor}(x) \Leftrightarrow \text{IsMajor}(y))$

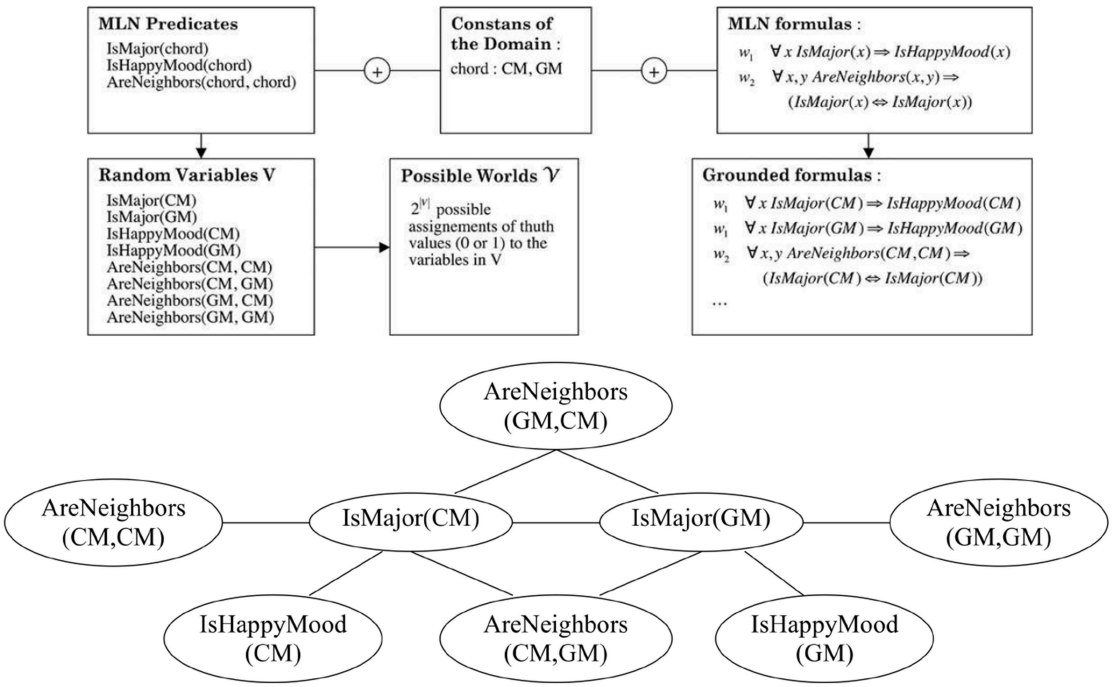

FIGURE 8 | Markov network **(Bottom)** obtained by applying two logical formulas **(Top)** (that describe knowledge about the relationship between chords and mood) to the two constants CM chord and GM chord. The logical formulas can be viewed as a template for constructing the Markov networks: given different set of constants, it will produce different networks; the size of the Markov network grows as the number of constants increases but the number of the formulas (or the templates) stays the same. The grounding process is illustrated in the **(Middle)**. For more details see Papadopoulos and Tzanetakis (2017).

particular recurrent neural networks have allowed considering ACE systems handling large chord vocabularies (217 chords classes in Deng and Kwok, 2017, 170 classes in McFee and Bello, 2017). A shortcoming of these approaches is that MIR neural-network algorithms still heavily rely on supervised learning and their efficiency is usually conditioned upon a large amount of annotated data.

Human beings are capable of learning from very few data examples and resolve inconsistencies, something that machines cannot do in general. Annotated corpora are often unbalanced and inevitably contain annotation errors, which make computational models behave in unforeseen ways. For instance, in the context of ACE estimation using a large chord vocabulary, the authors in Deng and Kwok (2016) observe a drop in the classification when training their model using multiple datasets from different sources. They argue that the reason is that chords inversions in different datasets are annotated differently, which makes the classifier getting confused when trained on the combined dataset.

2.3.3. Taking Into Account Human Subjectivity

Furthermore, annotations are largely prone to human subjectivity, with experts disagreeing in the annotations (Ni et al., 2013; Koops et al., 2017). Evaluation of the algorithms suffers from some inherent methodological pitfalls that make difficult to reliably evaluate and compare the behavior of ACE algorithm (Humphrey and Bello, 2015). For instance, the Beatles corpus, widely used for ACE evaluation contains 66 chord types, with cardinalities ranging from zero to six (Harte, 2010). Evaluation is done by mapping reference chords of the ground truth to the classes of the chord lexicon handled by the algorithms. Very often, the chord lexicon reduces to two chord types (major/minor). For a 24 chords minor/major dictionary, as in MIREX competition[5], a C minor seventh chord $C:min7$ (C-Eb-G-Bb) will be mapped to C minor $C:min$ (C-Eb-G). Template-based ACE chord algorithms that do not consider

[5]See http://www.music-ir.org/. Note that if early MIREX competitions considered only major/minor chords, current editions include seventh chords with inversions.

tonal context may estimate $C:min7$ as a $C:min$ as well as an E major $E:maj$ (Eb-G-Bb), since both chords are contained in $C:min7$. However, in the evaluation, $E:maj$ will be considered as an error. It has been advocated that this subjectivity should be embraced, possibly by incorporating domain knowledge (Humphrey and Bello, 2015), but current evaluation schemes do not allow such possibilities.

2.3.4. Some Possible Directions to Operate With Complementary Heterogeneous Data

As more complementary heterogeneous sources of music-related information are becoming available, an important question that remains open is how to exploit this knowledge, how to combine it with learning, how to deal with insufficient or inconsistent training data and how to reconcile large-scale modeling with human-level understanding.

Recent work in the statistical relational leaning area offer interesting reflections on this aspect:

- **Ability to flexibly incorporate prior knowledge**: As discussed in De Raedt et al. (2016), statistical relational AI aims at developing approaches that allow incorporating domain knowledge in the models in a flexible way: prior knowledge could be used to overcome the lack of training data, but should be refutable if there is enough strong evidence against it. Also, it is needed to be able to learn from multiple heterogeneous data sets, derived from different type of data, different contexts or different experts, at different levels of abstraction and possibly with overlapping or contradictory concepts (Thimm, 2016). The combination of logic and probabilities allows such flexibility. For instance pure first-order logic does not suffer contradictions. Markov logic that extends first-order logic with probabilities allows handling contradictions between formulas by weighting the evidence on both sides (Richardson and Domingos, 2006). Another interesting aspect is that Markov logic can be used to handle collective knowledge bases since it allows merging several knowledge bases, even if they are partly incompatible (Richardson, 2004).
- **Using expert knowledge to leverage lack of training data:** In some situations we have little or no training data, but instead we have access to the subjective probabilities of a domain expert. For example, we may have little data about certain rare chords, but we may have a subjective notion of what percentage of a rare chord occurs among the others. The possibility of incorporating such domain knowledge combined with training (Pápai et al., 2012), or to use domain knowledge for guiding the learning process (Kuzelka et al., 2016), may help leverage music complexity.
- **Incorporating knowledge to enhance purely data-driven approaches:** Data-driven MIR approaches would benefit from extensions of deep learning based approaches that allow integrating logical knowledge: it has been shown in other domains that incorporating logical knowledge adds robustness to the learning system when errors are present in the labels of the training data (Donadello et al., 2017a).
- **Interactive machine learning:** Flexible and expressive approaches, that allow an intuitive description of knowledge, open the possibility to put the human in the loop and take into account the users feedback to iteratively refine the developed models and improve its performances, as done for instance for the task of machine reading (Poon and Domingos, 2010). As an example of interactive machine learning we cite the software Orchidea, by the authors (Cella, 2018). Orchidea is tool to perform static and dynamic assisted orchestration by means of multi-objective optimization on various multi-dimensional features. Assisted orchestration can be thought as the process of searching for the best combinations of orchestral sounds to match a target sound under specified metric and constraints. Although a solution to this problem has been a long-standing request from many composers, it remains relatively unexplored because of its high complexity, requiring knowledge and understanding of both mathematical formalization and musical writing. The approach adopted by Orchidea is by putting the composer (main type of user of the tool) in the loop of optimization: he/she can guide the optimization process by introduction in-step symbolic constraints that reduce the search space of the algorithm. At any give number of epochs of the optimization the software presents the intermediate results to the user that can validate, eliminate or correct solutions from high level by introducing constraints in the cost function. We refer the reader to the vision paper (Gurevych et al., 2018) for possible directions for interactive machine learning.

2.4. (Dis)ability to Generalize Toy Problems to Real Tasks

2.4.1. Getting Stuck to Toy Problems

A typical approach in MIR is to tackle simplified versions of a problem in order to develop algorithms and techniques to be subsequently transferred to more real problems (the usual *Future developments* section in papers). More than often, the developed techniques make very strong assumptions about the problem itself and cannot really be transferred to more general cases. As a consequence, some MIR research has been devoted to solve problems that are not large enough to be applied to real tasks and the developed methodologies got stuck to a specific context and cannot be reused.

While the common two-step feature extraction/classification scheme for ACE has been useful in several contexts (for example for pop music) it also shows inherent limitations that make it impossible to transfer the used concepts to other contexts. For example, using beat-synchronous features are only meaningful in some kind of music: it makes no sense to talk about them in the context of contemporary classical music in which most of the language develops around different interpretations of time. In the same way, using chroma features makes impossible to analyze music that is based on different tuning systems, such as Indian music. Also, the fact that chord templates are used means that only a possible subset of harmonic choices is considered, making impossible to analyze contemporary jazz, that also exits normal functional rules for harmony.

2.4.2. Scalability at the Expense of Explainability

Deep learning approaches have allowed scalability and handling real problems. For instance, starting from the premises that using appropriate features is an essential aspect of chord estimation systems (Cho and Bello, 2014), the authors in Korzeniowski and Widmer (2016a) use deep neural networks to compute harmonically relevant chroma features, that are robust to irrelevant interferences (e.g., overtones, percussive instruments, etc.), compared to handcrafted ones. To avoid temporal smooth post-processing, they learn the deep-chroma features using some context audio frames. They conclude that a context of 1.5 s is adequate for local harmony estimation. They show the robustness of the proposed deep-chroma features for chord recognition, but acknowledge that it is difficult to understand what the neural network learned, and on which basis they generated their output. They also explain that they chose to derive a data-driven approach because *it is close to impossible to find the rules or formulas that define harmonic relevance of spectral content manually*.

Deep learning approaches have also allowed constructing sophisticated end-to-end data pipelines. The authors in Korzeniowski and Widmer (2016b) develop an end-to-end chord recognition system that employs fully convolutional neural network for feature extraction combined with a CRF for chord sequence decoding. The model is evaluated considering 25 chord classes. They investigate if the system extracts musically interpretable features, looking at questions such as whether *the network learned to distinguish major and minor modes independently of the root note*. They look at feature maps that have a high average contribution to minor and major chords and are able to identify some kind of a zig-zag pattern that discriminates between chords that are next to each other in the circle of fifths, but they are not able to explain why and how the feature maps contribute to the learning.

Ad-hoc representations (such as chroma features) created by MIR are defined by sound mathematical models and embody many years of research on a specific domain but fail to achieve the expressivity of deep learning. On the other hand, deep learning proved to be valuable in incredibly different domains and showed that some learning techniques are indeed general and can be transferred to different contexts, but does not always permit a logical understanding of the problem. Deep learning techniques can discover multiple hierarchical layers of data and generate new feature combinations, but those intermediate layers are not always easily interpretable.

2.4.3. Some Possible Directions Toward General, Scalable, Modular, and Explainable Models

From what is stated above, it appears that good representations are essential to build robust and expressive systems for musical processing. It is not easy, anyway, to define the concept of good representations. Among important properties for musical representations, as seen in the previous sections, there are milestones such as the capacity of handling multiple abstraction levels and the capacity of acting on multiple time scales. Also, the description of a musical signal usually targets a particular degree of abstraction.

Low-level representations (such as waveforms) are more generic and have very high dimensionality; mid-level representations, such as chromas, are often related to perceptual concepts (Ellis and Rosenthal, 1995) and have a higher level of abstraction and a smaller dimensionality; they allow for transformations on specific concept (variables), usually defined on discrete domains (for example 12 equal-tempered pitch classes in $0, 1, \ldots, 11$ comprising the chromatic scale) (Lostanlen and Cella, 2016). Very abstract representations (for example compositional schemes such as the one in **Figure 9**), finally, are more expressive and have a low dimensionality; these representations deal with almost stationary entities such as musical ideas and unfortunately it is very difficult to know which mathematical structure stays behind and to incorporate them into approaches for music processing without an adequate formal expressive language.

Oversimplified approaches generate solutions that can never handle the complexity of real problems: in order to cope with them, we need a new language that combines learning, logic, and probability. Such approaches would open perspectives to overcome the shortcomings described above. Several points are of particular interest to the music community, and would help designing systems that are more general, scalable, modular, and explainable:

- **Explainability:** Despite the tremendous impact of deep learning, limited progress has been made until now toward understanding the principles and mechanisms underlying this language, and how to integrate or learn human interpretable rules. This need has been a long-standing problem since the beginning of AI. The interpretability and expressive power of logic combined with the effectiveness of neural network learning would open interesting perspectives toward this demand. These last years, the question of integrating (deep) learning and reasoning methods into a unified explainable foundation has received more and more attention in particular in the StarAI field (d'Avila Garcez et al., 2015; Šourek et al., 2015; Donadello et al., 2017b; Wang and Poon, 2018). The MIR field would benefit from such insight .
- **Structure learning:** In the aim of improving systems for musical signal processing and enlarging their perspectives, it would be interesting to take advantage of *structure learning* algorithms developed in the field of StarAI that are able to learn the relational structure of the data (Friedman et al., 1999; Kok and Domingos, 2005). Approaches developed in the StarAI field combine elements of statistical and relational learning, and the structure can be learned in a humanly interpretable way. Of particular interest is *statistical predicate invention* (Kok and Domingos, 2007) that allows the discovery of new concepts, properties and relations in structured data, and that generalizes hidden variable discovery in statistical models and predicate invention in Inductive Logic Programming.
- **Deep transfer learning:** Another direction of interest is *deep transfer learning*. Human knowledge is modular, composable

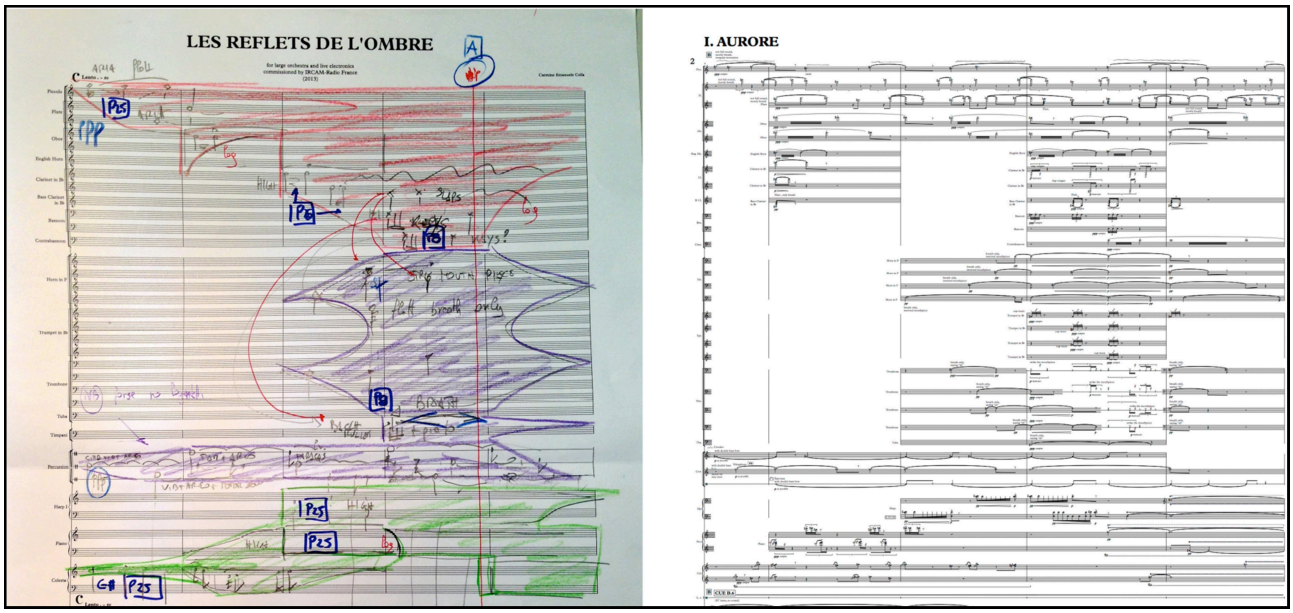

FIGURE 9 | The figure shows a high-level compositional representation used by the author in his piece *Reflets de l'ombre* (2013) for large orchestra and electronics. In the left part of the image, there is a drawing aimed at outlining the spectral morphology of the piece, together with high-level specific information such as musical figures and harmonies. It sketches a high-level visualization of an imaginary musical situation before the actual realization by means of notes, durations and so on. From a composer's perspective, such image is a static mono-dimensional entity that is related to other musical ideas during the development of the piece: for example, this drawing represents a sound that the composer defines *hollow*. In the right part, there is the corresponding actual orchestral score: the actual *realization* of the musical idea lies in a high-dimensional space given by the multiple entities involved: notes, durations, playing styles and so on. Each entity has a large space of variations that interacts with other entities: each time-point of the piece can be used for any combination of notes with any duration thus having a large space of possibilities. It is very difficult (if possible) to find a formal relation between the two representations. Even if these mental representations probably cannot be fully captured and described even by StarAI methods, we believe that these methods can be used to approach better such kind of representations: a logic-based language in connection with sound-related description helps in defining the connection between musical ideas and their realizations.

and declarative: as such it can be reused. This stands in contrast with most of the approaches used today in MIR. Deep learning and reinforcement learning made some substantial progress in the direction of modular representations but nonetheless this is just the beginning of the road. The MIR field has started embracing *transfer learning* techniques (Pan and Yang, 2010) for feature transfer that use knowledge gained from learning a source task (Hamel et al., 2013), or multiple source tasks (Kim et al., 2018), to aid learning in a related target task performed on the same type of raw input data. This is particularly interesting when having few training examples for a task. The statistical relational AI field has opened the path to *deep transfer learning* (Mihalkova et al., 2007; Davis and Domingos, 2009), that generalizes transfer learning across domains, by transferring knowledge between relational domains where the types of objects and variables are different. In deep transfer learning, the knowledge to be transferred is the relationship among the data. For instance we could use knowledge learned in the movie industry to help solve tasks in the music domain (see **Figure 10**). These directions are steps toward designing a generalized framework for music processing that can discover its own representations at once, as humans do, and is able to integrate prior knowledge.

3. STARAI, A POSSIBLE DIRECTION FOR AN INTEGRATED FRAMEWORK

The previous discussion outlines the necessity of a general framework able to learn, reason logically, and handle uncertainty for processing music data. Real data such as music signals exhibit both uncertainty and rich relational structure. Probabilistic graphical models can cope with the uncertainty of the real world, but they cannot handle its complexity. Logic can handle the complexity of the real world but not its uncertainty. Prior representations such as chroma features formally define the level of abstraction, but cannot reach the same level of aggregate information gathered by deep learning networks. These networks, on the other hand, are not capable of explaining the concepts they discover. As advocated in Cella (2017), for such reasons, it is interesting to make a bridge between these approaches by immersing all these aspects in a more general framework, made by the unification of learning, logical, and probabilistic knowledge representations.

As stated at the beginning of this article, and illustrated through our analysis of existing content-based MIR approaches, logic and probability have been generally treated separately. Probability being the classic way to represent uncertainty in

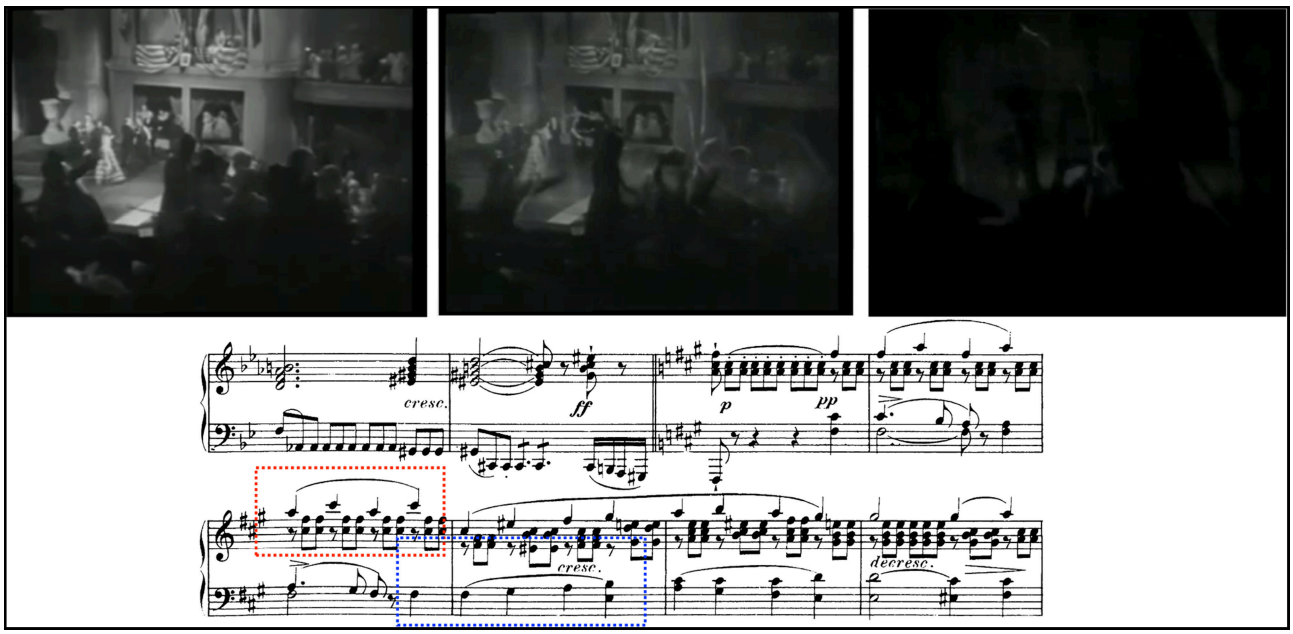

FIGURE 10 | Although the data are of different natures, images in films and audio samples in music signals share common semantic relations. For instance the phenomenon of fade-out/fade-in in a film, where images are superimposed during the transition [see the example of D. W. Griffith's *Abraham Lincoln* film **(Top)**] finds a correspondence in music when one subphrase finishes while another begins [see the extract of Schubert D.960 sonata **(Bottom)**]. Transfer of knowledge of transitions in movies could help finding transitions in audio signals. Source of the images: screenshots of the movie. No permission is required for their use: the film entered the public domain in 1958 when the initial copyright expired. See Paolo Cherchi Usai (2008), *The Griffith Project: Essays on D.W. Griffith*, British Film Institute. p. 208. Retrieved January 16, 2016.

knowledge, while logical representation being the classic way to represent knowledge and complex relational information. Both practice converged to the same idea: that the world is both uncertain and it is made of structured relational objects in it, and that logic and probability should be unified to deal with it. This path has already been embraced by a number of other area emerging from both traditions, evolving to the field of Statistical Relational Artificial Intelligence (StarAI). The recent tremendous resurgence of neural networks and deep learning approaches have joined the same agreement. Approaches developed in the StarAI field are able to handle concurrently time large-scale, structured and uncertain domains.

We advocate that it is time that the MIR community considers the perspectives offered by this promising research field . Music is by nature complex, relational and ambiguous, it would thus benefit from approaches that allow representing, reasoning and learning under uncertainty, complex relational structure, rich temporal context and large existing knowledge. Most models currently used in MIR are special cases of StarAI approaches and can find powerful extensions that combine logic and probability.

In this section, we present some mainstream approaches developed in the fields of StarAI (section 3.1). We summarize the potential benefits to embrace approaches that unify logic, learning and probability (section 3.2), and we present a case study chord recognition model using StarAI approach (section 3.3). We then point out some obstacles and challenges that remain to be addressed (section 3.4). We conclude with the perspectives offered by the use of integrated approach for the MIR field (section 3.5).

3.1. Toward Unification With StarAI

With the growing field of StarAI, many representations in which learning, statistical, and relational knowledge are unified within a single representation formalism have been proposed. The abundance of these approaches illustrates the richness and maturity of the field. As reviewed in De Raedt et al. (2016), early approaches in the 1990's consisted in pairwise combinations of the three key ingredients of AI (logic, probability, and learning) resulting in *probabilistic learning* (see Koller and Friedman, 2009; Murphy, 2012 for a review), *logic learning* such as inductive logic learning and relational learning (see De Raedt, 2008 for an overview), or *probabilistic logic* (see Nilsson, 1986). More recent approaches such as probabilistic relational models (Getoor, 2001), Bayesian logic programs (Kersting and Raedt, 2001), or Markov logic networks (Richardson and Domingos, 2006) combine the three ingredients. StarAI approaches have been successfully applied to a multitude of problems in various structured, uncertain domains. We refer the reader to Getoor and Taskar (2007), Raedt and Kersting (2008), Domingos and Lowd (2009), Blockeel (2013), Kimmig et al. (2015), De Raedt et al. (2016), and Besold and Lamb (2017) for surveys of StarAI.

We present here three relational probabilistic model formalisms that are probably the most prominent of the main streams in StarAI, and that illustrate how each tradition has

converged to the idea of unifying all aspects together: *Markov logic networks* that extend a probabilistic graphical model with logic, *ProbLog* that extends logic programming language with probabilities, and *logic tensor networks* (LTNs) (Donadello et al., 2017b) that integrates symbolic knowledge with deep networks.

3.1.1. Markov Logic Networks (MLNs)

Markov logic networks (MLNs), introduced by Richardson and Domingos (2006), combine first-order logic and probabilistic graphical models (Markov networks). They have received considerable attention in recent years. A MLN is a set of weighted first-order logic formulas, that can be viewed as a template for the construction of probabilistic graphical models.

A *first-order knowledge base* (KB) is a set of formulas in first-order logic, constructed from predicates using logical connectives and quantifiers. A first-order KB can be seen as a set of hard constraints on the data. In a real world scheme, logic formulas are *generally* true, but not *always* true. The basic idea in Markov logic is to soften these constraints to handle uncertainty by adding weights to formulas. The weight associated with each formula reflects how strong a constraint is.

MLNs and their extensions [e.g., Slice Normalized Dynamic Markov Logic Networks (Papai et al., 2012) or Bayesian Logic Networks (Jain et al., 2011)] encompass in an elegant and compact way the probabilistic models that are used in the MIR literature (HMM, CRF, DBN, etc.), while allowing modeling more flexible and complex relational structures. MLNs are intuitive representations of real-world scenarios: weighted first-order logic formulas are first used to express knowledge (weight express the belief in the truth of the corresponding formula); a Markov network is then constructed from the instantiation of these FOL formulas; finally inference is performed on the Markov network.

Instead of using standard learning and inference algorithms for probabilistic graphical models, far more scalable techniques that exploit logical structures and symmetries that are encoded in the MLN representation have been developed (Venugopal, 2015; Sarkhel et al., 2017; Al Farabi et al., 2018), which enables solving very large and complex real-world MLNs.

First-order logic is the special case of MLNs obtained when all weights are equal and tend to infinity. From a probabilistic point of view, Markov logic allows very complex models to be represented very compactly. It also facilitates the incorporation of rich domain knowledge that can be combined with purely empirical learning, and allows reasoning with incomplete data (Papai et al., 2012; Snidaro et al., 2015).

Open-source implementations, as well as further materials on how to use MLNs, are available, for example the *Alchemy*[6] and *ProbCog*[7] software packages.

3.1.2. ProbLog

ProbLog (De Raedt et al., 2007) is a probabilistic extension of the logic programming language *Prolog* (Flach, 1994). ProbLog is arguably the simplest probabilistic extension of Prolog one can design. ProbLog is essentially Prolog where all clauses are labeled with the probability that they are true, these probabilities being mutually independent. Compared to Prolog, ProbLog allows intuitively building programs that not only encode complex interactions between large sets of heterogeneous data, but also handles uncertainties of the real world. Prolog allows determining whether a query succeeds or fails, whereas ProbLog allows computing the *probability* that it succeeds. This allows modeling and reasoning in a real world scheme, and deal with the degree of belief about relational objects.

Open-source implementations of ProbLog can be found at https://dtai.cs.kuleuven.be/problog/. ProbLog and related formalisms have a number of important features and have been used for a variety of applications (De Raedt and Kimmig, 2015), especially in the field of bioinformatics.

3.1.3. Logic Tensor Networks

Finally, *Logic Tensor Networks* (LTNs) (Serafini and d'Avila Garcez, 2016; Donadello et al., 2017a,b) provide a model that combines first-order logic and neural network-based approaches. Learning, based on tensor networks (Socher et al., 2013), is integrated with reasoning using first-order many-valued logic (Bergmann, 2008). Data (logical constants) are described as feature vectors of real numbers. Using relational symbolic knowledge (specified compactly using first-order logic), they are then translated into soft and hard constraints on the subsymbolic level (implemented as a tensor network). The network learns from numerical data and logical constraints to approximate a solution to the constraint-optimization problem called best satisfiability when faced with new data. LTNs thus enable relational knowledge infusion into deep networks and approximate reasoning on unseen data to predict new facts. Compared to purely neural-network based approaches, where if no examples exist in the training data, the network generally fails to represent the corresponding concept, LTNs can also handle exceptions (Donadello et al., 2017b). There are strong connections between LTNs and neural-symbolic paradigms that address the problem of combining symbolic and connectionist approaches for knowledge representation, learning, and reasoning (Besold and Lamb, 2017).

In deep learning, the possibility of creating abstract representations from raw data is a major need; nonetheless this possibility is not fully understood (Mallat, 2016). In a recent Dagstuhl seminar (d'Avila Garcez et al., 2015), major general opportunities for neural-symbolic learning have been outlined, that are obviously valid for music processing: the mechanisms for structure learning remain to be understood; the learning of the generalization of symbolic rules is an essential process and is still not well understood; retrieval of knowledge from large-scale networks remains a challenge. Logic Tensor Networks and related approaches (see Besold and Lamb, 2017 for an overview) can indeed be useful toward this goal. Methodologically, they create bridges between gaps through changes of representation and configure as a promising path to answer questions about knowledge representation, reasoning, and learning (d'Avila Garcez et al., 2015).

[6]http://alchemy.cs.washington.edu
[7]http://ias.cs.tum.edu/research/probcog

3.2. On the Benefits of an Integrated Framework Over Traditional Formalisms

In this article, we have critically reviewed standard approaches to content-based MIR and we have pointed out four limitations that call for a general framework able to integrate different approaches for the representation of music signals into a common perspective:

- Designing expressiveness and flexibility are obtained to the expense of robustness and vice versa;
- Current architectures are unable to cope with the complex relational structure of music at multiple abstraction levels and multiple time scales;
- Available heterogeneous multimodal sources of information and users feedback are little exploited;
- Simplified versions of MIR problems cannot be generalized to real problems, scalability is often obtained to the detriment of explainability.

Standard approaches for content-based MIR are not able to cope both with its inherent uncertainty and complex relational structure, nor to cope with its rich temporal context and large existing knowledge. The benefits of using approaches that unify logic, learning and probability are manifold.

It is to be noted that, in general, the unified approaches proposed in the statistical relational AI area are extensions of existing probabilistic or logic-based approaches. For instance Markov logic networks (Richardson and Domingos, 2006), represented as sets of weighted first-order rules, generalize first-order logic and encompass probabilistic graphical models. In the same way, feed-forward neural network can be seen as a special case of Lifted Relational Neural Networks (Šourek et al., 2015), that are also represented as sets of weighted first-order rules. Probabilistic context-free grammars (PCFG), that have been used in MIR (see e.g., Kameoka et al., 2012; Tsushima et al., 2018), can be represented as Bayesian logic programs that unify Bayesian networks with logic programming (Kersting and de Raedt, 2007). It is thus generally easy, from a conceptual point of view, to move from traditional approaches to starAI approaches.

We have identified several advantages for the MIR field to develop algorithm in an integrated framework able to learn, reason logically, and handle uncertainty. The main ideas are recalled below:

- A unification of logic and probabilistic graphical models allows combining the strength of these two synergistic formalisms: the representational richness and expressivity of logics and the ability of reasoning with uncertainty of probability theory;
- It allows a compact, expressive and intuitive description of music, and a modeling of music at multiple levels of abstraction and detail;
- The distinction is not only relevant at the representational level: it allows improving the learning of the parameters and the structure of the models by combining principles of both statistical and logical learning. It also allows improving scalability and efficient inference by reasoning about symmetries and regularities across different situations, e.g., using *lifted inference*.
- It allows the flexible incorporation of prior knowledge, from multiple heterogeneous data sets, derived from different types of data, different contexts, from different experts, at different levels of abstraction and possibly with overlapping or contradictory concepts, to leverage lack of training data or enhance data-driven approaches by improving the quality of learning;
- It opens perspectives to design models that are more general and explainable (e.g., design neural network that are interpretable through the inception of first-order logic for the specification of their latent relational structures);
- It allows the discovery of new concepts, properties and relations in structured data with *structure learning*. StarAI algorithms should be able to discover their own representations;
- It allows making use of user feedback (e.g., following the work in *interactive machine learning*), to accumulate new knowledge and refine existing knowledge to converge on a solution;
- It allows developing models that are modular and reusing knowledge, e.g., by using *deep transfer learning* that generalizes transfer learning across domains.

3.3. A Case Study for MIR

Among all StarAI approaches, Markov logic networks (MLNs), that we briefly presented in section 3.1.1 have received considerable attention in recent years and found applications in a wide variety of fields. The potential of this powerful framework for music processing has been investigated in a recent article (Papadopoulos and Tzanetakis, 2017). We detail here this system for chord estimation as a proof-of-concept case study to demonstrate the potential of StarAI methods for music processing.

3.3.1. Formal Definition of MLNs

Syntactically MLNs extend first-order logic by adding weights to logical formulas. Semantically, they induce a probability distribution over the set of all possible worlds, where a world is an assignment of truth values to every grounded predicate.

Definition 1. *A* Markov network*, is a model for the joint distribution of a set of variables* $V = (V_1, V_2, \ldots, V_n) \in \mathcal{V}$ *(Pearl, 1988), often represented as a log-linear model:*

$$p(V = v) = \frac{1}{Z} \exp\left(\sum_j w_j f_j(v)\right) \tag{1}$$

where Z is a normalization factor, and $f_j(v)$ *are features of the state v. A feature may be any real-valued function of the state, but in our case, and in most literature of Markov logic, we focus on binary features,* $f_j(v) \in \{0, 1\}$.

Definition 2. *Formally, a* Markov logic network *L (Richardson and Domingos, 2006) is defined as a set of pairs* (F_i, w_i)*, where* F_i *is a formula in first-order logic and* w_i *is a real number associated with the formula. Applied to a finite set of constants C (to which the predicates appearing in the formulas can be applied), it defines a ground Markov network* $M_{L,C}$ *as follows:*

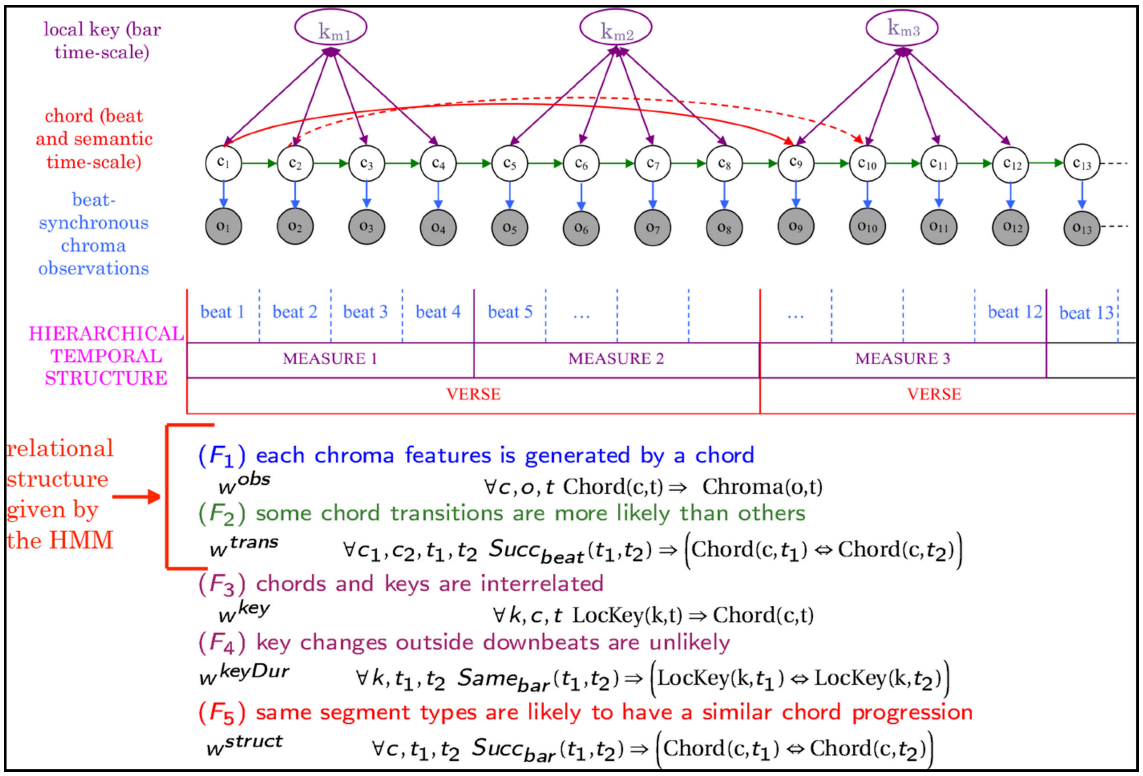

FIGURE 11 | Markov logic network for tonal harmony analysis of music. The model combines harmony-related information (chords, local key, and semantic structure) that interact at various time scales (analysis frame, phrase and global structure) in a single unified formalism. Relational structure is expressed in terms of logical formulas that are used as a template to build a probabilistic graphical model. The first two formulas compactly encode the relational structure that can be modeled with a HMM.

1. *$M_{L,C}$ contains one binary node for each possible grounding of each predicate (i.e., each atom) appearing in L . The value of the node is 1 if the ground predicate is true, and 0 otherwise.*
2. *$M_{L,C}$ contains one feature f_j for each possible grounding of each formula F_i in L. The feature value is 1 if the ground formula is true, and 0 otherwise. The weight w_j of the feature is the weight w_i associated with the formula F_i in L.*

A MLN can be viewed as a *template* for constructing Markov networks: given different set of constants, it will produce different networks. Each of these networks is called a *ground Markov network*. A ground Markov network $M_{L,C}$ specifies a joint probability distribution over the set $\mathcal{V}$ of possible worlds, i.e., the set of possible assignments of truth values to each of the ground atoms in V[8]. From Def. (2) and Equation (1), the joint distribution of a possible world V given by $M_{L,C}$ is:

$$p(V = v) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(v)\right) \tag{2}$$

[8]The ground Markov network consists of one binary node for each possible grounding of each predicate. A world $V \in \mathcal{V}$ is a particular assignment of truth value (0 or 1) to each of these ground predicates. If $|V|$ is the number of nodes in the network, there are $2^{|V|}$ possible worlds.

where the sum is over indices of MLN formulas and $n_i(v)$ is the number of true groundings of formula F_i in v [i.e., $n_i(v)$ is the number of times the *ith* formula is satisfied by possible world V], and $Z = \sum_{v' \in \mathcal{V}} \exp\left(\sum_i w_i n_i(v')\right)$.

3.3.2. Chord Estimation Markov Logic Network

The model proposed for automatic chord estimation in Papadopoulos and Tzanetakis (2017) is depicted in **Figure 11**. It combines different types of harmony-related (chords, local key, global key) information at various time scales (analysis frame, phrase and global structure) in a single unified formalism, resulting in a more elegant and flexible model compared to existing more *ad-hoc* approaches.

The structure of the domain is represented by a set of weighted logical formulas (described by the sentences $F_1, \ldots, F_5$ in **Figure 11**). The constants of the domain are the 24 major and minor triads, and the 24 major and minor keys. The logical formulas applied to these constants produce a Markov network illustrated in **Figure 11**. In addition to this set of rules, a set of evidence literals represents the observations (chroma vectors) and prior information (the temporal structure). Given this set of rules with attached weights and the set of evidence literals, Maximum A Posteriori (MAP) inference is used to infer the most

likely state of the world. For a detailed description of the model, we refer the reader to Papadopoulos and Tzanetakis (2017).

3.3.3. Potential of the Method

This model, although exploiting only a small part of the possibilities offered by the framework of MLNs, already demonstrates how combining learning, logic and probabilities would help overcome some of the shortcomings of current approaches for music processing outlined in section 2.

3.3.3.1. Robustness, flexibility, compactness

In this article, we have discussed the necessity of being able to handle both uncertainty and complex relational structure, stressing the complementarity between probabilistic graphical models and logic, and the current inability to have models that are both robust, flexible and expressive (see section 2.1.1). MLNs allow handling these aspects together. Complex graphical models can be specified compactly in Markov logic. For instance, in **Figure 11**, formulas F_1 and F_2 compactly encode the relational structure of a HMM. This model can then easily be extended to combine various kind of harmony-related information at various time-scales in a single unified formalism, by adding logical formulas and corresponding weights.

Compared to pure logic, weighted logic allows dealing directly with the uncertainty of audio, without the need of an intermediate transcription step. Also, it is not necessary to have absolute correctness and completeness of logical formulas, which makes writing large knowledge base that reflects real world data often an impossible procedure in pure logic. MLNs can handle inconsistencies, incompleteness and contradictions between formulas.

3.3.3.2. Expressiveness

We have also previously underlined the benefit of being able to model music rules in an intuitive and human-readable way, so that developers, researchers and end-users can easily incorporate complex knowledge from diverse sources. MLNs allow such possibility. For instance, it is possible to annotate manually only a few chords, say 10%: $c_0^{\mathrm{A}}, c_9^{\mathrm{A}}, c_{19}^{\mathrm{A}}, \cdots$, and add this evidence to the model given by F_1 and F_2, by adding evidence predicates of the form: $State(c_0^{\mathrm{A}}, 0)$, $State(c_9^{\mathrm{A}}, 9)$, $State(c_{19}^{\mathrm{A}}, 19)$, $\cdots$. This scenario tested on the model described by formulas F_1 and F_2 on the *Fall out boy* song *This ain't a scene its an arms race* results in an increase in the chord estimation results from 60.5 to 76.2%, showing how additional evidence can easily be added and have a significant impact.

3.3.3.3. Multiple abstraction levels, multiple time scales

We have also emphasized the need to be able to reason jointly at multiple abstraction levels and multiple time scales (see section 2.2). The model depicted in **Figure 11** combines various kinds of harmony-related information (chords, global and local keys). Also, it has been possible to incorporate dependencies between music events at various time-scales (beat-synchronous analysis frame, phrase and global structure).

The combination of all the previously described rules results in a single unified formalism for chord estimation at multiple temporal and semantic levels. It is found in Papadopoulos and Tzanetakis (2017) that such multi-faceted and multi time-scale analysis allows significantly improving the results of the various attributes (chords, key, local key) compared to existing more *ad-hoc* approaches. This work is a new step toward a unified multi-scale description of audio, and toward the modeling of complex tasks such as music functional analysis.

3.3.3.4. Other perspectives

Although not yet explored, such a model offers many perspectives for music processing, that include:

- Context information (metrical structure, instrumentation, chord patterns, etc.) could be compactly and flexibly embedded in the model (adding additional weighted logical rules) moving toward a unified analysis of music content.
- It can be extended to handle rich domain knowledge that can be combined with purely empirical learning (Pápai et al., 2012) to help leveraging music complexity.
- It would allow dealing with very large networks since inference and learning can be performed on graphical models with millions of variables and billions of features with highly efficient algorithms that combine probabilistic methods with ideas from logical inference, such as lazy inference methods (Poon et al., 2008) (that takes advantage of the sparseness of relational domain by only grounding atoms that are needed, since only a small fraction of all possible relations are actually true), or lifted inference (Van Haaren et al., 2016) (where queries are answered without materializing all the objects in the domain by grouping random variables that are symmetrical give the first-order structure, and then sampling over the high-level representation).
- It creates the possibility to find useful features and relational structure, and discover new concepts from the data through statistical predicate invention.
- It gives the prospect to use experience in other domains to learn faster in the domain of music processing through deep transfer learning.

3.4. Challenges of StarAI for the MIR Community

The authors hope that the previously outlined work developed in the statistical relational AI area may inspire the MIR community to embrace this path. However there are a few difficulties that may prevent at first glance the MIR researchers to take the plunge in this direction.

First, since StarAI is built on ideas developed within many different fields, it is quite demanding for beginners to acquire the required background. It is necessary to have a combination of competencies from various fields (including for instance machine learning, knowledge representation and logic or probability theory), that often use different views to tackle a given problem, and use different vocabulary and concepts.

Secondly, the newcomers may feel lost face to the plethora of proposed formalisms for statistical relational AI that attempt to unify logic, probability and learning. This (still growing) mass of approaches is often referred to as the *alphabet soup* of statistical

relational AI and it is difficult to understand how they are related or differ from another, and which formalism is more suitable for a given task. There has been some work toward this goal, but a framework for comparing the various existing formalisms is still missing.

Thirdly, although offering many exciting perspectives compared to traditional formalisms, and although having already provided interesting results for many different tasks, it may not be easy to make statistical relational AI approaches work for a given problem, since there are often many subtleties to understand. For instance, even if the principle of Markov logic networks is very simple (expressing knowledge with weighted first-order logical formulas), using the formalism without any understanding of its semantics may result in unexpected results in practice. Indeed, weights do not have a direct correspondence with probabilities, except in some very special cases. However, with the increasing number of people attracted by starAI approaches, knowledge engineering guidelines have been developed (Jain, 2011).

Finally what is maybe lacking now is a generalized symbolic language: low-level features are usually described as numeric values while logic laws are defined with more formal operators. Despite the huge amount of work done in StarAI to permit expressive kind of descriptions, the underlying language and related symbolic machinery are still an open problem. A generalized symbolic language that can be used by users and researchers still needs to be found.

3.5. Potential Perspectives for MIR

The MIR field needs to reason under uncertainty and learning in the presence of data and rich knowledge. Using approaches that combine learning, logic, and probabilities would help leveraging with the complexity of music. In particular it would allow pursuing the five challenges that were stated almost one decade ago (Downie et al., 2010), when reviewing the first 10 years of ISMIR conferences, and in which the MIR community has still not fully engaged yet:

- Dig deeper into the music itself by incorporating multiple features, but in a way that these combinations can be understood in musical terms. StartAI approaches aim at representing complex relational structure at multiple levels of representation.
- Interact with potentials *users* of MIR technology (e.g., musicologists, sound archivists, performing musicians) and get advantage from the multi-disciplinary knowledge they would bring. Putting together several sources of knowledge such as psychoacoustics, cognitive musicology, computational neurobiology, signal processing and machine learning is a key for future development for intelligent machines listening (Cella, 2017). StarAI algorithms are directed toward incorporating such knowledge.
- Expand musical horizons by conducting research on various types of music. Progress toward this goal have already been made by exploring non-Western music traditions such as Indian (Srinivasamurthy et al., 2017) or Chinese (Repetto and Serrá, 2014) music, but other facets of music, such as contemporary classical music remain unexplored, and, as stated above, existing computational models cannot be applied to them. Taking advantage of relational structure learning algorithms, that would allow finding explainable representation from the data, would be of particular interest.
- Boost engagement with data other than audio, such as symbolic data and metadata. We need structured representations that would permit to deal with many different types of knowledge, and use the available large-scale music data that combine multiple data modalities (MIDI, scores, lyrics, user tags etc.).
- Not only focus on subcomponents of MIR systems, but develop full-featured, multifaceted, robust, and scalable real-world-useable systems. We need to develop an end-to-end approach to complexity.

4. CONCLUSIONS

In this article, we have critically looked at the problem of automatic chord estimation from audio recordings as a case study to better understand the current standpoint of content-based MIR research. We have pointed out several deficiencies in current approaches for music processing: the inability to handle simultaneously uncertainty and rich relational structure; the incapacity to handle multiple abstraction levels and the incapability to act on multiple time scales; the unemployment of available multimodal information; and the ineptitude to generalize simplified problems to complex tasks.

In a general sense, music processing algorithms must be able to learn, reason logically with complex relational structure, and handle uncertainty. Existing approaches fail to capture these aspects simultaneously. Probabilistic graphical models can handle uncertainty but cannot deal with complex relational structure. Logic can handle the complexity of music but not its uncertainty. Deep architectures are well suited to characterize the hierarchical nature of music, but they cannot explain the concepts they discover. There have been recently attempts of integrating all these aspects into a common framework.

Unifying logic learning and probability has been a long-standing goal for Artificial Intelligence. This allows "classical AI," based predominantly on first-order logic, to deal with uncertainty and learn from real data, while "modern AI," based mostly on probability theory, can acquire enough expressive power to handle complex relational domains and incorporate prior knowledge (Russell, 2014). Today, the field of StarAI has opened several promising directions toward this goal. We believe that the combination of logic, probability, and learning will allow deploying end-to-end systems for music processing that are expressive enough to represent and learn in a human-readable way complex relational knowledge, and to discover explainable representations. These representations should be able to incorporate rich prior knowledge of various types and interact with heterogeneous data. Moreover they should handle uncertainty and cope

with various forms of noise, imprecision and incomplete information. These approaches would allow efficient and scalable inference, and learning, including learning from other domain. We encourage the MIR community to consider this path.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. Both authors have contributed equally to this work.

REFERENCES

Al Farabi, M. K., Sarkhel, S., and Venugopal, D. (2018). "Efficient weight learning in high-dimensional untied mlns," in *International Conference on Artificial Intelligence and Statistics* (Playa Blanca), 1637–1645.

Anglade, A., Benetos, E., Mauch, M., and Dixon, S. (2010). Improving music genre classification using automatically induced harmony rules. *J. New Music Res.* 39, 349–361. doi: 10.1080/09298215.2010.525654

Anglade, A., and Dixon, S. (2008a). "Characterisation of harmony with inductive logic programming," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (Philadelphia, PA), 63–68.

Anglade, A., and Dixon, S. (2008b). "Towards logic-based representations of musical harmony for classification, retrieval and knowledge discovery," in *MML* (Helsinki).

Anglade, A., Ramirez, R., and Dixon, S. (2009). "Genre classification using harmony rules induced from automatic chord transcriptions," in *International Society of Music Information Retrieval* (Kobe).

Arabi, A. F., and Lu, G. (2009). "Enhanced polyphonic music genre classification using high level features," in *ICSIPA* (Kuala Lumpur), 101–106.

Aucouturier, J., and Pachet, F. (2004). "Improving timbre similarity: how high is the sky?," in *Journal of Negative Results in Speech and Audio Sciences*, 1.

Bartsch, M., and Wakefield, G. (2001). "To catch a chorus using chroma-based representations for audio thumbnailing," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, NY), 15–18.

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Trans. Audio Speech Lang. Process.* 13, 1035–1047. doi: 10.1109/TSA.2005.851998

Bello, J., and Pickens, J. (2005). "A robust mid-level representation for harmonic content in music signal," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (London).

Bengio, Y. (2016). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Bergmann, M. (2008). *An Introduction to Many-Valued and Fuzzy Logic: Semantics, Algebras, and Derivation Systems*. Cambridge University Press.

Besold, T. R., d'Avila Garcez, A., and Lamb, L. (2017). Human-like neural-symbolic computing (Dagstuhl Seminar 17192). *Dagstuhl Rep.* 7, 56–83.

Blockeel, H. (2013). "Statistical relational learning," in *Handbook on Neural Information Processing*, 241–281.

Böck, S., Krebs, F., and Widmer, G. (2016). "Joint beat and downbeat tracking with recurrent neural networks," in *ISMIR* (New York City).

Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. (2013). "Audio chord recognition with recurrent neural networks," in *ISMIR* (Curitiba), 335–340.

Burgoyne, J., Pugin, L., Kereliuk, C., and Fujinaga, I. (2007). "A cross validated study of modeling strategies for automatic chord recognition in audio," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (Vienna), 251–254.

Burgoyne, J., and Saul, L. (2005). "Learning harmonic relationships in digital audio with dirichlet-based hidden markov models," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (London).

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: current directions and future challenges. *Proc. IEEE* 96, 668–696. doi: 10.1109/JPROC.2008.916370

Cella, C. (2017). "Machine listening intelligence," in *In Proceedings of the First International Workshop on Deep Learning and Music Joint With IJCNN Anchorage* (Anchorage), 50–55.

Cella, C.-E. (2018). *Orchidea: Intelligent Assisted Orchestration*. Available online at: http://www.carminecella.com/orchidea/ (accessed April, 2019)

Cho, T., and Bello, J. (2014). On the relative importance of individual components of chord recognition systems. *IEEE Trans. Audio Speech Lang. Process.* 22, 477–492. doi: 10.1109/TASLP.2013.2295926

Crane, R., and McDowell, L. (2012). "Investigating markov logic networks for collective classification," in *ICAART* (Vilamoura).

Dannenberg, R. (2005). "Toward automated holistic beat tracking, music analysis, and understanding," in *ISMIR* (London).

d'Avila Garcez, A., Gori, M., Hitzler, P., and Lamb, L. (2015). Neural-symbolic learning and reasoning (Dagstuhl Seminar 14381). *Dagstuhl Rep.* 4, 50–84.

Davis, J., and Domingos, P. (2009). "Deep transfer via second-order markov logic," in *ICML* (Montreal, QC).

De Raedt, L. (2008). *Logical and Relational Learning: From ILP to MRDM (Cognitive Technologies)*. Secaucus, NJ: Springer-Verlag New York, Inc.

De Raedt, L., Kersting, K., Natarajan, S., and Poole, D. (2016). *Statistical Relational Artificial Intelligence: Logic, Probability, and Computation*, *Volume 32 of Synthesis Lectures on Artificial Intelligence and Machine Learning*. San Rafael, CA: Morgan & Claypool Publishers.

De Raedt, L., and Kimmig, A. (2015). Probabilistic (logic) programming concepts. *Mach. Learn.* 100, 5–47. doi: 10.1007/s10994-015-5494-z

De Raedt, L., Kimmig, A., and Toivonen, H. (2007). "ProbLog: a probabilistic Prolog and its application in link discovery," in *IJCAI* (Hyderabad), 2462–2467.

De Raedt, L., Kimmig, A., and Toivonen, H.(2007). "ProbLog: a probabilistic prolog and its application in link discovery," in *IJCAI* (Hyderabad), 2462–2467.

Deng, J., and Kwok, Y.-K. (2016). "A hybrid gaussian-hmm-deep-learning approach for automatic chord estimation with very large vocabulary," in *ISMIR*.

Deng, J., and Kwok, Y.-K. (2017). Large vocabulary automatic chord estimation using deep neural nets: design framework, system variations and limitations. *arXiv [Preprint]. arXiv:1709.07153*.

Dobrian, C. (1993). *Music and Artificial Intelligence*. Published on Internet.

Domingos, P., and Lowd, D. (2009). *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan & Claypool.

Donadello, I., Serafini, L., and d'Avila Garcez, A. (2017a). "Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation," in *SAC* (Marrakech: ACM), 125–130.

Donadello, I., Serafini, L., and d'Avila Garcez, A. (2017b). "Logic tensor networks for semantic image interpretation," in *International Joint Conferences on Artificial Intelligence* (Melbourne), 1596–1602.

Dovey, M., Lavrac, N., and Wrobel, S. (1995). *Analysis of Rachmaninoff's piano Performances Using Inductive Logic Programming (Extended abstract)*, Vol. 912. Berlin; Heidelberg: Springer.

Downie, S., Byrd, D., and Crawford, T. (2010). "Ten years of ismir: reflections on challenges and opportunities," in *International Society for Music Information Retrieval* (Utrecht), 13–18.

Ellis, D. (2007a). Beat tracking by dynamic programming. *J. New Music Res.* 36, 51–60. doi: 10.1080/09298210701653344

Ellis, D. (2007b). "Classifying music audio with timbral and chroma features," in *Austrian Computer Society (OCG)*.

Ellis, D., and Rosenthal, D. (1995). "Mid-level representations for computational auditory scene analysis," in *International Joint Conferences on Artificial Intelligence* (Montreal, QC).

Flach, P. (1994). *Simply Logical - Intelligent Reasoning by Example*. Wiley Professional Computing. John Wiley.

Foote, J. (1997). "Content-based retrieval of music and audio," in *Proceedings of the SPIE 3229, Multimedia Storage and Archiving Systems II*, 138–147.

Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). "Learning probabilistic relational models," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (Stockholm), 1300–1309.

Fujishima, T. (1999). "Real-time chord recognition of musical sound: a system using common lisp music," in *International Catholic Migration Commission* (Beijing).

Gaudefroy, C., Papadopoulos, H., and Kowalski, M. (2015). "A multi-dimensional meter-adaptive method for automatic segmentation of music," in *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)* (Prague), 1–6.

Getoor, L. (2001). *Learning Statistical Models from Relational Data*. PhD thesis, Stanford, CA.

Getoor, L., and Taskar, B. (2007). *Introduction to Statistical Relational Learning*. Adaptive Computation and Machine Learning. The MIT Press, 608.

Grosche, P., Müller, M., and Serrà, J. (2012). "Audio content-based music retrieval," in *Multimodal Music Processing*, eds M. Müller, M. Goto, and M. Schedl (Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum für Informatik), 157–174.

Gurevych, I., Meyer, C., Binnig, C., Fürnkranz, J., Roth, S., and Simpson, E. (2018). "Interactive data analytics for the humanities," in *Computational Linguistics and Intelligent Text Processing: Proceedings of the 18th International Conference. Part I, Volume 10761 of Lecture Notes in Computer Science*, ed A. Gelbukh (Cham: Springer), 527–549.

Haack, S. (1978). *Philosophy of Logics*. New York, NY: Cambridge University Press.

Hamel, P., Davies, M., Yoshii, K., and Goto, M. (2013). "Transfer learning in MIR: sharing learned latent representations for music audio classification and similarity," in *International Society for Music Information Retrieval* (Curitiba).

Harte, C. (2010). *Towards Automatic Extraction of Harmony Information From Music Signals*. PhD thesis, University of London.

Herremans, D., and Chuan, C. H. (2017). "A multi-modal platform for semantic music analysis: visualizing audio-and score-based tension," in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 419–426.

Humphrey, E., and Bello, J. (2012). "Rethinking automatic chord recognition with convolutional neural networks," in *2012 11th International Conference on Machine Learning and Applications*, Vol. 2, 357–362.

Humphrey, E., and Bello, J. (2015). "Four timely insights on automatic chord estimation," in *International Society for Music Information Retrieval* (Malaga), 673–679.

Humphrey, E., Bello, J., and LeCun, Y. (2013). Feature learning and deep architectures: new directions for music informatics. *J. Intell. Inform. Syst.* 41, 461–481. doi: 10.1007/s10844-013-0248-5

Jain, D. (2011). *Knowledge Engineering with Markov Logic Networks: A Review*. DKB.

Jain, D., von Gleissenthall, K., and Beetz, M. (2011). "Bayesian logic networks and the search for samples with backward simulation and abstract constraint learning," in *KI 2011: Advances in Artificial Intelligence, 34th Annual German Conference on AI, Volume 7006 of Lecture Notes in Computer Science* (Springer), 144–156.

Jernite, Y., Rush, A. M., and Sontag, D. (2015). "A fast variational approach for learning markov random field language models," in *International Conference on Machine Learning* (Corvallis), 2209–2216.

Kameoka, H., Ochiai, K., Nakano, M., Tsuchiya, M., and Sagayama, S. (2012). "Knowledge engineering with markov logic networks: a review," in *International Society for Music Information Retrieval* (Porto).

Kempf, D. (1996). What is symmetry in music? *Int. Rev. Aesthet. Sociol. Music* 27:155. doi: 10.2307/3108344

Kernfeld, B. (2007). "Blues progression," in *The New Grove Dictionary of Jazz, 2nd Edn.* Oxford: Oxford University Press.

Kersting, K., and de Raedt, L. (2007). "Bayesian logic programming: theory and tool," in *Introduction to Statistical Relational Learning,* eds L. Getoor and B. Taskar (Cambridge: MIT Press), 291–322.

Kersting, K., and Raedt, L. D. (2001). "Towards combining inductive logic programming with bayesian networks," in *ILP '01 Proceedings of the 11th International Conference on Inductive Logic Programming,* eds C. Rouveirol and M. Sebag (Berlin: Springer), 118–131.

Kim, J., Urbano, J., Liem, C. C. S., and Hanjalic, A. (2018). One deep music representation to rule them all?: a comparative analysis of different representation learning strategies. *arXiv [Preprint]. arXiv:1802.04051*.

Kimmig, A., Mihalkova, L., and Getoor, L. (2015). Lifted graphical models: a survey. *Mach. Learn.* 99, 1–45. doi: 10.1007/s10994-014-5443-2

Kindermann, R., and Snell, J. L. (1980). *Markov Random Fields and Their Applications*. Providence, RI: American Mathematical Society, 142.

Kok, S., and Domingos, P. (2005). "Learning the structure of markov logic networks," in *International Conference on Machine Learning* (Bonn), 441–448.

Kok, S., and Domingos, P. (2007). "Statistical predicate invention," in *International Conference on Machine Learning* (Corvallis).

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models - Principles and Techniques*. MIT Press. I–XXXV, 1–1231.

Koops, H., de Haas, W., Bransen, J., and Volk, A. (2017). "Chord label personalization through deep learning of integrated harmonic interval-based representations," in *Proceedings of the First International Conference on Deep Learning and Music, US, May, 2017* (Anchorage, AK).

Korzeniowski, F., Sears, D., and Widmer, G. (2018). "A large-scale study of language models for chord prediction," in *ICASSP* (Calgary).

Korzeniowski, F., and Widmer, G. (2016a). "Feature learning for chord recognition: the deep chroma extractor," in *International Society for Music Information Retrieval* (New York City).

Korzeniowski, F., and Widmer, G. (2016b). "A fully convolutional deep auditory model for musical chord recognition," in *MLSP* (Salerno).

Korzeniowski, F., and Widmer, G. (2017). "On the futility of learning complex frame-level language models for chord recognition," in *AES International Conference on Semantic Audio*.

Krumhansl, C. (1990). *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Kuzelka, O., Davis, J., and Schockaert, S. (2016). Constructing markov logic networks from first-order default rules. *Induct. Logic Program.* 9575, 91–105. doi: 10.1007/978-3-319-40566-7_7

Lafferty, J. (2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning* (Morgan Kaufmann), 282–289.

Lee, K., and Slaney, M. (2008). "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," in *IEEE Trans. Audio Speech Lang. Proc.* 16, 291–301.

Leivant, D. (1994). *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume2, Deduction Methodologies*. Clarendon Press. 229–322.

Lew, M. (2006). Content-based multimedia information retrieval: state of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.* 2, 1–19. doi: 10.1145/1126004.1126005

Lewin, D. (1987). *Generalized Musical Intervals and Transformations*. New Haven, CT: Yale University Press.

Liu, Y., Carbonell, J., Weigele, P., and Gopalakrishnan, V. (2005). "Segmentation conditional random fields (scrfs): a new approach for protein fold recognition," in *Research in Computational Molecular Biology*, eds S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, and M. Waterman (Berlin; Heidelberg: Springer), 408–422.

Lostanlen, V., and Cella, C.-E. (2016). "Deep convolutional networks on the pitch spiral for music instrument recognition," in *International Society for Music Information Retrieval* (New York City).

Malkin, R. (2006). *Machine Listening for Context-Aware Computing*. PhD thesis, Language Technologies, Institute School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Mallat, S. (2016). Understanding deep convolutional networks. *Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci.* 374:2065. doi: 10.1098/rsta.2015.0203

Mallory, E., Zhang, C., Ré, C., and Altman, R. B. (2016). Large-scale extraction of gene interactions from full-text literature using deepdive. *Bioinformatics* 32, 106–113.

Maresz, Y. (2013). On computer-assisted orchestration. *Contemp. Music Rev.* 32, 99–109. doi: 10.1080/07494467.2013.774515

Marsík, L., Rusek, M., Slaninová, K., Martinovic, J., and Pokorný, J. (2017). "Evaluation of chord and chroma features and dynamic time warping scores on cover song identification task," in *Computer Information Systems and Industrial Management Applications* (Bialystok), 205–217.

Mauch, M., and Dixon, S. (2010). Simultaneous estimation of chords and musical context from audio. *IEEE Trans. Audio Speech Lang. Process.* 18, 1280–1289. doi: 10.1109/TASL.2009.2032947

Mauch, M., Noland, K., and Dixon, S. (2009). "Using musical structure to enhance automatic chord transcription," in *International Society for Music Information Retrieval* (Kobe).

McFee, B., and Bello, J. (2017). "Structured training for large-vocabulary chord recognition," in *International Society for Music Information Retrieval* (Suzhou), 188–194.

McVicar, M., Santos-Rodriguez, R., Ni, Y., and Bie, T. D. (2014). Automatic chord estimation from audio: a review of the state of the art. *IEEE Trans. Audio Speech Lang. Process.* 22, 556–575. doi: 10.1109/TASLP.2013.2294580

Mihalkova, L., Huynh, T., and Mooney, R. (2007). "Mapping and revising markov logic networks for transfer learning," in *Association for the Advancement of Artificial Intelligence* (Vancouver, BC).

Minsky, M. L., and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press. 308.

Mishkin, H. (1978). Schubert's last year, 1828. *Massachusetts Rev.* 19, 229–244.

Morales, E. (1997). Pal: a pattern-based first-order inductive system. *Mach. Learn.* 26, 227–252. doi: 10.1023/A:1007373508948

Morales, E., and Morales, R. (1995). "Learning musical rules," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (Montreal, QC), 81–85.

Muggleton, S. (1991). Inductive logic programming. *New Generat. Comput.* 8, 295–318. doi: 10.1007/BF03037089

Müller, M., Chew, E., and Bello, J. (2016). Computational music structure analysis (Dagstuhl Seminar 16092). *Dagstuhl Rep.* 6, 147–190.

Muller, M., Ellis, D., Klapuri, A., and Richard, G. (2011). Signal processing for music analysis. *IEEE J. Select. Top. Signal Process.* 5, 1088–1110. doi: 10.1109/JSTSP.2011.2112333

Müller, M., Goto, M., and Schedl, M., editors (2012). *Multimodal Music Processing, Volume 3 of Dagstuhl Follow-Ups*. Dagstuhl: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press.

Ni, Y., McVicar, M., Santos-Rodríguez, R., and Bie, T. D. (2013). Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Trans. Audio Speech Lang. Process.* 21, 2607–2615. doi: 10.1109/TASL.2013.2280218

Nilsson, N. (1986). Probabilistic logic. *Artif. Intell.* 28, 71–87. doi: 10.1016/0004-3702(86)90031-7

Ojima, Y., Nakano, T., Fukayama, S., Kato, J., Goto, M., Itoyama, K., et al. (2017). "A singing instrument for real-time vocal-part arrangement of music audio signals," in *IEEE International Conference on Systems, Man, and Cybernetics* (Espoo), 443–449.

Orio, N., Rizo, D., Miotto, R., Schedl, M., Montecchio, N., and Lartillot, O. (2011). "Musiclef: a benchmark activity in multimodal music information retrieval," in *International Society for Music Information Retrieval* (Miami, FL).

Oudre, L., Grenier, Y., and Févotte, C. (2009). "Template-based chord recognition: influence of the chord types," in *International Society for Music Information Retrieval* (Kobe), 153–158.

Pachet, F. (2016). A joyful ode to automatic orchestration. *ACM Trans. Intell. Syst. Technol.* 8, 18:1–18:13.

Paiement, J.-F., Eck, D., Bengio, S., and Barber, D. (2005). "A graphical model for chord progressions embedded in a psychoacoustic space," in *Proceedings of the International Conference on Machine Learning (ICML)* (Bonn), 641–648.

Pan, S., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Papadopoulos, H. (2010). *Joint Estimation of Musical Content Information From an Audio Signal*. PhD thesis, University Paris 6, Paris.

Papadopoulos, H., and Peeters, G. (2007). "Large-scale study of chord estimation algorithms based on chroma representation and HMM," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI)* (Bordeaux).

Papadopoulos, H., and Peeters, G. (2011). Joint estimation of chords and downbeats. *IEEE Trans. Audio Speech Lang. Process.* 19, 138–152. doi: 10.1109/TASL.2010.2045236

Papadopoulos, H., and Tzanetakis, G. (2013). "Exploiting structural relationships in audio signals of music using markov logic," in *International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC).

Papadopoulos, H., and Tzanetakis, G. (2017). Models for music analysis from a markov logic networks perspective. *IEEE Trans. Audio Speech Lang. Process.* 25, 19–34. doi: 10.1109/TASLP.2016.2614351

Pápai, T., Ghosh, S., and Kautz, H. (2012). "Combining subjective probabilities and data in training markov logic networks," in *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I, ECML PKDD'12* (Berlin; Heidelberg: Springer-Verlag), 90–105.

Papai, T., Kautz, H., and Stefankovic, D. (2012). "Slice normalized dynamic markov logic networks," in *Advances in Neural Information Processing Systems 25*, eds P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, 1916–1924.

Paulus, J., Müler, M., and Klapuri, A. (2010). "State of the art report: audio-based music structure analysis," in *International Society for Music Information Retrieval* (Utrecht).

Pauwels, J., and Peeters, G. (2013). "Segmenting music through the joint estimation of keys, chords and structural boundaries," in *MM* (France), 741–744.

Pawar, S., Bhattacharyya, P., and Palshikar, G. K. (2017). "End-to-end relation extraction using markov logic networks," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*, 818–827.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann.

Pereira, R., and Silla, C. (2017). "Using simplified chords sequences to classify songs genres," in *IEEE International Conference on Multimedia and Expo* (Hong Kong), 1446–1451.

Poole, D. (2003). "First-order probabilistic inference," in *International Joint Conferences on Artificial Intelligence* (Acapulco).

Poon, H. (2011). *Markov Logic for Machine Reading*. PhD thesis, University of Washington.

Poon, H., Domingos, D., and Sumner, M. (2008). "A general method for reducing the complexity of relational inference and its application to MCMC," in *AAAI* (Chicago, IL).

Poon, H., and Domingos, P. (2010). "Machine reading: a "killer app" for statistical relational ai," in *AAAI* (Atlanta), 76–81.

Prince, J., Schmuckler, M. A., and Thompson, W. F. (2009). The effect of task and pitch structure on pitch-time interactions in music. *Mem. Cogn.* 37, 368–381. doi: 10.3758/MC.37.3.368

Raedt, L. D., and Kersting, K. (2008). "Probabilistic inductive logic programming," in *Probabilistic Inductive Logic Programming, Volume 4911 of Lecture Notes in Computer Science*, eds L. De Raedt, P. Frasconi, K. Kersting, and S. Muggleton (Berlin; Heidelberg: Springer), 1–27.

Rameau, J. P., and Gossett, P. (1971). *Treatise on Harmony*. Translated with an introduction and notes by Philip Gossett (Original French version first published in 1722). Dover Publications.

Ramirez, R., Hazan, A., Maestre, E., Serra, X., Petrushin, V., and Khan, L. (2007). *A Data Mining Approach to Expressive Music Performance Modeling*. London: Springer.

Ramirez, R., and Palamidessi, C. (2003). *Inducing Musical Rules with ILP*, Vol. 2916. Berlin; Heidelberg: Springer.

Repetto, R. C., and Serrá, X. (2014). "Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis," in *International Society for Music Information Retrieval* (Taipei).

Richardson, M. (2004). *Learning and Inference in Collective Knowledge Bases*. PhD thesis, University of Washington.

Richardson, M., and Domingos, P. (2006). Markov logic networks. *Mach. Learn.* 62, 107–136. doi: 10.1007/s10994-006-5833-1

Riedel, S., and Meza-Ruiz, I. (2008). "Collective semantic role labelling with markov logic," in *CoNLL* (Manchester, UK).

Riemann, H. (1896). *Harmony Simplified: Or the Theory of the Tonal Functions of Chords* (Augener).

Russell, S. J. (2014). "Unifying Logic and Probability: A New Dawn for AI?," in *IPMU (1), Communications in Computer and Information Science* (Cham: Springer) 10–14.

Salamon, J., Gómez, E., Ellis, D. P. W., and Richard, G. (2013). "Melody extraction from polyphonic music signals: approaches, applications and challenges," in *IEEE Signal Processing Magazine,* 118–134.

Sarkhel, S., Venugopal, D., Ruozzi, N., and Gogate, V. (2017). "Efficient inference for untied mlns," in *International Joint Conferences on Artificial Intelligence* (Melbourne), 4617–4624.

Schedl, M., Gómez, E., and Urbano, J. (2014). Music information retrieval: recent developments and applications. *J. Found. Trends Inform. Retrieval* 8, 127–261. doi: 10.1561/1500000042

Schedl, M., Knees, P., and Gouyon, F. (2017). "New paths in music recommender systems research," in *RecSys* (Como), 392–393.

Schoenberg, A. (1969). *Structural Functions in Harmony*. New York, NY: Norton.

Schuller, B. (2013). "Applications in intelligent music analysis," in *Intelligent Audio Analysis*, Signals and Communication Technology (Berlin; Heidelberg: Springer) 225–298.

Serafini, L., and d'Avila Garcez, A. (2016). "Logic tensor networks: deep learning and logical reasoning from data and knowledge," in *CEUR Workshop Proceedings,* CEUR-WS.org

Serrà, J. (2011). *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*. PhD thesis, Universitat Pompeu Fabra, Barcelona.

Serrá, J., Gómez, E., and Herrera, P. (2010). Audio cover song identification and similarity: background, approaches, evaluation, and beyond. *Stud. Comput. Intell.* 274, 307–332. doi: 10.1007/978-3-642-11674-2_14

Sheh, A., and Ellis, D. (2003). "Chord segmentation and recognition using EM-trained HMM," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (Baltimore).

Shenoy, A., and Wang, Y. (2005). Key, chord and rhythm tracking of popular music recordings. *Comput. Music J.* 3, 75–86. doi: 10.1162/0148926054798205

Sigtia, S., Boulanger-Lewandowski, N., and Dixon, S. (2015). "Audio chord recognition with a hybrid recurrent neural network," in *International Society for Music Information Retrieval* (Malaga), 127–133.

Singla, P., and Domingos, P. (2006). "Memory-efficient inference in relational domains," in *Association for the Advancement of Artificial Intelligence* (Palo Alto, FL).

Smith, J. B. L., Hamasaki, M., and Goto, M. (2017). "Classifying derivative works with search, text, audio and video features," in *IEEE International Conference on Multimedia and Expo* (Hong Kong), 1422–1427.

Snidaro, L., Visentini, I., and Bryan, K. (2015). Fusing uncertain knowledge and evidence for maritime situational awareness via markov logic networks. *Inform. Fusion* 21, 159–172. doi: 10.1016/j.inffus.2013.03.004

Socher, R., Chen, D., Manning, C., and Ng, A. (2013). "Reasoning with neural tensor networks for knowledge base completion," in *Conference on Neural Information Processing Systems* (Lake Tahoe, CA), 926–934.

Šourek, G., Aschenbrenner, V., Železny, F., and Kuželka, O. (2015). "Lifted relational neural networks," in *COCO*, Vol. 1583 (Montreal, QC), 52–60.

Srinivasamurthy, A., Holzapfel, A., Ganguli, K. K., and Serrá, X. (2017). Aspects of tempo and rhythmic elaboration in hindustani music: a corpus study. *Front. Digit. Humanit.* 4:20. doi: 10.3389/fdigh.2017.00020

Sutton, C., and McCallum, A. (2007). "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning,* eds L. Getoor and B. Taskar (Cambridge: MIT Press), 1–35.

Sztyler, T., Civitarese, G., and Stuckenschmidt, H. (2018). "Modeling and reasoning with problog: an application in recognizing complex activities," in *PerCom* (Athens), 220–225.

Thimm, M. (2016). *Uncertainty and Inconsistency in Knowledge Representation*. Universität Koblenz-Landau, Department of Computer Science, Institute for Web Science and Technologies. Habilitation Thesis.

Tsushima, H., Nakamura, E., Itoyama, K., and Yoshii, K. (2018). Generative statistical models with self-emergent grammar of chord sequences. *J. New Music Res.* 47, 226–248. doi: 10.1080/09298215.2018.1447584

Van Baelen, E., Raedt, L. D., and Muggleton, S. (1997). *Analysis and Prediction of Piano Performances Using Inductive Logic Programming*, Vol. 1314, 55–71. Berlin; Heidelberg: Springer.

Van Haaren, J., Van den Broeckn, G., Wannes Meert, W., and Davis, J. (2016). Lifted generative learning of markov logic networks. *Mach. Learn.* 103:27–55. doi: 10.1007/s10994-015-5532-x

Venugopal, D. (2015). "Scaling-up inference in markov logic," in *AAAI* (Austin, TX), 4259–4260.

Wang, H., and Poon, H. (2018). "Deep probabilistic logic: a unifying framework for indirect supervision," in *Empirical Methods in Natural Language Processing* (Brussels).

Widmer, G. (2003). Discovering simple rules in complex data: a meta-learning algorithm and some surprising musical discoveries. *Artif. Intell.* 146, 129–148. doi: 10.1016/S0004-3702(03)00016-X

Wu, Y., and Li, W. (2018). "Music chord recognition based on midi-trained deep feature and blstm-crf hybird decoding," in *International Conference on Acoustics, Speech, and Signal Processing*.

Zalkow, F., Weiß, C., and Müller, M. (2017). "Exploring tonal-dramatic relationships in richard Wagner's ring cycle," in *International Society for Music Information Retrieval* (Suzhou), 642–648.

Zhou, X., and Lerch, A. (2015). "Chord detection using deep learning," in *International Society for Music Information Retrieval* (Malaga), 52–58.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.