Haizi Yu | University of Chicago, Chicago, IL 60637 USA | E-mail: haiziyu@uchicago.edu
Lav R. Varshney, Senior Member, IEEE | University of Illinois Urbana-Champaign, Urbana, IL 61801 USA | E-mail: varshney@illinois.edu
Heinrich Taube | University of Illinois Urbana-Champaign, Urbana, IL 61801 USA | E-mail: taube@illinois.edu
James A. Evans | University of Chicago, Chicago, IL 60637 USA | E-mail: jevans@uchicago.edu

(Re)discovering Laws of Music Theory Using Information Lattice Learning

Abstract—Information lattice learning (ILL) is a novel framework for knowledge discovery based on group-theoretic and information-theoretic foundations, which can rediscover the rules of music as known in the canon of music theory and also discover new rules that have remained unexamined. Such probabilistic rules are further demonstrated to be human-interpretable. ILL itself is a rediscovery and generalization of Shannon's lattice theory of information, where probability measures are not given but are learned from training data. This article explains the basics of the ILL framework, including both how to construct a lattice-structured abstraction universe that specifies the structural possibilities of rules, and how to find the most informative rules by performing statistical learning through an iterative student–teacher algorithmic architecture that optimizes information functionals. The ILL framework is finally shown to support both pedagogy and novel patterns of music co-creativity.

Introduction

Is it possible for an artificial intelligence (AI) system to learn the laws of music theory in the same human-interpretable form as a textbook? How little prior knowledge and how little data is needed to do so? Do novel conceptual discoveries also emerge, or is the system restricted to rediscovery (where algorithm designers' conscious and unconscious biases might have helped [1])? Can the underlying approach to such knowledge discovery also provide representations that make it easy for people to decompose and recompose novel music as a form of social co-creativity?

At the Royal Society on September 27, 1950, Claude Shannon presented "The Lattice Theory of Information" as part of the first London Symposium on Information Theory [2]. He aimed to describe the fundamental nature of information beyond just characterizing its amount, as in his seminal 1948 paper [3]. The basic idea he developed was that all translations or ways of describing the same information should be regarded as equivalent. Initially, unbeknownst to us, we discovered a generalization of Shannon's lattice theory of information in the context of music intelligence [4], [5], [6], and have used it to build AI systems that can indeed rediscover much of the music theory curriculum at the University of Illinois Urbana-Champaign in the same basic form as a textbook [7], [8], [9], find new laws of music that music theorists find compelling [9], and support a co-creativity platform we are building to compose completely new music via a novel language of music fragments [10]. Moreover, all of this can be done with no musical knowledge built in (just universal priors consistent with human innate cognition—the Core Knowledge priors in cognitive science [11]), and on the basis of just the sheet music from 370 chorales by Johann Sebastian Bach, a German composer and musician of the late Baroque period.

In a sense, the knowledge discovery algorithms we develop in our *information lattice learning* (ILL) framework yield results that parallel the celebrated theoretical book *Gradus ad Parnassum* (1725) by Johann Joseph Fux, an Austrian music theorist and pedagogue of the late Baroque period. Since the ILL approach is directly human-interpretable (neither inscrutable nor nonintuitive [12]), it can also be used to teach people music theory. Indeed, the compositional rules that emerge by distilling sheet music can be used to deliver personalized lessons on

Digital Object Identifier 10.1109/MBITS.2022.3205288
Date of publication 9 September 2022; date of current version 8 December 2022.

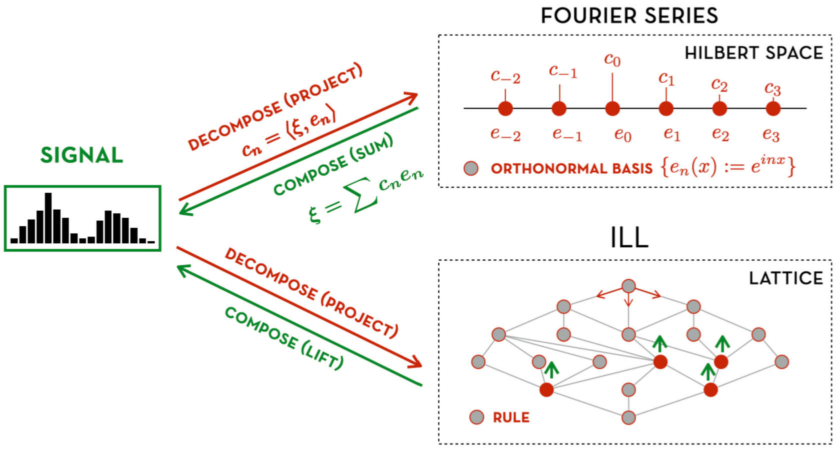

FOURIER SERIES
HILBERT SPACE
c_{-2} c_{-1} c_0 c_1 c_2 c_3
e_{-2} e_{-1} e_0 e_1 e_2 e_3
ORTHONORMAL BASIS $\{e_n(x) := e^{inx}\}$
SIGNAL
DECOMPOSE (PROJECT)
$c_n = \langle \xi, e_n \rangle$
COMPOSE (SUM)
$\xi = \sum c_n e_n$
ILL
LATTICE
DECOMPOSE (PROJECT)
COMPOSE (LIFT)
RULE

Figure 1

Fourier analysis decomposes a signal into simple components in an orthogonal basis with coefficients $\{c_n\}$ whereas ILL breaks a signal into human-interpretable rules that are part of a lattice-structured human-interpretable hierarchy.

music composition, to enhance content-based music search, to support creative music composition, and to discover new music knowledge about different styles/genres. Interestingly, Shannon himself was both very musical (e.g., playing jazz clarinet in Greenwich Village [13]) and believed “the most promising new developments in information theory will come from work on very complex machines, especially from research into artificial intelligence” [14]. So it is very appropriate that we find information lattices arising in an automatic music theorist, pedagogue, and composer.

Musicology is the scholarly analysis and research-based study of music, and its subdisciplines of historical musicology and ethnomusicology tend to focus on performances, traditions, genres, and the people who produce and engage with them, such as musicians and composers. Systematic musicology, on the other hand, refers to these specific realizations but tends to focus on more general questions about music by analyzing empirical data and developing theory. Indeed, music theory is often cast as the study of possibilities in music. As noted in *The Oxford Companion to Music*, music theory includes rudiments of music notation, scholars’ views on music from antiquity to the present, and definitions of processes and general principles in music. The high-level idea is that the starting point is not individual works or performances, but the “fundamental materials from which it is built” [15]. There has been little prior work on automating music theory, besides our own [16]. Indeed work in AI for automatic theory development has focused on physical sciences, cf., [1], [17], [18], rather than humanities or social sciences (which some regard as more complicated). Notwithstanding, much of music theory is quite mathematical, drawing on statistical and information-theoretic concepts, together with geometric, topological, and algebraic ones.

Whether in music or in these other domains (like chemistry [9], genetics [19], and quantum physics), ILL aims to solve the basic question of what makes a given object that object, as a form of conceptual knowledge discovery. It emphasizes the *what* in explaining the object rather than focusing on the ability to generate similar objects as in generative modeling or predicting labels for the object as in supervised learning. ILL explains a signal to people via human-like and human-interpretable abstractions of that signal called *rules*. As a signal may be viewed from several different perspectives, one should aim to find several rules that collectively explain most of the signal, with each rule explaining a unique aspect. Solving this fundamental question enables rule-based knowledge discovery designed to help people understand complex signals. Note that whereas classic machine learning problems (classification or generation) may sometimes reveal knowledge during the classification or generation process, their central target is not knowledge discovery, nor its interpretability. That is, a music AI that classifies concertos by composer or generates new concertos that mimic a given composer does not necessarily produce human insight about what makes a concerto a concerto or the best rules a novice composer might employ to write one. The rules we seek may be used in classification later, but they are primarily built for understanding. So, instead of optimizing a task-specific objective like classification error, ILL balances among objectives favoring *fewer*, *simpler* rules for interpretability, as well as more *essential* rules for effectiveness.

The mathematical intuition behind ILL is *to break the whole into simple pieces*, somewhat akin to decomposing a signal into its Fourier representation (see Figure 1). Whereas Fourier analysis decomposes a signal in a Hilbert space via the inner product (i.e., projection to orthonormal basis) and synthesizes it via a weighted sum, ILL decomposes a signal in a hierarchical space called a *lattice*. We aim for human-like, hierarchical rule abstraction-and-realization via signal decomposition-and-synthesis in a lattice, called *projection-and-lifting*, resulting in more than the sum of parts.

As noted, the ILL approach generalizes Shannon's original information lattices to a hierarchical distribution of representations and importantly brings statistical learning into the lattice. Shannon's original work (further formalized by Li and Chong [20]) assumed a probability space with a given probability measure, rather than allowing statistical learning. Moreover, numerous hierarchical decompositions of information have been proposed in the literature that all assume a fixed probability measure. Examples include lattices of Huffman codes ordered with respect to code tree imbalance [21]; partition lattices stemming from submodular functions in the context of multivariate information measures [22]; integrated information lattices in the temporal decomposition of information in complex systems [23]; and information hierarchies in economics considering Blackwell partial ordering in decision making [24].

There are two phases to ILL. First, the information lattice (a kind of abstraction universe) is algorithmically constructed from group-theoretic foundations using techniques from computational group theory [6], [25]. Second, the learning algorithm is realized by an iterative discovery cycle that has a student-teacher architecture. The iterative structure is reminiscent of other student-teacher approaches to conceptual learning [26], [27], but specifically operates on the lattice through alternating optimization of information measures.

In the remainder of this article, we will start with an exposition of information lattices and then get into the ILL framework for self-exploratory and self-explanatory AI. After that preparation, we will get into human-interpretable music knowledge discovery, teaching, and creativity.

As a separate line of research, note that there is a significant history of computer-based music composition, largely drawing on ideas from stochastic processes [28]. As an example, Betty Shannon and John Pierce wrote a 1949 Bell Labs technical memorandum on "Composing Music by a Stochastic Process" which (at the time unknowingly) expanded on works by W. A. Mozart, J. Haydn, M. J. K. D. Stadler, and K. P. E. Bach. It introduced stochastic models to describe the generating process of the chord progression in four-part harmonies by using known music theory rules and was implemented using three specially made dice and a table of random numbers [28], [29]. The ILLIAC Suite, a 1957 composition for string quartet coming from a program by Lejaren Hiller and Leonard Issacson for the ILLIAC I computer at the University of Illinois Urbana-Champaign, is generally agreed to be the first score composed by an electronic computer [30]; like the Shannon–Pierce work, it also followed a rule-based Markov generative process. These approaches were highly dependent on the rules used, and further did not capture long-range dependence. Later work in computer music composition focused on imitating particular styles and drew on much more intricate hand-designed rule sets [31]. Modern deep learning methods for music composition, such as the Music Transformer [32], capture much longer ranges of dependence than simple Markov models through their attention mechanisms, but may lack the strong ability for human creative control. Yet, human intentionality is central to creativity [33], [34]. The ILL framework focuses on learning the laws of music theory from data—rather than relying on hand-designed rules—that can then be used in a human-approachable way for music composition [35], whether by stochastically sampling or by much more human-controllable methods of co-creativity.

Information Lattices and Learning

In his 1950 work, Claude Shannon attempted to describe the nature of information beyond just quantifying its amount [2]. With the specific context of communication problems in mind, he developed the term *information element* to denote the nature of information, which is invariant under "(language) translations" or different encoding–decoding schemes. He further introduced a partial order between a pair of information elements, eventually yielding a lattice of information elements, the *information lattice*.

Here, we first briefly review Shannon's original work and then cast the information lattice in our abstraction-generation framework without needing to introduce information-theoretic functionals, such as entropy, or even probability spaces $(\Omega, \mathcal{F}, P)$ with sample spaces Ω, σ-algebras $\mathcal{F}$, and probability measures P. Our abstraction-generation framework for knowledge discovery not only generalizes Shannon's information lattice, but more importantly presents a generating chain that brings learning into the picture. This eventually opens up the opportunity for data-driven concept learning, which aims to discover human-interpretable rules from data.

Theoretical Generalization: A Separation of Clustering and Statistics

Consider Shannon's original work and a follow-up work [20] that formalizes Shannon's idea in a more principled way. Is the nature of information an

$$\text{information element or chance variable?} \tag{1}$$

We say two chance variables are informationally equivalent if they induce the same σ-algebra (of the sample space). An *information element* is an equivalence class of chance variables (of a common sample space) with respect to the "being-informationally-equivalent" relation. Under this definition, the notion of an information element—essentially a probability space—is more abstract than that of a chance variable: an information element can be realized by different chance variables. The relationship between different but informationally equivalent chance variables and their corresponding information element is analogous to the relationship between different faithful translations (say, Chinese and Hindi) of a message and the actual content/meaning of that message. Since different but faithful translations are viewed as different ways of describing the same information, the information itself is then regarded as the equivalence class of all translations or ways of describing the same information. Therefore, the notion of the information element is said to reveal the fundamental nature of the information.

The idea of information lattices can be given a group-theoretic interpretation:

$$\text{information lattice} \rightarrow \text{partition lattice} \rightarrow \text{subgroup lattice } (\rightarrow \text{interpretation}). \tag{2}$$

An *information lattice* is a lattice of information elements, where the partial order is defined by $x \leq y \Leftrightarrow H(x|y) = 0$ where H denotes the conditional Shannon entropy. The join of two information elements $x \vee y = x + y$ is called the *total information* of x and y; the meet of two information elements $x \wedge y = xy$ is called the *common information* of x and y, and is in fact the Gács–Körner common information [20], [36]. By definition, every information element can be uniquely determined by its induced σ-algebra. Also, it is known that every σ-algebra of a countable sample space can be uniquely determined by its generating (via union operation) sample-space-partition. Thus, an information lattice has a one-to-one correspondence to a partition lattice. Further, given a partition of a sample space, [20] constructed a unique permutation subgroup whose group action on the sample space produces orbits that coincide with the given partition. Therefore, under this specific construction, any partition lattice has a one-to-one correspondence to the constructed subgroup lattice (see the General Isomorphism Theorem in [20]). This yields the abovementioned Chain (2) which further achieves group-theoretic interpretations of various information-theoretic results, bringing together information theory and group theory [37].

Now, we cast the abovementioned results into our framework and point out the key differences. Is the nature of abstraction

$$\text{clustering or classification?} \tag{3}$$

Generalizing Shannon's insight on (1) reveals an essential difference between clustering and classification in machine learning. Following the "being-informationally-equivalent" relation, we can similarly define an equivalence relation on the set of all classifications where two classifications are equivalent if they yield the same set of classes and only differ by class labels. For example, given a set of animals, classifying them into {fish, amphibians, reptiles, birds, mammals} is equivalent to classifying them into {poisson, amphibians, reptiles, oiseaux, mammifères}, where the different class labels are only English and French translations of the same animal classes. So, the relationship between clustering and classification is analogous to that between information elements and chance variables. Clustering rather than classification captures the nature of abstraction. This explains why we formalize abstraction and knowledge discovery as a clustering problem in ILL.

TABLE 1. Partition Lattice and Information Lattice: The Main Difference Comes From the Fact that a Partition Lattice is not Coupled with a Measure; Whereas an Information Lattice is Coupled with a Probability Measure, so Both the Partial Order and the Metric can be Defined in Terms of Entropies

	Partition lattice	*Information lattice*
Element	Partition $(\mathcal{P})$;	Information element (x);
	Clustering $(X, \mathcal{P})$;	Probability space (X, Σ, P);
	Equiv. class of classifications	Equiv. class of chance variables
Partial order	$\mathcal{P} \preceq \mathcal{Q}$	$x \leq y \Leftrightarrow H(x\|y) = 0$
Join	$\mathcal{P} \vee \mathcal{Q}$	$x + y$
Meet	$\mathcal{P} \wedge \mathcal{Q}$	xy
Metric	Undefined	$\rho(x, y) = H(x\|y) + H(y\|x)$

We summarize major connections between a partition lattice and an information lattice in Table 1. The differences are rooted in the *separation of clustering from statistics*, so roughly speaking, a partition lattice—which is measure-free—can be thought of as an information lattice without probability measure. In this sense, abstraction is a more general concept than information by not being specific to communication problems, and in particular, is not attached to stochastic processes or information-theoretic functionals such as entropy.

This leads to our view of group-theoretic learning

$$\begin{aligned} &\text{subgroup lattice} \to \text{partition lattice} \to \\ &\quad \text{information lattice } (\to \text{learning}). \end{aligned} \tag{4}$$

The separation of clustering and statistics is important since it opens the opportunity for *interpretable statistical learning*, where *interpretability* is achieved by the explicit construction of a partition lattice (symmetry-generated hierarchical clustering), and *learning* is achieved by subsequent statistical inference on this lattice.

This is more precisely presented in Chain (4) aiming for learning, which at first glance, is just a reverse process of Chain (2) aiming for reinterpretation. However, the subgroup lattices in both chains are in stark contrast: the subgroups considered in Chain (4) are based on certain symmetries—the underlying mechanism of abstraction for knowledge discovery—whereas the subgroups considered in Chain (2) are merely (isomorphic) restatements of the given partitions. That is to say, among many possible subgroups that generate the same partition, we only pick the one that provides us explanations using the types of symmetries under consideration. The preservation of interpretable symmetries through Chain (4) makes the subsequent learning transparent. Therefore, when abstraction does meet statistics, it will yield interpretable machine learning and knowledge discovery, which is beyond simply a reinterpretation of known results.

Constructing Abstraction Universes and Learning on Lattices

We formalize ILL as a single optimization problem and then solve it practically in two phases. Given a signal/dataset to explain, optimize over rule sets such that the best rule set a) recovers the signal well and b) is simple. Since exact recovery may not always be possible, use a divergence function like relative entropy (Kullback–Leibler divergence) to measure loss, which is to be minimized. Information loss may occur if the abstraction universe is insufficient, or if we make algorithmic choices like favoring uniformity inappropriately. We say a rule set is simpler if it has fewer and simpler rules; whereas a rule is informationally simpler if it has smaller entropy, so it is more deterministic, easier to remember, and closer to common notions of a "rule." Note that the lattice structure induces a tradeoff between the two goals, which is hard to address computationally since full partition lattices are superexponential in size. Our two-phase approach is as follows.

Constructing Abstraction Universes

The first phase is to construct a restricted partition lattice based on group-theoretic foundations and using domain-specific or domain-agnostic priors on which symmetries might be most relevant. The need for restriction is to ensure human interpretability: not every set partition is easily explained but those that come, e.g., from group-theoretic invariances are naturally explained by the mechanism of the group action. Moreover, since there are a superexponential Bell number of possible set partitions (also patterns in universal source coding [38]) to say nothing of the subgroup structure that relates them to one another in the lattice, working with all possibilities would be computationally infeasible.

Lattice construction plays a role similar to building a function class in machine learning, which is sometimes called metalearning. While its importance is commonly understood, the construction phase in many data-driven models is often treated cursorily—using basic templates and/or ad hoc priors—leaving most computation to the learning phase. In contrast, we put substantial computational and mathematical effort into our prior-driven construction phase. Pursuing generality and interpretability, we want universal, simple priors that are domain-agnostic and close to innate human cognition [39]. Thus, we draw from Core Knowledge in cognitive science [11], [40], where we have studied the following two main categories:

- "the (small) natural numbers and elementary arithmetic prior," and
- "the elementary geometry and topology prior"

which largely correspond to isometries (rigid body transformations) in group theory.

We have developed computational group theory algorithms, which we detail elsewhere [6], [25] to construct abstractions from these priors, and consider such a construction prior-efficient if it is interpretable, expressive, and systematic. The main technical problem addressed by our algorithms is as follows: given an input space and a class of symmetries, explicitly compute a hierarchical family of symmetry-driven abstractions of the input space.

To go from mathematically formalized abstractions to computationally explicit abstractions, we introduce two general principles—a top-down approach and a bottom-up approach—to algorithmically generate hierarchical abstractions from hierarchical symmetries enumerated in a systematic way. The two principles leverage different dualities developed in the formalism and lead to practical algorithms that realize the abstraction-generating process.

In the top-down approach, we start from all possible symmetries and gradually restrict to certain types of symmetries. For example, symmetries common in crystallography, symmetries induced from affine transformations, or isometries that are all human-interpretable as per Core Knowledge. In these examples, a large symmetry-enumeration problem not only decomposes into smaller enumeration subproblems, but also suggests ways of adding restrictions to obtain desired symmetries. This approach from general symmetries to more restrictive ones corresponds to top-down paths in the symmetry hierarchy.

In the bottom-up approach, we start from a set of atomic symmetries (the seeds, corresponding to simple functions like sort or mod$_{12}$) and generate all symmetries that are seeded from the given set. A strong duality result we develop yields an induction algorithm to compute a hierarchical family of abstractions without explicitly enumerating the corresponding symmetries. This induction algorithm allows abstractions to be made from earlier abstractions and is therefore more efficient than generating all abstractions from scratch. This approach from atomic symmetries to more complicated ones corresponds to bottom-up paths in the symmetry hierarchy.

The explicitly constructed partition lattice is the universe of possible abstractions that are considered in the second phase, the statistical learning phase.

Learning on Lattices

The learning phase of the ILL framework starts with the constructed partition lattice and a (small) data set, such as a small number of chorales by J. S. Bach, which are canonical in music theory. Learning in an information lattice means searching for a minimal subset of simple rules from the information lattice of a signal (dataset) to best explain that signal (dataset). We adopt a (greedy) idea much like principal component analysis by first finding the most essential rule in explaining the signal, then the second most essential rule in explaining the rest of the signal, and so on. Learning in the lattice proceeds iteratively, as depicted in Figure 2 according to a student–teacher architecture. The student is a music generator and the teacher is a discriminator. The two components form a loop where the teacher guides the students toward a target style (using the input dataset, e.g., Bach) through iterative feedback and exercise, which map onto the process of extracting and applying rules. As detailed elsewhere [5], [8], [9], this is done by optimizing information functionals in the two components. Namely, the student tries to find the most random (largest entropy) probability distribution under the current rule set, whereas the teacher tries to find the most discriminative (largest relative entropy) rule to add to the rule set, which satisfies lattice-structured constraints, ensuring the added rule is not redundant with existing rules, and trying to find rules that have highly concentrated probability mass (e.g., low entropy). This iterative procedure aims to solve the optimization problem of searching for a minimal subset of simple rules from the information lattice of a signal so as to best explain that signal.

More specifically, we start with an empty rule set. Then the teacher (discriminator) takes as input the student's latest style $p_{\text{stu}}^{\langle k-1\rangle}$ and the input style p from the training corpus and identifies a feature through which the two styles manifest the largest gap $D(p_{\text{stu}}^{\langle k-1\rangle}\|p)$, subject to the rule corresponding to the feature not already being in the rule set or being too hierarchically close to a rule in the rule set. The identified feature is then made into a rule Γ_k and is added to the rule set $\{\Gamma_i\}_{i=1}^{k}$. Adding the rule that maximizes divergence to the rule set tends to minimize the original objective. Computationally this maximization is a discrete optimization, which also corresponds to the optimization of a divergence quantity called Bayesian surprise. The student (generator) takes as input the augmented rule set to update its writing style into $p_{\text{stu}}^{\langle k\rangle}$ and favors creativity, i.e., more possibilities, by maximizing the entropy subject to the rule constraints. When the entropy function is taken as the Tsallis entropy, the problem is least-squares optimization for which there are efficient algorithms.

In short, the teacher extracts rules while the student applies rules; both perform their tasks by solving optimization problems. ILL outputs not just a human-interpretable hierarchy of human-interpretable rules, where the hierarchy is interpretable due to the lattice structure and the rules are interpretable due to their mechanistic explanation from symmetries. It also outputs a rule trace comprising an evolving sequence of rules, rule sets, and recovered signals, which is useful as a curriculum for teaching, among other uses.

Automatic Music Theorist

Let us return to the real application in music, where we build an automatic music theorist to develop a hierarchy of music theory rules and further to teach students personalized lessons on music composition [4], [29]. It implements the "student $\rightleftharpoons$ teacher" model in the music setting: the student is a music generator and the teacher is a music discriminator. The two components form a loop where the teacher guides the students toward a target style through iterative feedback and exercise, which map onto our process of extracting and applying rules. This application reaches beyond basic illustrations by considering:

- many *music voices (or channels)*, so signals are in higher dimensions and rules are on more complex chord structure; and
- *temporal structure*, so signals include various (un)conditional chord distributions (i.e., n-grams for $n = 1, 2, \ldots$),

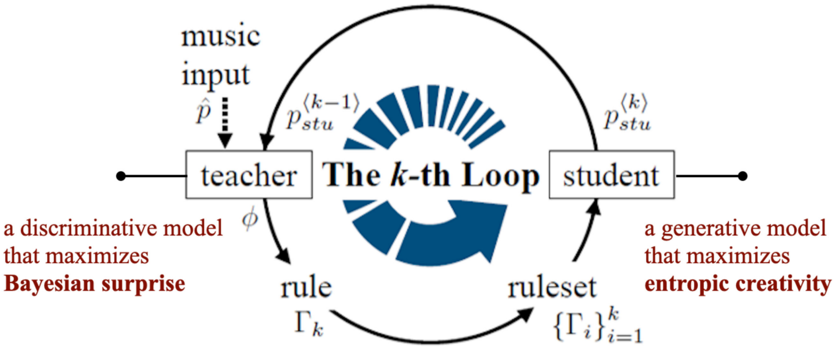


Figure 2

Student–teacher iterative structure for learning an information lattice, given a partition lattice and a music input dataset p. The student tries to be most random by maximizing an entropic quantity under the current rule set $\{\Gamma_i\}_{i=1}^k$ to produce a probability distribution $p_{\text{stu}}^{\langle k\rangle}$. The teacher tries to find the most discriminative rule Γ_k by maximizing a relative entropy between the most recent student probability $p_{\text{stu}}^{\langle k-1\rangle}$ and the music input p under constraints from the lattice and from the desire for concentrated rules.

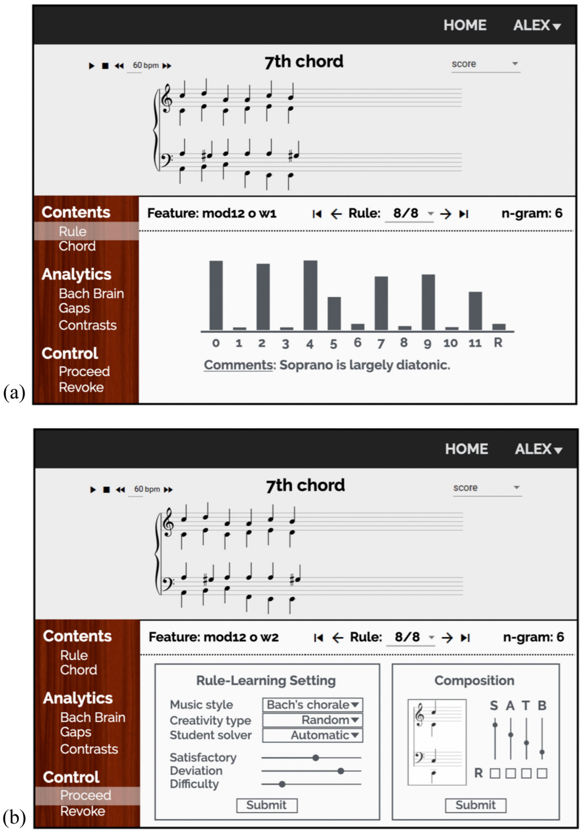
yielding both context-free and context-dependent rules, but new challenges too, namely *rare contexts/conditionals* and *contradictory rules.*

ILL's core idea of *abstraction* makes rare context common and a redesigned lifting operator solves contradiction [35]. Further, ILL parameters are made into knobs for human learners to personalize pace.

User Interface

We designed a web interface (see Figure 3) for the music application so users can more easily control the rule-learning process and how the learned results are displayed. Users learn music rules—each rendered as a histogram over a tagged partition (i.e., machine-codified music concepts)—and personalize their learning pace using knobs in the interface. These include rule difficulty, satisfactory level (a high value indicates high fidelity of the recovered signal to the original), and deviation level (a high value indicates a larger perturbation to the rule). Their set values are automatically converted to internal parameters.

Music is highly contextual. To model context, we consider more than one signal simultaneously, including multiple n-grams with varying n values and varying conditionals. In this way, ILL projects n-grams to lattices and aims for rules that characterize not only individual chord formation but also melodic and harmonic progression. Accordingly, ILL produces both context-free and context-dependent rules, each indexed by a partition and a conditional under that partition. For example, given a partition that abstracts chords into roman numerals and conditioned on the previous two chords being $\text{I}_4^6 \rightarrow \text{V}$, an ILL rule specifies the probability distribution of the next roman numeral rather than the next chord, and in this case, reproduces the music rule on Cadential-64. Note that in a context-dependent rule, not only is the query chord abstracted, but also the conditional. This *abstracted n-gram* differs from plain n-gram models. The latter may suffer from rare context, i.e., a conditional occurs very few or even zero times in the training set. Yet, the core



Feature: mod12 o w1 Rule: 8/8 n-gram: 6

Comments: Soprano is largely diatonic.

Feature: mod12 o w2 Rule: 8/8 n-gram: 6

Music style: Bach's chorale
Creativity type: Random
Student solver: Automatic

Figure 3

Music web interface has (a) a rule histogram and (b) a user control panel.

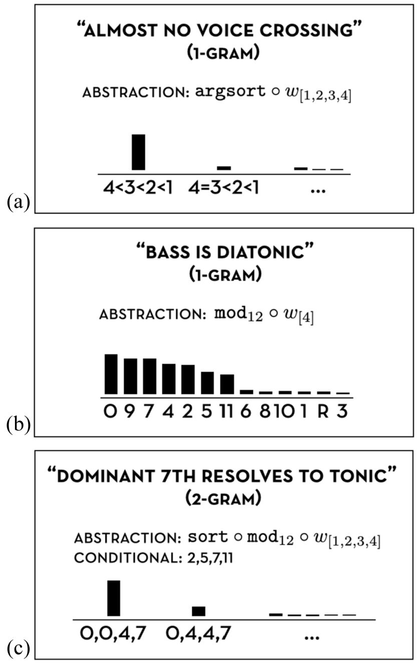


Figure 4

Examples of ILL rules: (a), (b) context-free; (c) context-dependent, where the $w_{[\cdot]}$ window operator selects voices (soprano, alto, tenor, bass).

idea of abstraction makes small data large and rare contexts common. Under appropriately discovered equivalence classes, rare things become prevalent: a theorist might think that "Although I have never seen this exact chord progression before, I have seen this type." Figure 4 exemplifies two context-free rules and a context-dependent one. These rule histograms are generated by ILL based on 370 of Bach's four-part chorales.

Knowledge Rediscovery

Making use of the music application's web interface, we conduct studies to evaluate two important performance metrics of an ILL application, or indeed a knowledge-discovery task in general, namely *rule-learning capability* and *human-interpretability*. The first study assesses both how much known knowledge an AI can reproduce (common in automatic knowledge discovery settings, such as [41] and [42]) and how much new it can discover. The second dimension normally involves studies with human evaluators, which we return to in the next section. Figure 5 summarizes the main results for both performance dimensions.

To assess rule-learning capability, let us compare machine-discovered rules with human-codified domain knowledge to identify how much is rediscovered and also what new can be discovered. In our context, we compare ILL-distilled rules to the standard undergraduate music theory curriculum at the University of Illinois Urbana-Champaign. The initial idea is to use known theory as a benchmark. Yet, we emphasize the ultimate goal is not to use known theory as supervision to reconstruct only what we know, but also to discover new rules, new understandings of existing rules, and new composition possibilities, and to teach rules in a personalized way.

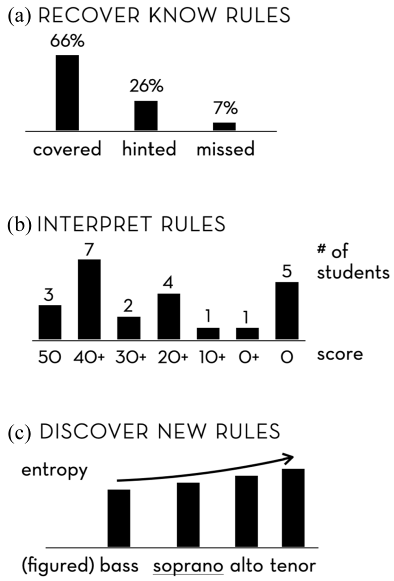


Figure 5

Assessments of ILL on knowledge-discovery tasks. (a) Trained on 370 chorales, ILL explicitly reproduced 66% and implicitly hinted at 26% of a standard music theory curriculum. (b) In an interpretation-focused assignment, the majority (2/3) of the music theory students who did the assignment succeeded (w.r.t. 30/50 passing score) in interpreting ILL-discovered rules. (c) ILL revealed a new way of building chords, namely figured soprano, which is confirmed independently by other music theorists. More examples of new rules are in the text.

Before proceeding, note three major differences between human-codified music theory and ILL-generated rules.

- *Raw music representation (input):* Known music theory is derived from all aspects of sheet music whereas ILL-generated rules are currently derived only from MIDI pitches and their durations in digital sheet music. This is because we currently study ILL as a general framework. One can later include more music raw information, such as spelling, meter, measure, beaming, and articulation.
- *Rule format (output):* Known music theory and ILL-generated rules have two different styles. The former is more descriptive and absolute (hard), whereas the latter is more numerical and stochastic (soft). For instance, a music rule that strictly bans consecutive

TABLE 2. Comparison of ILL's Symmetry-Induced Rule Abstractions to Music OPTIC

Operation	*Music description*	*Subgroup*	
Octave shift	"Move any note into a new octave."	$\langle\{t_{12\boldsymbol{e_1}}, t_{12\boldsymbol{e_2}}, t_{12\boldsymbol{e_3}}, t_{12\boldsymbol{e_4}}\}\rangle$	✓
Permutation	"Reorder the object, changing which voice is assigned to which note."	$\langle\{r_{P(1,2)}, r_{P(2,3)}, r_{P(3,4)}\}\rangle$	✓
Transposition	"Transpose the object, moving all of its notes in the same direction by the same amount."	$\langle\{t_{\mathbf{1}}\}\rangle$	✓
Inversion	"Invert the object by turning it 'upside down'."	$\langle\{r_{-I}\}\rangle$	✓
Cardinality Change	"Add a new voice duplicating one of the notes in the object."		✗

fifths is softened in an ILL rule that assigns a small non-zero probability. So, while it is possible to "translate" a probabilistic rule in ILL to a verbal rule in known theory, it may not make sense to "translate" the other way. Furthermore, it may not be wise to hard-code known rules as categorical labels in a supervised setting, as music rules are inherently flexible and hard-coding may lead to a rule-based AI that generates "mechanical" music like the Illiac Suite [30].

Dataset specificity: Music theory is often intended for educational purposes, rather than to reflect the style of a musical oeuvre. For instance, while consecutive fifths are banned in homework and exams, they are used in the real-world composition. Even in our dataset of Bach's chorales, which are supposed to follow the known rules quite well, we see some consecutive perfect intervals. ILL-generated rules are specific to the input data set. We may find datasets that follow the known rules quite well (e.g., Bach's chorales), but others that break known rules and set their own.

Keeping these three differences in mind and isolating them from the comparison results, we discuss the remaining differences due to the rule-learning process itself. To come up with the benchmark, we compiled a comprehensive syllabus of laws from music theory taught in our music school's theory review course, which runs through the full series of theory classes at a fast pace. This human-codified music knowledge is organized as a running list of 75 topics and subtopics indexed by lecture number. On the other hand, ILL-generated rules are indexed by partition (ID) and n-gram (n). Results are summarized in Table 3, where the colored crosses in the last column indicate topics missed by ILL for different reasons. Among the total 75 topics in Table 3, we first ignore seven of them (red crosses), which require raw music representations beyond MIDI pitches and durations (e.g., accents and enharmonic respellings of some augmented sixth chords). ILL covered 45 out of the remaining 68 topics, i.e., 66%. Among the 23 missed topics, 18 (blue crosses) are related to deeper-level temporal abstractions, such as harmonic functions (the tendency of certain chords to progress to other chords or to remain at rest) and forms (long-range structure of a musical composition). These temporal abstractions may be better modeled as *abstractions of transitions*, which are implicitly captured but not explicitly recovered from our current multiabstraction multi-n-gram language model, modeling only *transitions of abstractions*. The other five missed topics (black crosses) are tricky and require *ad hoc* encodings, not explicitly learnable, but may be implicitly captured from our current ILL implementation. Accordingly, the $30 = 7 + 18 + 5$ uncovered topics suggest three future directions to raise ILL's rule-learning capacity:

1) include more raw music representations;
2) model abstractions of transitions; and
3) make music-specific adjustments to ILL (or find a more expressive and general framework).

Recall, however, that the goal here is not to reproduce what we know but also to augment it. We may stop after enabling abstractions of transitions, which may improve coverage to 84% (i.e., 93% of the topics from MIDI notes only) and may be sufficient for meaningful understanding and training.

TABLE 3. Comparison of ILL-Generated Rules to Human-Codified Laws of Music Theory Taught in Standard Undergraduate Music Theory Courses. Checks (45) in the Last Column Denote Topics Recovered by ILL

Lecture	*Music theory*	*Partition IDs*	n*-gram*	
1	Music accents			✗
2	Pitch	1–4	1	✓
2	Pitch class	16–19	1	✓
2	Interval	31–36	1	✓
2	Interval class	97–102	1	✓
3	Stepwise melodic motion (counterpoint)	1–4	2	✓
3	Consonant harmonic intervals (counterpoint)	97–102	1	✓
3	Beginning scale degree (counterpoint)	16–19	2	✓
3	Ending scale degree (counterpoint)	16–19	2	✓
3	Beginning interval class (counterpoint)	97–102	2	✓
3	Ending interval class (counterpoint)	97–102	2	✓
3	Parallel perfect intervals (counterpoint)	97–102	2	✓
3	Directed perfect intervals (counterpoint)			✗
3	Law of recovery (counterpoint)	1–4	$\geq$3	✓
3	Contrapuntal cadence (counterpoint)	1–4, 97–102	2,3	✓
3	Melodic minor ascending line (counterpoint)			✗
4	Triads and seventh chords	26–30	1	✓
4	Triads and seventh chords: quality	140–144	1	✓
4	Triads and seventh chords: inversion	113–117	1	✓
5	Figured bass	113–117	1,2	✓

TABLE 3. (*Continued*) Comparison of ILL-Generated Rules to Human-Codified Laws of Music Theory Taught in Standard Undergraduate Music Theory Courses. Checks (45) in the Last Column Denote Topics Recovered by ILL

Lecture	*Music theory*	*Partition IDs*	*n-gram*	
5	Roman numerals	81–85,129–133	1	✓
6	Melodic reduction (Schenkerian analysis)			✗
7	Passing tone (tones of figuration)	1–4, 134–144	3	✓
7	Neighbor tone (tones of figuration)	1–4, 134–144	3	✓
7	Changing tone (tones of figuration)	1–4, 134–144	4	✓
7	Appoggiatura (tones of figuration)	1–4, 134–144	3	✓
7	Escape tone (tones of figuration)	1–4, 134–144	3	✓
7	Suspension (tones of figuration)	1–4, 134–144	3	✓
7	Anticipation (tones of figuration)	1–4, 134–144	3	✓
7	Pedal point (tones of figuration)	1–4	≥ 3	✓
7	(Un)accented (tones of figuration)			✗
7	Chromaticism (tones of figuration)			✗
8	Tonic (function)			✗
8	Dominant (function)			✗
8	Authentic cadence	1,4,81–85,129–133	2,3	✓
8	Half cadence	81–85,129–133	2,3	✓
9	Voice range (four-part texture)	1–4	1	✓
9	Voice spacing (four-part texture)	31–41	1	✓
9	Voice exchange (four-part texture)	20–25	2	✓
9	Voice crossing (four-part texture)	53–63	1	✓

TABLE 3. (*Continued*) Comparison of ILL-Generated Rules to Human-Codified Laws of Music Theory Taught in Standard Undergraduate Music Theory Courses. Checks (45) in the Last Column Denote Topics Recovered by ILL

Lecture	*Music theory*	*Partition IDs*	*n-gram*	
9	Voice overlapping (four-part texture)			✗
9	Tendency tone (four-part texture)	16–19	1,2	✓
9	Doubling (four-part texture)	86–91	1	✓
10	Harmonic reduction (second-level analysis)			✗
11	Expansion chord			✗
12	Predominant (function)			✗
13	Phrase model			✗
14	Pedal or neighbor (six-four chord)	4,113–117	3	✓
14	Passing (six-four chord)	4,113–117	3	✓
14	Arpeggiated (six-four chord)			✗
14	Cadential (six-four chord)	85,113–117,133	3,4	✓
15	Embedded phrase model			✗
16	Nondominant seventh chord (function)			✗
17	Tonic substitute (submediant chord)			✗
17	Deceptive cadence (submediant chord)	81–85,129–133	2,3	✓
18	Functional substitute (mediant chord)			✗
19	Back-relating dominant	81–85,129–133	2,3	✓
20	Period (I)			✗
21	Period (II)			✗
22	Period (III)			✗

TABLE 3. (*Continued*) **Comparison of ILL-Generated Rules to Human-Codified Laws of Music Theory Taught in Standard Undergraduate Music Theory Courses. Checks (45) in the Last Column Denote Topics Recovered by ILL**

Lecture	*Music theory*	*Partition IDs*	n*-gram*	
23	Applied chords (I)	81–85,129–133	2,3	✓
24	Applied chords (II)	81–85,129–133	2,3	✓
25	Applied chords (III)	81–85,129–133	2,3	✓
26	Modulation (I)			✗
27	Modulation (II)			✗
28	Binary form (I)			✗
29	Binary form (II)			✗
30	Modal mixture			✗
31	Neapolitan	81–85,129–133	1	✓
32	Italian sixth chord	140–144	1	✓
32	French sixth chord	144	1	✓
32	German sixth chord			✗
32	Swiss sixth chord			✗
33	Ternary form			✗
34	Sonata form			✗

Notes: Red crosses (7) denote topics not recoverable from our raw music representations; blue crosses (18) denote topics not recoverable from our n-gram transitions of abstractions/partitions; black crosses (5) denote topics not recoverable from the constructed lattice of abstractions.

Let us also consider another music theory source focused on music symmetries [43]. We compare ILL-generated rules with a set of commonly used music operations, known as the OPTIC operations: octave shifts (O), permutations (P), transpositions (T), inversions (I), and cardinality changes (C). As summarized in Table 2, ILL covers the major four types of operations (OPTI). The C operation is not recovered because it is not a transformation in the mathematical sense of being unambiguous and bijective. Notationally, t_v denotes a translation by the translation vector v, i.e., $t_v(x) := x + v$; r_A denotes a rotation (can be proper or improper) by the rotation matrix A, i.e., $r_A(x) := Ax$ As a special type of rotation matrix, $P^{(\cdots)}$ denotes a permutation matrix where the superscript is the cycle notation of a permutation. Note that ILL, as a general framework, considers a much larger universe of generic symmetries (from Core Knowledge) beyond those already considered in music. Therefore, ILL not only recovers existing music symmetries, but also suggests new symmetries to be exploited in new styles.

New Knowledge Discovery

Besides the rediscovery of existing music theory knowledge, the ILL approach also discovers new knowledge. We mention a few new rules discovered by ILL that piqued the interest of our colleagues in the School of Music.

a) Tritone resolution plays a key role in tonal music and is an epitome of many more general harmonic resolutions. However, in Bach's chorales, tritones sometimes do not resolve in typical ways, but rather consistently transition to other dissonances, such as a minor seventh, behaving like a harmonic version of an escape or changing tone.

b) A new notion of "the interval of intervals" has been consistently extracted in several ILL-generated rule traces. This "second derivative," like acceleration in mechanics, might suggest a new microscopic chord structure that has not been considered before.

c) New symmetry patterns reveal new harmonic foundations, hence new composition possibilities. As a parallel concept of harmony traditionally built on *figured bass* (the dominant pattern in Bach's chorales, as confirmed by ILL), ILL reveals the presence of "figured soprano" as the next alternative in explaining Bach's music [see Figure 5(c)]. Although not the best way to explain Bach's chorales according to ILL and also not included in any standard music theory class, it may be a more efficient perspective to view or create music starting deviating from classical (e.g., in Jazz). Indeed, this idea was developed, contemporaneously with us, by Casey Sokol [44], a music professor at York University, which we quote in the following: "The idea of Figured Soprano is simply a way of taking this thinking from the top-down and bringing it into greater prominence as a creative gesture. So these exercises are not anything new in their ideation, but they can bring many new ideas, chord progressions and much else. It's a somewhat neglected area of harmonic study and it's a lot of fun to play with."

Human Interpretability

To assess human interpretability, we ask people to interpret machine-generated rules. Such human evaluation is considered the gold standard in the human–computer interaction (HCI) and explainable AI literatures. In particular, to characterize the degree to which ILL-generated rules are interpretable, we assess human-generated verbal interpretations of ILL rules, which are originally sophisticated symbolic and numeric objects. Let us detail the collection and assessment procedure.

The evaluation was conducted in the form of a two-week assignment for 23 students from the CS+Music degree program at the University of Illinois Urbana-Champaign. Each student had a basic knowledge of computer science, math, and music theory, but no student had read any ILL-generated rules before. By *interpretability*, we mean interpretable to these students.

The assignment had three parts. Part I gave detailed instructions on the format of rules as shown in Figure 4, including both feature-related and probability-related instructions (symmetries were excluded from tags because group theory is unfamiliar to these students). We provided verbal definition, mathematical representation, and typical examples for each of the following terms: chord, window (for coordinate selection), seed feature, feature, rule, n-gram, histogram, and dataset. An understanding of these eight terms was the only prerequisite for the assignment. The estimated reading time of instructions was one hour.

Part II had eleven 1-gram rules—a histogram specified by window and seed feature(s); Part III contained 14 2-gram rules—a histogram specified by the window, seed feature(s), and additionally, a conditional. Students were asked to write what they saw in each of the histograms in response to the following two prompts:

- Does the histogram agree/disagree with any of the music rules and concepts you know (write in music-theoretic terms when possible)?
- Does the histogram suggest something new (i.e., neither an agreement nor a disagreement, with no clear connection to anything you know)?

Responses to each of the 25 rules were to be given as text, containing word descriptions that "decode" the histogram. Students were explicitly instructed that a literal repetition of the histogram (e.g., taking a modulo 12 of a chord results in a 91.2% chance of being 0,0,4,7) was unacceptable and only qualitative descriptions to reveal the music behind the math was requested. Students were also specifically instructed to only attend to the relative values of the probabilities (e.g., what are most likely, more likely, or nearly impossible). This students were asked to complete the assignment independently with no group work or office hours.

The assignment was designed such that every rule histogram encoded at least one music concept/rule consistent with standard music theory. In addition, every histogram contained either one additional known music rule or something strange that either conflicted with a known rule or represented something new. Each rule was scored as two points.

To score the assignments, we prepared an initial rubric containing the (authoritative) music keywords used to describe every rule histogram. To ensure the credibility and fairness of the initial rubric, we held a discussion session (after the assignment was submitted) with all students and teaching staff, as a form of peer review. During that session, we discussed all 25 rules individually. For each, we first announced keywords in the initial rubric and explained that these keywords would later be used to score their assignment. Every student was encouraged to object to any of our announced keywords or to propose new keywords accompanied with a convincing explanation. New/modified keywords commonly agreed upon by the students and teaching staff were added to the initial rubric. By the end of the discussion, there was a more inclusive rubric containing broadly accepted keywords. This rubric-generating process was transparent to all students. Every student's response sheet was scored against keywords in the inclusive rubric, and the resulting scores are summarized in Figure 5(b). Besides not doing the assignment, a major score deduction was due to misunderstanding the n-gram (e.g., the probability of the present condition on the past was mistakenly interpreted as the probability of the past conditioned on the present). This may be largely due to unfamiliarity with the n-gram models for new CS+Music students. Notwithstanding, most students that completed the assignment succeeded in exceeding a 30/50 score, and several received perfect scores. This provides evidence for the interpretability of ILL rules.

Co-Creativity

Creativity is powerful. Regarded as one of our most sophisticated cognitive skills, this ability drives human progress by allowing us to perform nonroutine tasks, take advantage of novel opportunities, and invent new solutions to problems facing the world. Creativity is the hallmark of art and science, as well as engineering and technology that benefits wide swaths of society. Music composition is often thought of as an exemplary form of creativity, especially since music is engaging and central to human self-expression and culture. Creativity often builds on a knowledge base like music theory.

Although popular culture tends to lionize the lone genius and music composers traditionally work alone, group creativity often trumps individual creativity. Work in numerous creative domains is largely carried out by teams rather than heroic solo inventors, but comes with its own dynamics. Effective collaboration requires not only cooperation, i.e., having aligned goals, but also coordination mechanisms to enable effective alignment and adjustment to teammates' actions. Yet, group creativity is difficult: individual creative contributions are fundamentally complex and co-dependent; their combination requires more intelligence than a simple summation or independent voting. How to disaggregate/aggregate disparate contributions in complex tasks, such as composing music, has been an open question in HCI [45] and collective intelligence [46].

Based on the human-interpretable ILL framework for music, we have built a platform to support people working together with others and with AI to co-create new music, while preserving the autonomy of individual human contributors. The use of the ILL framework also enables a more abstract language by which creative communication can take place. For example, one might want to extract the melody and harmony from the song Happy Birthday and the rhythm and texture from Mozart's K545 piano sonata, and combine them together into a new song, specified just at this level, without note-by-note editing. To do so, we have developed lattice-based operations to decompose music into more abstract fragments, such as just the harmony, as well as ways to recompose several abstract fragments together. The decomposition operations go down the information lattice from raw representations to deeper-level abstractions as a lattice join operation, whereas recomposition operations go up the information lattice from abstractions to realize complete music in full raw representation as a lattice meet operation. This is a much more controlled alternative to simple stochastic sampling from the student phase of the iterative learning algorithm, which also leads to mellifluous results when a sufficient number of rules have been learned.

Outreach work with several youth groups focused on Hip Hop music is demonstrating the ease and engagement from ILL-based co-creativity, as well as the possible human well-being from social creativity activities [10].

Discussion

Music can be described in many different ways: in effusive and subjective statements such as "Haydn's Symphony No. 45, 'Farewell,' is a classic example of his *Sturm und Drang* writing (literally, *storm and stress*), wherein marvelously inventive and varied music is by turns dramatic and sublime," as well as in technical and clinically objective statements such as "The first movement of Haydn's Symphony No. 45, *Allegro assai*, is in 3/4 time and begins on the tonic of F# minor, cast in the sonata form and giving us *Sturm und Drang.*" Such contrasting descriptions reveal two distinct types of languages for different audiences: the former is widely understood by a general audience but somewhat vague, whereas the latter—filled with music theory terms—is more precise but also likely to be restricted to conversations among musicians.

The centuries-long effort in developing music theory aims to make music concepts precise and introduce tools/methods for people to describe, to understand, and further to make music in a more guided manner. In particular, theories on tonality have been well established as *de facto* standard materials taught in music conservatories, especially in the Western tradition. The precision gained through the language of music theory on the objective nature of music enables people who are miles away or even decades away to freely exchange their musical ideas.

It is noteworthy that modern music theory has stepped beyond its origins in the music community. There is a desire to make music theory more rigorous through mathematics and a desire to automate its development through information processing techniques. Indeed, modern music theory incorporates more advanced mathematics, such as set theory, abstract algebra, as well as geometry, and topology. Mathematical models of music concepts allow further fundamental discoveries, which embed known theory into a larger framework. Musically, this larger view suggests "gaps" in theory that further lead to new possibilities in composition, yielding exactly what we see in contemporary and modern music experimentation, as well as in new renderings of classic music.

Music in the sheet music symbolic (discrete) representation is unique to its compositional core since the same Haydn's Symphony No. 45 can be played by different orchestras and recorded by different means. If the goal were to appreciate the performance as a whole from a listener's perspective, one should faithfully study the specific sound recordings. However, if the focus is composition, as in this article, we should instead place less attention on the variations introduced by performers and recorders. As the music theorist Dmitri Tymoczko put it, "There is a potential for real divergence between what we might call composer's grammar and listener's grammar." This article focuses on interpretable music concepts that comprise the composer's grammar, but there are many theoretical questions on the listener side that remain wide open.

Operating on sheet music rather than audio signals, we have presented a framework to automatically learn the principles of music theory in a human-interpretable form. Such machine learning algorithms for symbolic music support a variety of applications in music pedagogy, music composition, music retrieval, musicology, and computational creativity for music. We have shown initial results in this new area of research that is promising, but also suggest new future opportunities for information processing researchers and practitioners.

Indeed, as music has entered the digital age, the idea of computational music has reshaped musical activities. The popularization of *notation software* has dramatically shifted composition from its traditional pencil-and-paper origins, making music much easier to share and reproduce in a variety of new ways. Among music applications, we have seen the rise of automatic composers that not only create music indistinguishable from works by people, but also establish modern pioneering styles. Music education has benefited from automatic music teachers (e.g., Harmonia), which deliver music theory lessons and exercises to students, producing grades and feedback in less than a second. The ILL framework is also enabling creative music composition support through a music mixing platform from Kocree, Inc. Going forward, there are numerous possibilities for not only information processing research on music, but also the application of information processing techniques to music applications.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant SMA-1829366, and Grant CCF-1717530, and in part by the IBM-Illinois Center for Cognitive Computing Systems Research (C3SR), a research collaboration as part of the IBM AI Horizons Network.

References

[1] M. Krenn et al., "On scientific understanding with artificial intelligence," *Nat. Rev. Phys.*, Apr. 2022. [Online]. Available: https://www.nature.com/articles/s42254-022-00518-3

[2] C. E. Shannon, "The lattice theory of information," *Trans. IRE Professional Group Inf. Theory*, vol. 1, no. 1, pp. 105–107, Feb. 1953.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul./Oct. 1948.

[4] H. Yu, L. R. Varshney, G. E. Garnett, and R. Kumar, "MUS-ROVER: A self-learning system for musical compositional rules," in *Proc. 4th Int. Workshop Musical Metacreation*, 2016.

[5] H. Yu and L. R. Varshney, "Towards deep interpretability (MUS-ROVER II): Learning hierarchical representations of tonal music," in *Proc. 6th Int. Conf. Learn. Representations*, 2017.

[6] H. Yu, I. Mineyev, and L. R. Varshney, "To abstract via algebraic innateness: Hierarchical, interpretable, and task-free clustering," in *Proc. Inf. Theory Appl. Workshop*, 2019.

[7] H. Yu, H. Taube, J. A. Evans, and L. R. Varshney, "Human evaluation of interpretability: The case of AI-generated music knowledge," in *Proc. ACM CHI Workshop Artif. Intell. HCI: A Modern Approach*, 2020.

[8] H. Yu, J. A. Evans, and L. R. Varshney, "Mimicking human minds via information lattices," in *Proc. NeurIPS 2020 Workshop BabyMind: How Babies Learn How Mach. Can Imitate*, 2020.

[9] H. Yu, J. A. Evans, and L. R. Varshney, "Information lattice learning," 2021, submitted.

[10] H. Yu, J. A. Evans, D. Gallo, A. J. Kruse, W. M. Patterson, and L. R. Varshney, "AI-aided co-creation for wellbeing," in *Proc. 12th Int. Conf. Comput. Creativity*, 2021, pp. 453–456.

[11] E. S. Spelke and K. D. Kinzler, "Core knowledge," *Devlop. Sci.*, vol. 10, no. 1, pp. 89–96, Jan. 2007.

[12] A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *Fordham Law Rev.*, vol. 87, no. 3, pp. 1085–1139, 2018.

[13] J. Soni and R. Goodman, *A Mind At Play: How Claude Shannon Invented the Information Age*. New York, NY, USA: Simon & Schuster, 2017.

[14] L. R. Varshney, "Mathematizing the world," *Issues Sci. Technol.*, vol. 35, no. 2, pp. 93–95, 2019.

[15] D. Fallows, "Theory," in *The Oxford Companion to Music*. A Latham, Ed., Oxford, U.K.: Oxford Univ. Press, 2011.

[16] H. Taube, "Automatic tonal analysis: Toward the implementation of a music theory workbench," *Comput. Music J.*, vol. 23, no. 4, pp. 18–32, 1999.

[17] J. Evans and A. Rzhetsky, "Machine science," *Science*, vol. 329, no. 5990, pp. 399–400, Jul. 2010.

[18] Z. Liu and M. Tegmark, "Machine learning hidden symmetries," *Phys. Rev. Lett.*, vol. 128, May 2022, Art. no. 180201.

[19] H. Yu, L. R. Varshney, and G. Stein-O'Brien, "Towards learning human-interpretable laws of neurogenesis from single-cell RNA-seq data via information lattices," in *Proc. Learn. Meaningful Representations Life Workshop at NeurIPS*, 2019.

[20] H. Li and E. K. P. Chong, "Information lattices and subgroup lattices: Isomorphisms and approximations," in *Proc. 47th Annu. Allerton Conf. Commun. Control Comput.*, 2007, pp. 1103–1110.

[21] D. S. Parker and P. Ram, "The construction of Huffman codes is a submodular ('convex') optimization problem over a lattice of binary trees," *SIAM J. Comput.*, vol. 28, no. 5, pp. 875–1905, 1999.

[22] C. Chan, A. Al-Bashabsheh, J. B. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," *Proc. IEEE*, vol. 103, no. 10, pp. 1883–1913, Oct. 2015.

[23] T. F. Varley, "Decomposing past and future: Integrated information decomposition based on shared probability mass exclusions," Feb. 2022, *arXiv:2202.12992*.

[24] B. Brooks, A. Frankel, and E. Kamenica, "Information hierarchies," *Econometrica*, vol. 90, no. 5, pp. 2187–2214, Sep. 2022.

[25] H. Yu, I. Mineyev, and L. R. Varshney, "Orbit computation for atomically generated subgroups of isometries of $\mathbb{Z}^n$," *SIAM J. Appl. Algebra Geometry*, vol. 5, no. 3, pp. 479–505, 2021.

[26] V. Vapnik and R. Izmailov, "Rethinking statistical learning theory: Learning using statistical invariants," *Mach. Learn.*, vol. 108, no. 3, pp. 381–423, Mar. 2019.

[27] K. Ellis et al., "DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning," in *Proc. 42nd ACM SIGPLAN Int. Conf. Program. Lang. Des. Implementation*, 2021, pp. 835–850.

[28] J. R. Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd ed. New York, NY, USA: Dover, 1980.

[29] H. Yu and L. R. Varshney, "On 'composing music by a stochastic process': From computers that are human to composers that are not human," *IEEE Inf. Theory Soc. Newslett.*, vol. 67, no. 4, pp. 18–19, Dec. 2017.

[30] J. Lejaren, A. Hiller, and L. M. Isaacson, *Illiac Suite for String Quartet*. Bryn Mawr, PA, USA: Theodore Presser Company, 1957.

[31] D. Cope, *Experiments in Musical Intelligence*. Madison, WI, USA: A-R Editions, 1996.

[32] C.-Z. A. Huang et al., "Music transformer," Sep. 2018, *arXiv:1809.04281*.

[33] L. R. Varshney, "Limits theorems for creativity with intentionality," in *Proc. 11th Int. Conf. Comput. Creativity*, 2020, pp. 390–393.

[34] A. A. Issak and L. R. Varshney, "Artistic autonomy in AI art," in *Proc. 13th Int. Conf. Comput. Creativity*, 2022, pp. 170–174.

[35] H. Yu, T. Li, and L. R. Varshney, "Probabilistic rule realization and selection," in *Advances in Neural Information Processing Systems 30*, I. Guyon et al., Eds., New York, NY, USA: Curran Associates, Inc., 2017, pp. 1562–1572.

[36] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems Control Inf. Theory*, vol. 2, no. 2, pp. 149–162, 1973.

[37] T. H. Chan and R. W. Yeung, "On a relation between information inequalities and group theory," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1992–1995, Jul. 2002.

[38] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–1481, Jul. 2004.

[39] G. Marcus, "Innateness, AlphaZero, and artificial intelligence," Jan. 2018, *arXiv:1801.05667*.

[40] F. Chollet, "On the measure of intelligence," Nov. 2019, *arXiv:1911.01547v2*.

[41] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, Apr. 2009.

[42] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," *Sci. Adv.*, vol. 6, no. 16, Apr. 2020, Art. no. eaay2631.

[43] D. Tymoczko, *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. Oxford, U.K.: Oxford Univ. Press, 2010.

[44] C. Sokol, "Figured soprano," 2016. [Online]. Available: https://caseysokol.com/?page_id=1067

[45] J. Frich, L. M. Vermeulen, C. Remy, M. M. Biskjaer, and P. Dalsgaard, "Mapping the landscape of creativity support tools in HCI," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–18.

[46] A. Kittur, B. Lee, and R. E. Kraut, "Coordination in collective intelligence: The role of team structure and task interdependence," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2009, pp. 1495–1504.

Haizi Yu received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2009, the M.S. degree in computer science and electrical engineering from Stanford University, Stanford, CA, USA, in 2014, and the Ph.D. degree in computer science from the University of Illinois Urbana-Champaign, Urbana, IL, USA, in 2019.

He is a Researcher with the University of Chicago, Chicago, IL, USA, and the University of Illinois Urbana-Champaign. His research interests include human-interpretable artificial intelligence, automatic knowledge discovery, and music intelligence.

Lav R. Varshney (Senior Member, IEEE) received the B.S. degree in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2004, and the S.M., E.E., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006, 2008, and 2010, respectively.

He is an Associate Professor of electrical and computer engineering with the University of Illinois Urbana-Champaign, Urbana, IL, USA. His research interests include information theory, artificial intelligence, and creativity.

Heinrich Taube received the B.A. and M.A. degrees from Stanford University, Stanford, CA, USA, and the Ph.D. degree from the University of Iowa, Iowa City, IA, USA, all in music composition.

He is a Professor with the School of Music, University of Illinois Urbana-Champaign, Urbana, IL, USA, where he is the Chair of the Music Composition, Music Theory, and Music Technology area and also of the CS+Music area.

James A. Evans received the B.A. degree in anthropology from Brigham Young University, Provo, UT, USA, in 1994, and the M.A. and Ph.D. degrees in sociology from Stanford University, Stanford, CA, USA, in 1999 and 2004, respectively.

He is the Max Palevsky Professor of Sociology, the Director of Knowledge Lab, and the Faculty Director of Computational Social Science with the University of Chicago, Chicago, IL, USA, and an External Professor with the Santa Fe Institute, Santa Fe, NM, USA. His research interests include large-scale data, machine learning, and generative models to understand how collectives think and what they know, which involves inquiry into the emergence of ideas, shared patterns of reasoning, and processes of attention, communication, agreement, and certainty, with a special interest on innovation.