Towards Deep Interpretability (MUS-ROVER II): Learning Hierarchical Representations of Tonal

Haizi Yu[1] Lav R. Varshney[1]

[1]University of Illinois at Urbana-Champaign

ICLR, 2017
Presenter: Xueying Bai

Outline

Outline

Task

Task: Study underlying patterns beneath the music surface.

- **Discovering compositional rules from raw music data:** music theorists develop concepts and rules to describe the regularity in music compositions;
  Computer scientists translated these rules into programs that automatically generate music.
- **Forming hierarchical concepts:** music theorists have devised multi-level analytical methods to emphasize the hierarchical structure of the music.

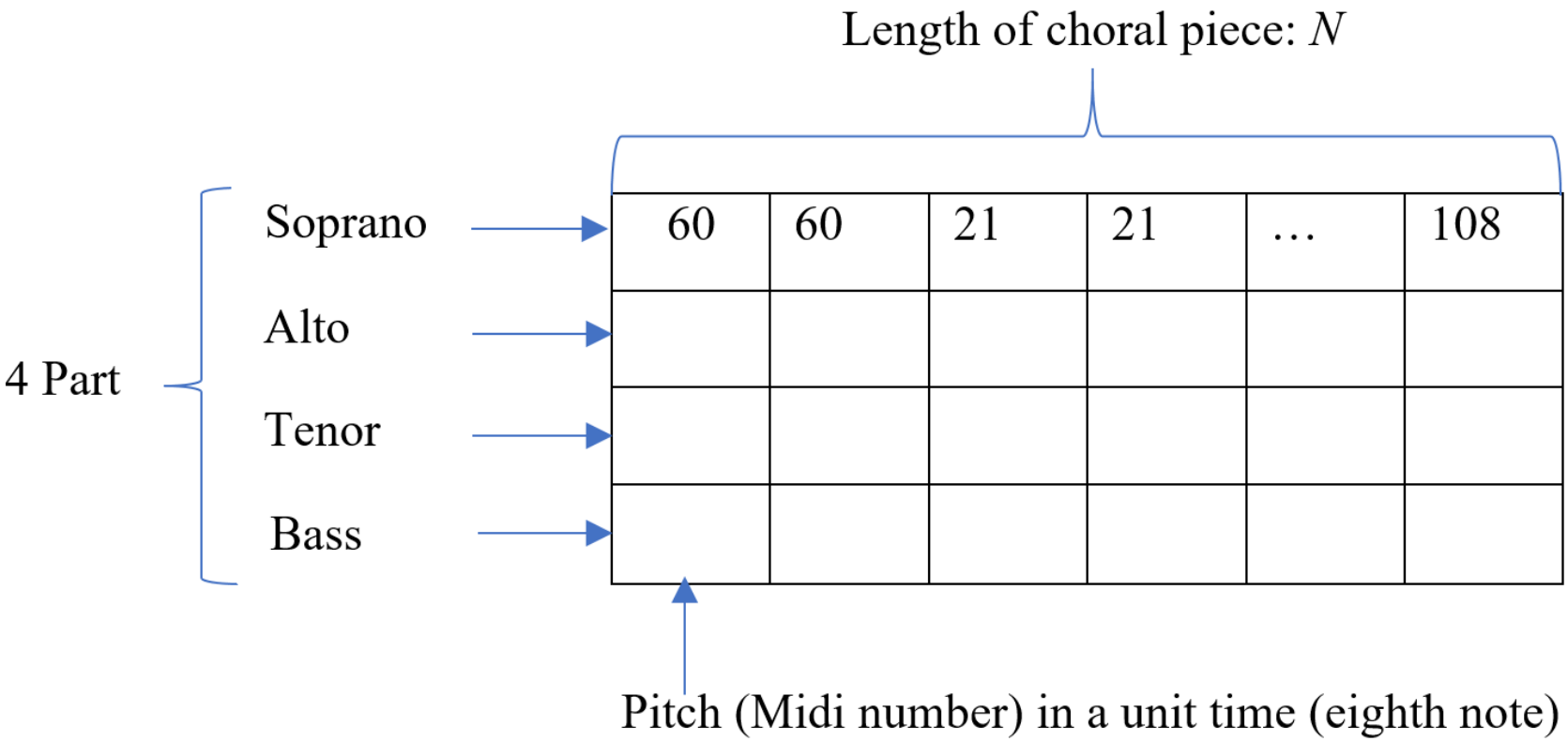
Outline

Related Work

- **Adversarial or Collaborative:** self-learning loop is similar to GAN.
- **Interpretable Feature Learning:** first recover disentangled representations, then interpret the semantics of the learned features.
- **Automatic Musicians:** models to automate the interaction of music theory and composition.

Outline

Musical Raw Representation



- Midi number set Ω: $\{21, 22, \ldots, 108\}$. Correspond to 88 piano keys.
- N: Length of choral piece.
- Midi matrix $X \in \Omega^{4 \times N}$: Musical raw representation.

Interpretable Feature Representation

- Each sonority p: $p \in \Omega^4$. Each column in Midi matrix.
- Part selection window $w_I : \Omega^4 \longmapsto \Omega^{|I|}$. $w_{1,4}(p) = (p_1, p_4)$.
- Basic descriptor B: $\{order, diff, sort, mod_{12}\}$. Atomic arithmetic operations.
- Descriptor with length k: $d_{(k)} = b_k \circ \cdots \circ b_1, b_i \in B$
- All windows: $W = \{w_i | I \in 2^{\{1,2,3,4\}} \backslash \{\emptyset\}\}$
- All descriptors with length k: $D^{[k]} = \{d_{(k')} | 0 \leq k' leq k\}$
- Feature universe: $\Phi = \{d \circ w | w \in W, d \in D^{[k]}\}$
- Any feature $\phi \in \Phi$, $\Omega^4 \longmapsto \phi\ (\Omega^4)$

Outline

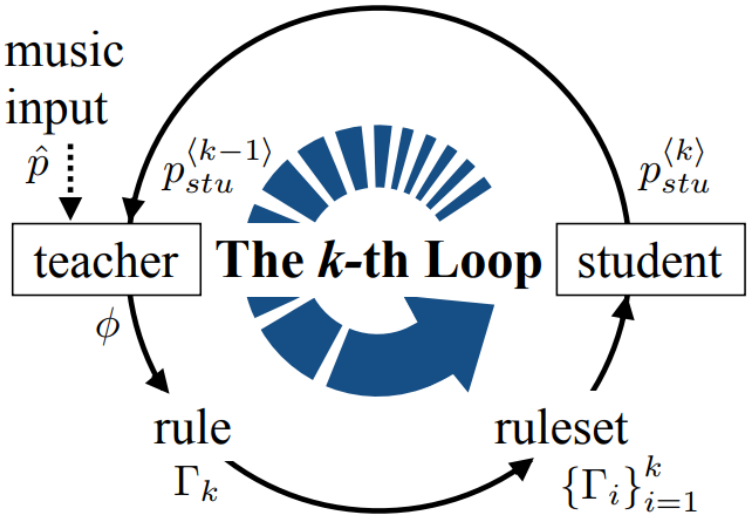
Self-Learning Loop

The teacher solves:

maximize $D\left(p_{\phi,stu}^{\langle k-1\rangle} \,||\, \hat{p}_{\phi}\right)$

subject to $\phi \in \Phi \backslash \Phi^{\langle k-1\rangle}$

(discrete optimization)



The student solves:

maximize $S_q\left(p_{stu}^{\langle k\rangle}\right)$

subject to $p_{stu}^{\langle k\rangle} \in \Gamma_1$

$\cdots$

$p_{stu}^{\langle k\rangle} \in \Gamma_k$

(linear least-squares)

- S_q: Tsallis entropy, which achieves a maximum when p is uniform.
- Γ_k: kth rule, containing $(\phi_i, \hat{p}_{\phi_i})$.
- p_{stu}^{k}: Sonority distribution. (n-gram)
- $p_{\phi,stu}^{<k-1>}$: Feature distribution.
- To get the feature distribution p_ϕ from sonority distribution $p(x)$:

$$p_\phi(y) = \sum_{x \in \phi^{-1}(\{y\})} p(x)$$

Outline

Feature Induced Partition

- **Feature induced partition of the input domain Ω^4:**

$$P_\phi = \{\phi^{-1}(\{y\}) | y \in \phi(\Omega^4)\}$$

- For two partitions $P, Q \in P_\phi$, P is finer than Q (Q is coarser), written as $P \succeq Q$, if for all p, $p' \in \Omega_4$, p, p' are in the same cluster under P $=>$ p, p' are in the same cluster under Q.
  P is strictly finer is written as $P \succ Q$.

Conceptual Hierarchy

Based on the relation $\succ$, construct the conceptual hierarchy for the parition family P_ϕ as a directed acyclic graph:

- Node: partitions.
- Edge: there's an edge between any pair of nodes v, v' if and only if the partition referred by v is (strictly) finer than v'.
- A higher level feature induces a coarser partition.

Informational Hierarchy

- Given a trace of extracted rules by the kth iteration of loop, a feature ϕ is informationally implied from the extracted rules with tolerance $\gamma > 0$ if:

$$gap(p^{<k>}_{\phi,dtu} || \hat{p}_\phi) := D(p^{<k>}_{\phi,stu} || \hat{p}_\phi) < \gamma$$

and

$$gap(p^{<k'>}_{\phi,stu} || \hat{p}_\phi) \geq \gamma, \forall k' < k$$

- This rule is beyond user's satisfaction.

Hierarchical Filters

- Add hierarchical filters when selecting rules to prune hierarchically entangled features and speed up feature selection.
- Teacher's optimization can be represented as:

$$\underset{\phi \in \mathbf{\Phi}}{maximize}\, gap(p_{\phi,stu}^{<k-1>} || \hat{p}_\phi)$$

$$H(\hat{p}_\phi) \leq \delta$$

$$\phi \notin C^{(k-1)} := \{\phi | P_\phi \prec P_{\phi'}, \phi' \in \mathbf{\Phi}^{<k-1>}\}$$

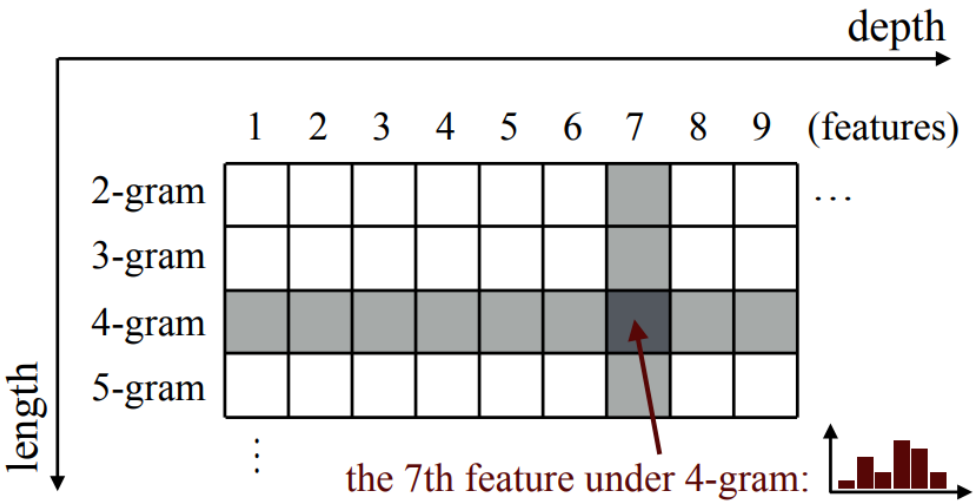
$$\phi \notin I^{(k-1)} := \{\phi | gap(p_{\phi,stu}^{<k-1>} || \hat{p}_\phi) < \gamma\}$$

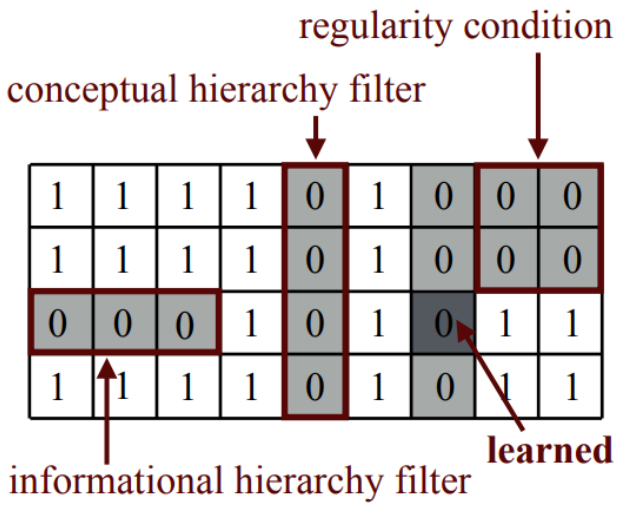
- $H(\hat{p}_\phi)$ constraint here is based on 'rules should be relatively easy to learn'.
- By selecting hyper-parameters γ and δ, can control the learning pace.

Outline

Two-Dimensional Memory selection

Rule selecting task: pick the right feature under the right *n* gram.





Outline

Application

Table 1: Customizing a syllabus (* signifies rules that are skipped in the faster pace)

	Rule Trace	**Faster Pace** ($\gamma = 0.5$)	**Slower Pace** ($\gamma = 0.1$)
1		$\texttt{order} \circ w_{\{1,2,3,4\}}$	$\texttt{order} \circ w_{\{1,2,3,4\}}$
2		$\texttt{mod}_{12} \circ w_{\{1\}}$	$\texttt{order} \circ \texttt{diff} \circ \texttt{sort} \circ w_{\{1,2,4\}}$*
3		$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{2,3\}}$	$\texttt{order} \circ \texttt{diff} \circ \texttt{mod}_{12} \circ w_{\{1,2,3\}}$*
4		$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{3,4\}}$	$\texttt{order} \circ \texttt{diff} \circ \texttt{diff} \circ w_{\{1,2,3,4\}}$*
5		$\texttt{diff} \circ \texttt{sort} \circ w_{\{2,3\}}$	$\texttt{order} \circ \texttt{sort} \circ \texttt{mod}_{12} \circ w_{\{2,3,4\}}$*
6		$\texttt{mod}_{12} \circ w_{\{3\}}$	$\texttt{order} \circ \texttt{sort} \circ \texttt{mod}_{12} \circ w_{\{1,3,4\}}$*
7		$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{1,2\}}$	$\texttt{order} \circ \texttt{sort} \circ \texttt{mod}_{12} \circ w_{\{1,2,3,4\}}$*
8		$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{2,4\}}$	$\texttt{mod}_{12} \circ w_{\{1\}}$
9		$\texttt{diff} \circ w_{\{1,2\}}$	$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{2,3\}}$
10		$\texttt{diff} \circ \texttt{sort} \circ w_{\{1,3\}}$	$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{3,4\}}$

Skip rules that are not so important.

Fundamentals: Hierarchical 1-gram

For fundamentals, knowledge independent of the context: 1-gram.

- With respect to a feature, record the gap between student and Bach for each iteration.

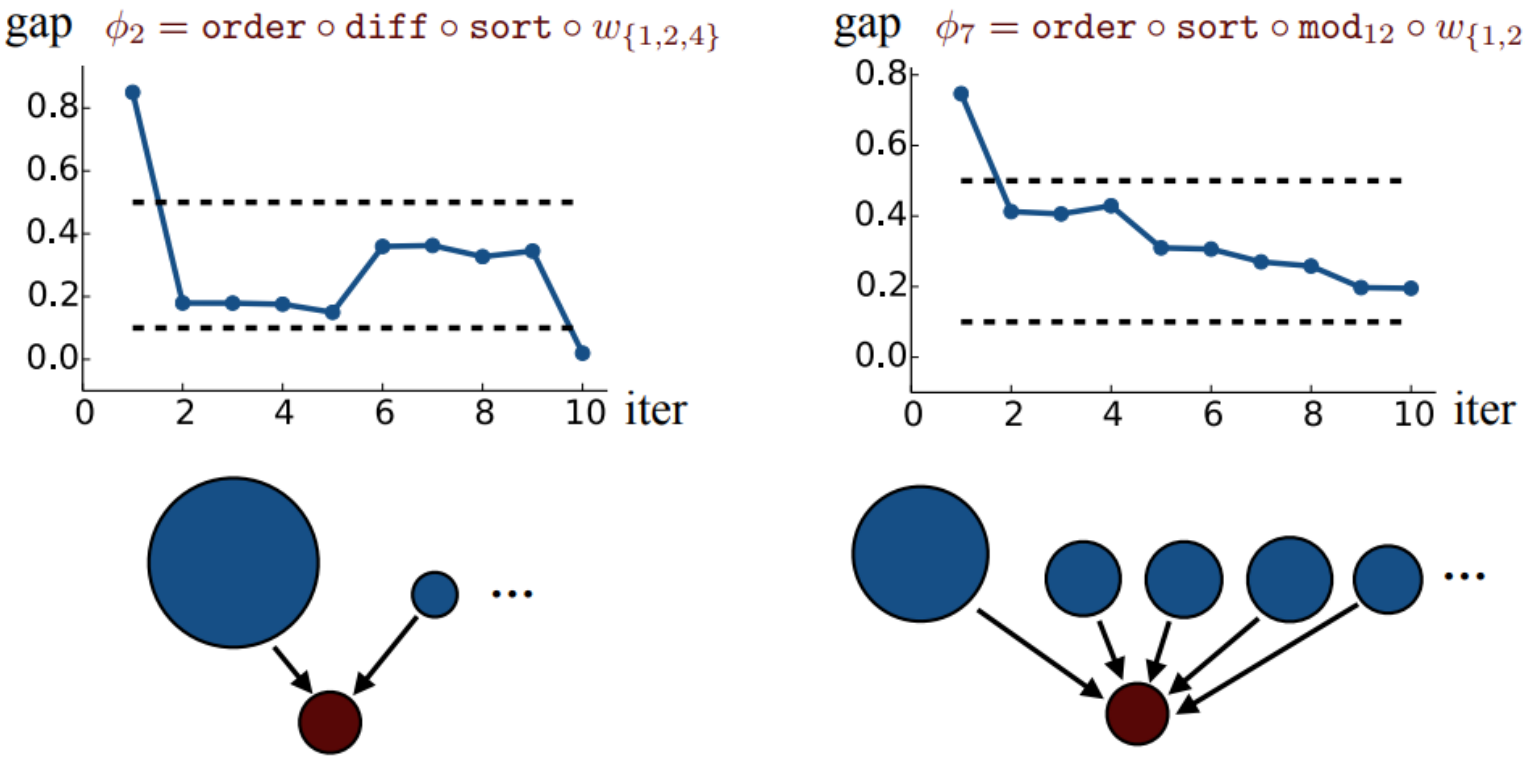


Figure 3: Gap trajectories for two features. The dashed black lines show two different satisfactory gaps ($\gamma = 0.5$ and 0.1). The bottom charts show the informationally implied hierarchies.

Part-writing: n-grams

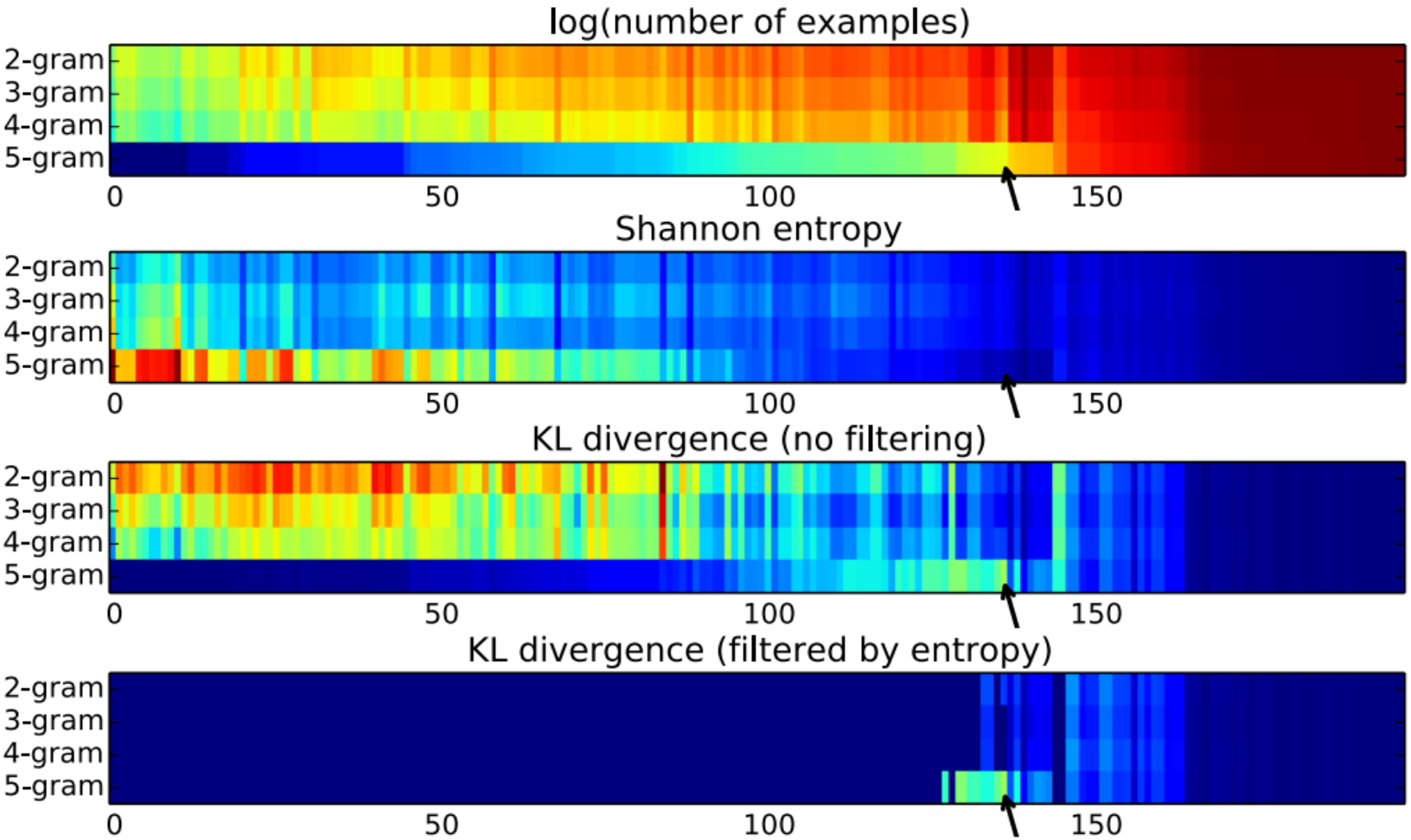


Figure 4: The relative performance of the selected rule (pointed) among the pool of all cells in the 2D memory. A desired rule has: higher confidence (measured by the number of examples, brighter regions in the first row), more regularity (measured by Shannon entropy, darker regions in the second row), and larger style gap (measured by KL divergence, brighter regions in the bottom two rows).

Visualizing Bach's Mind

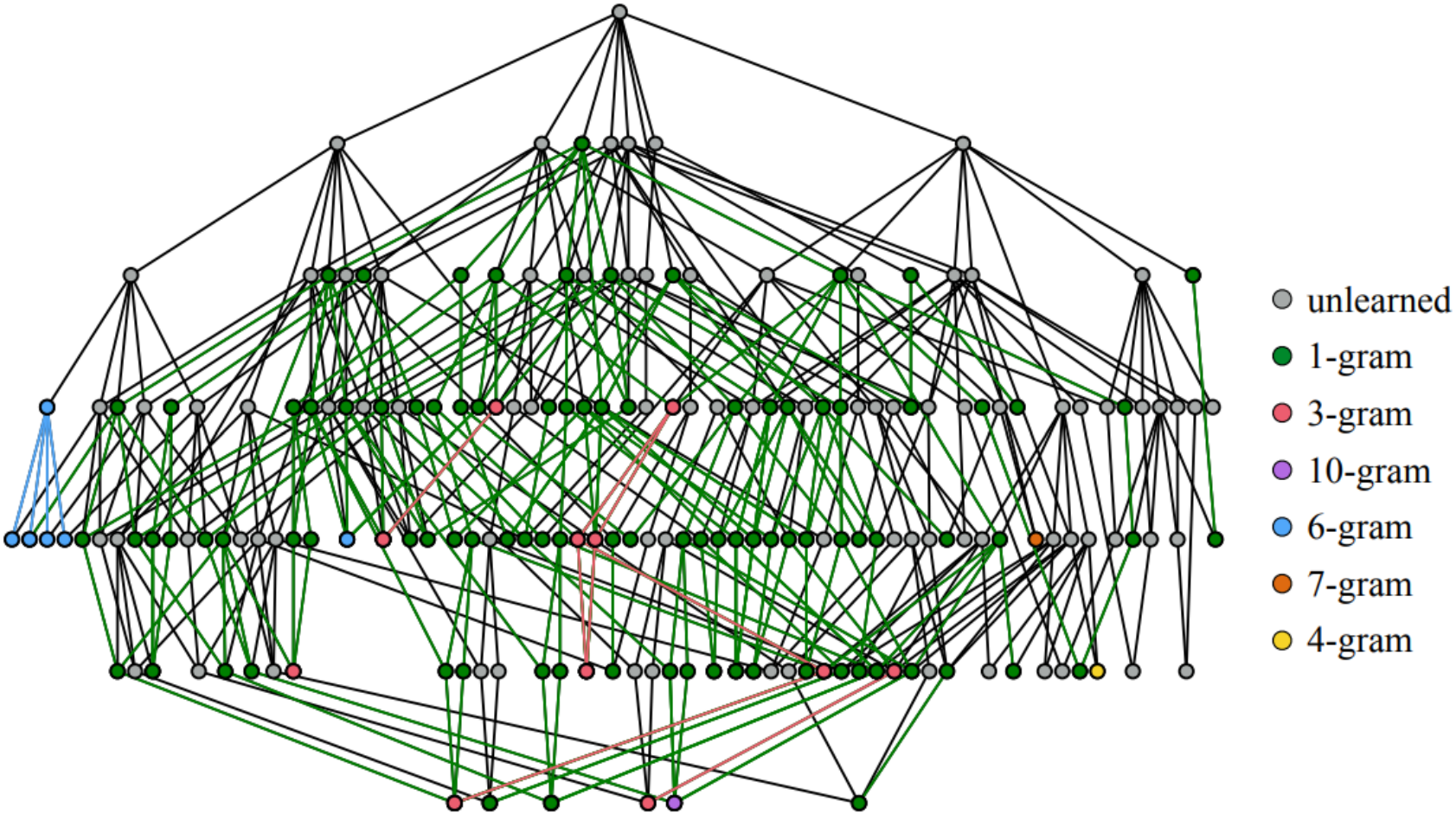


Figure 5: Visualization of Bach's music mind for writing chorales. The underlying DAG represents the conceptual hierarchy (note: edges always point downwards). Colors are used to differentiate rule activations from different n-gram settings. We have enlarged $N = \{1, 2, \ldots, 10\}$ to allow even longer-term dependencies.