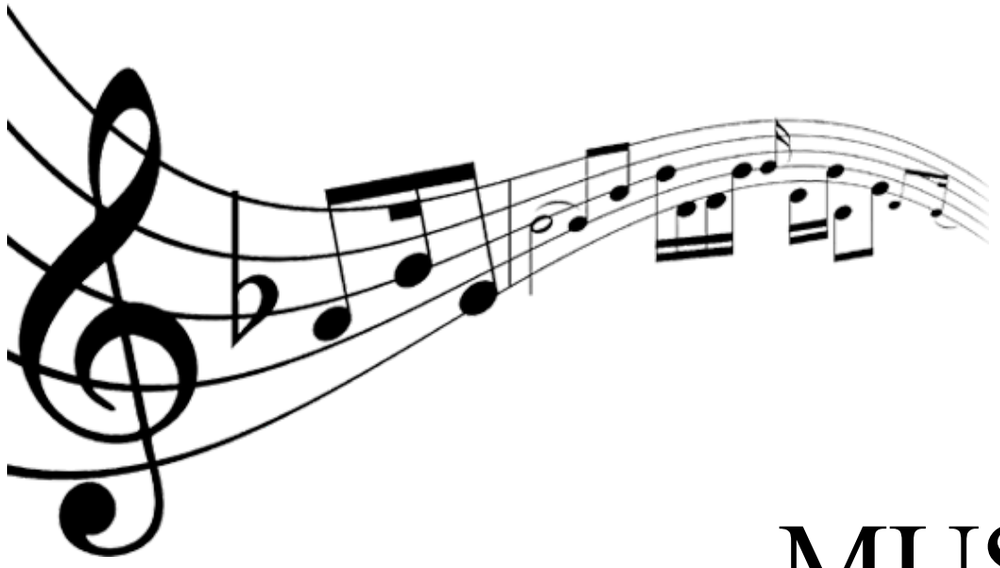
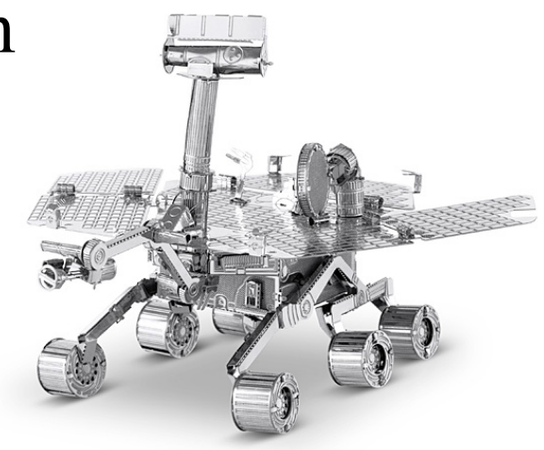


MUS-ROVER

An Automatic Music Theorist and Pedagogue

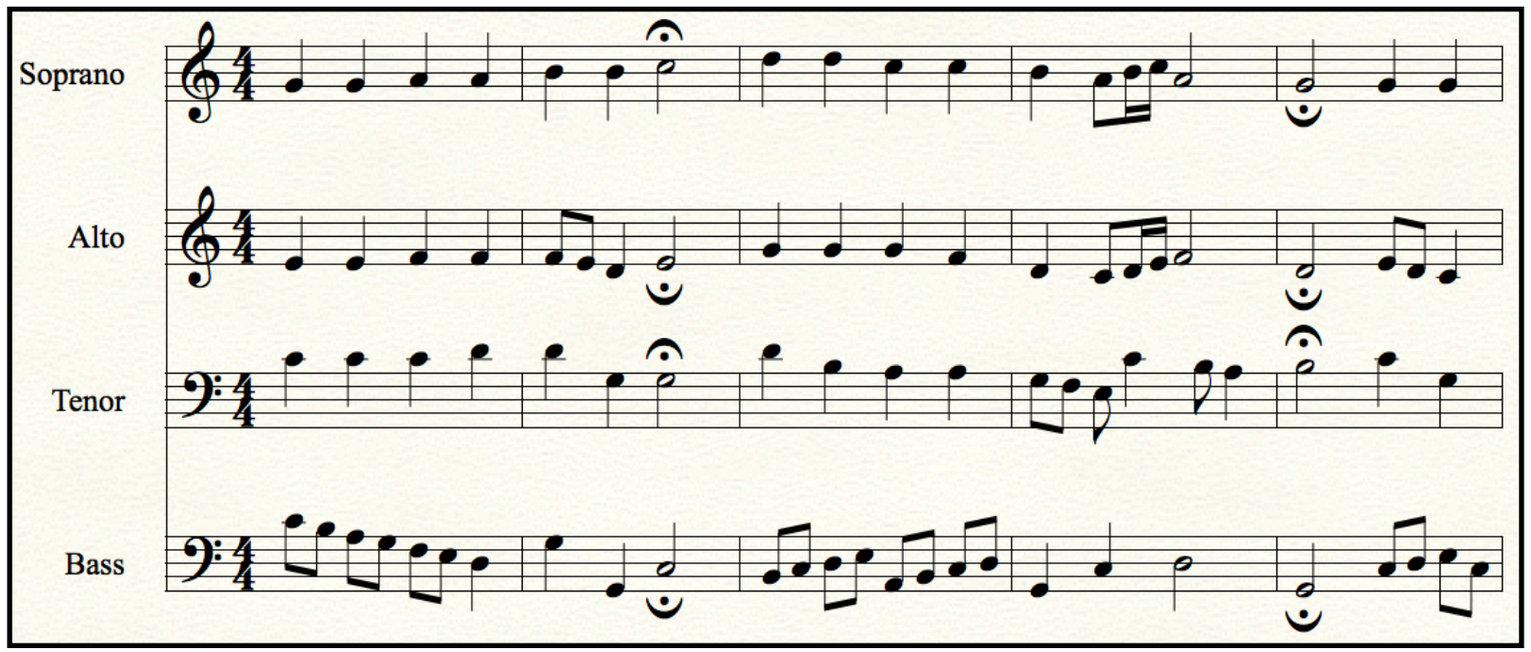
Haizi Yu, Lav R. Varshney

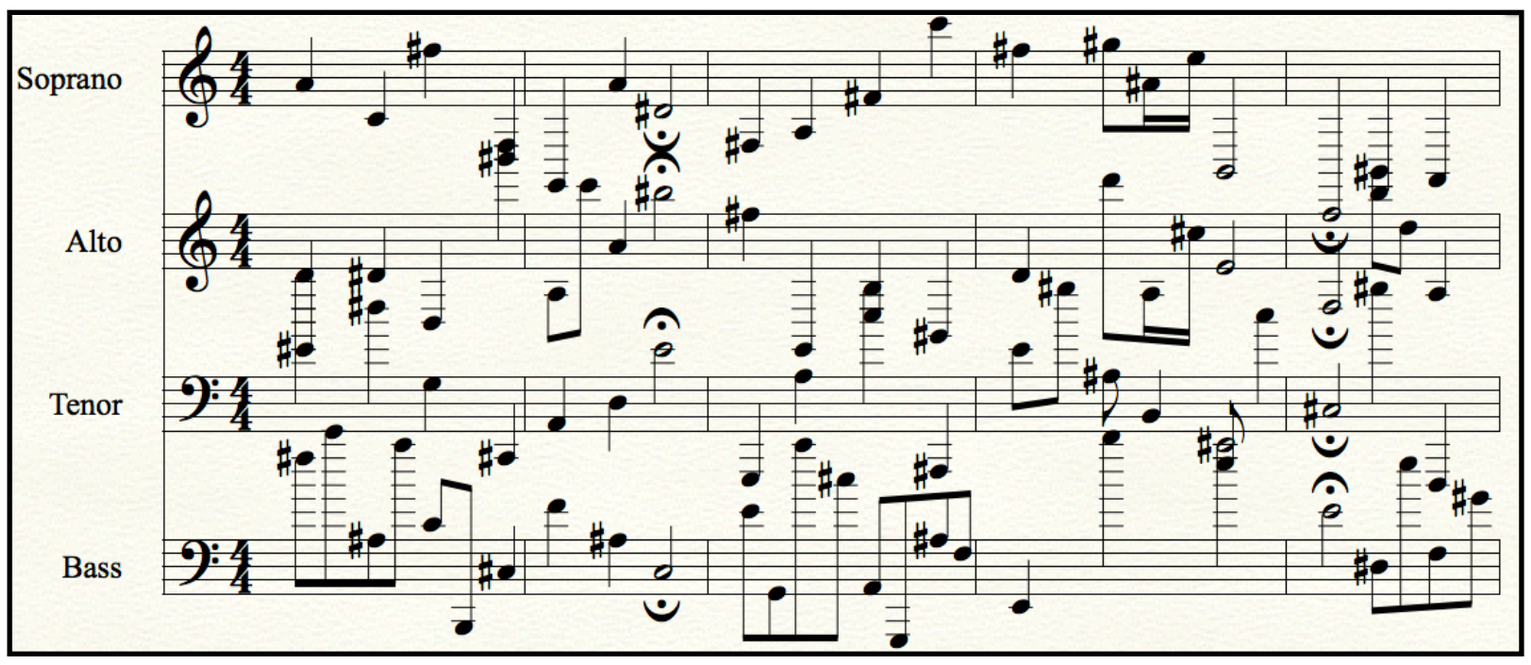
University of Illinois at Urbana-Champaign



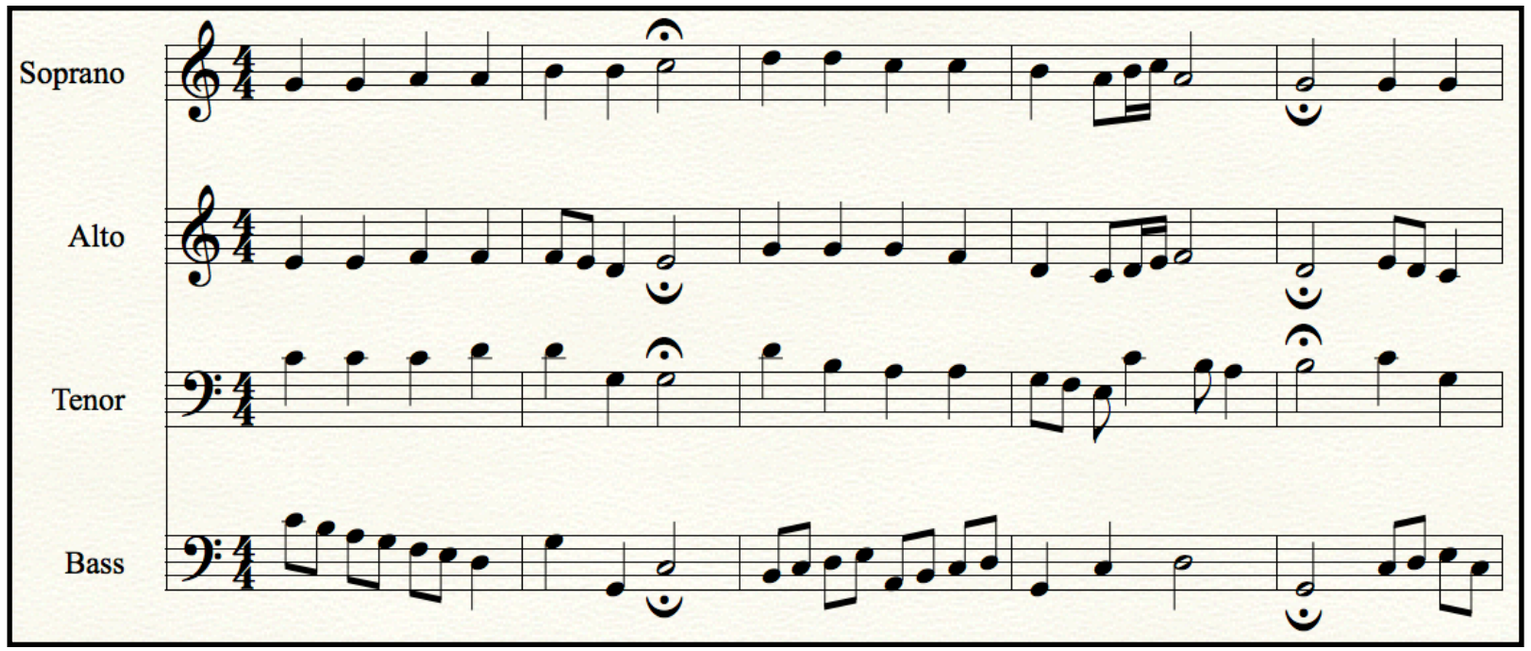
Supported in part by C3SR

A Music Story

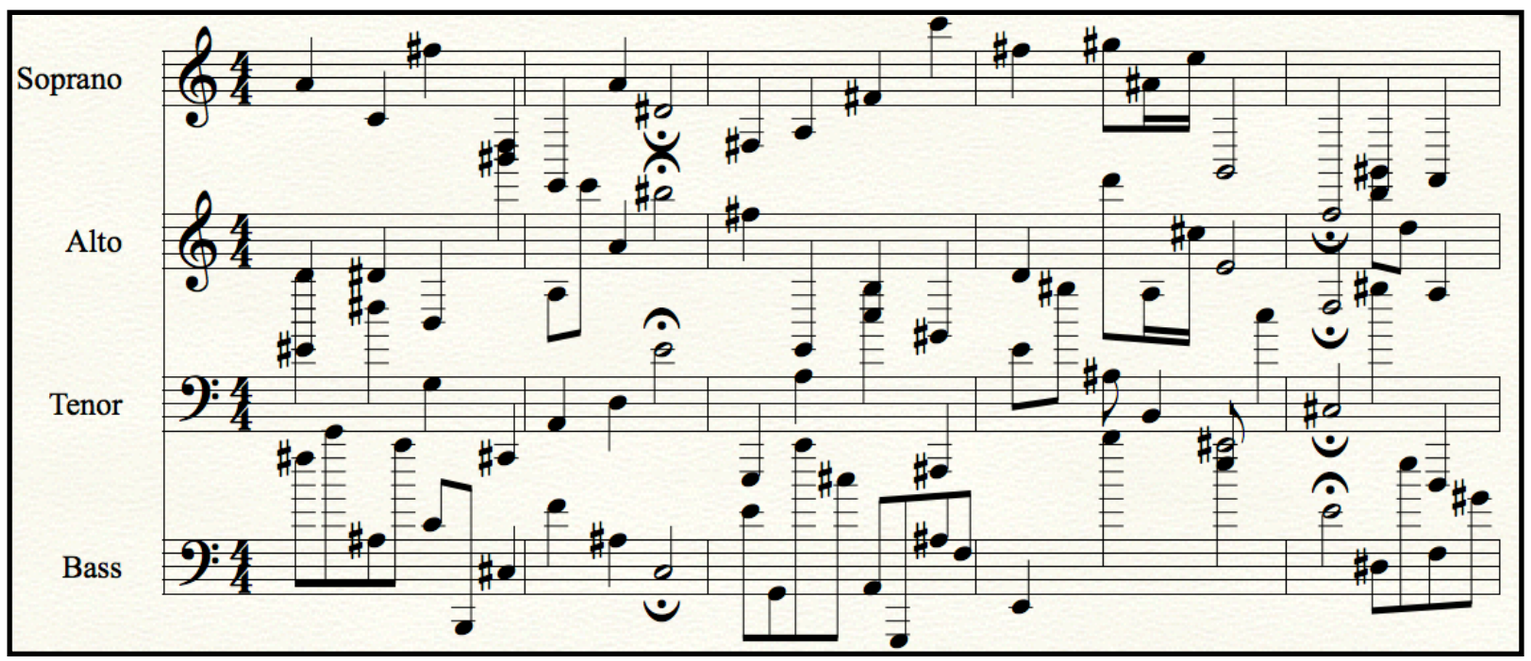




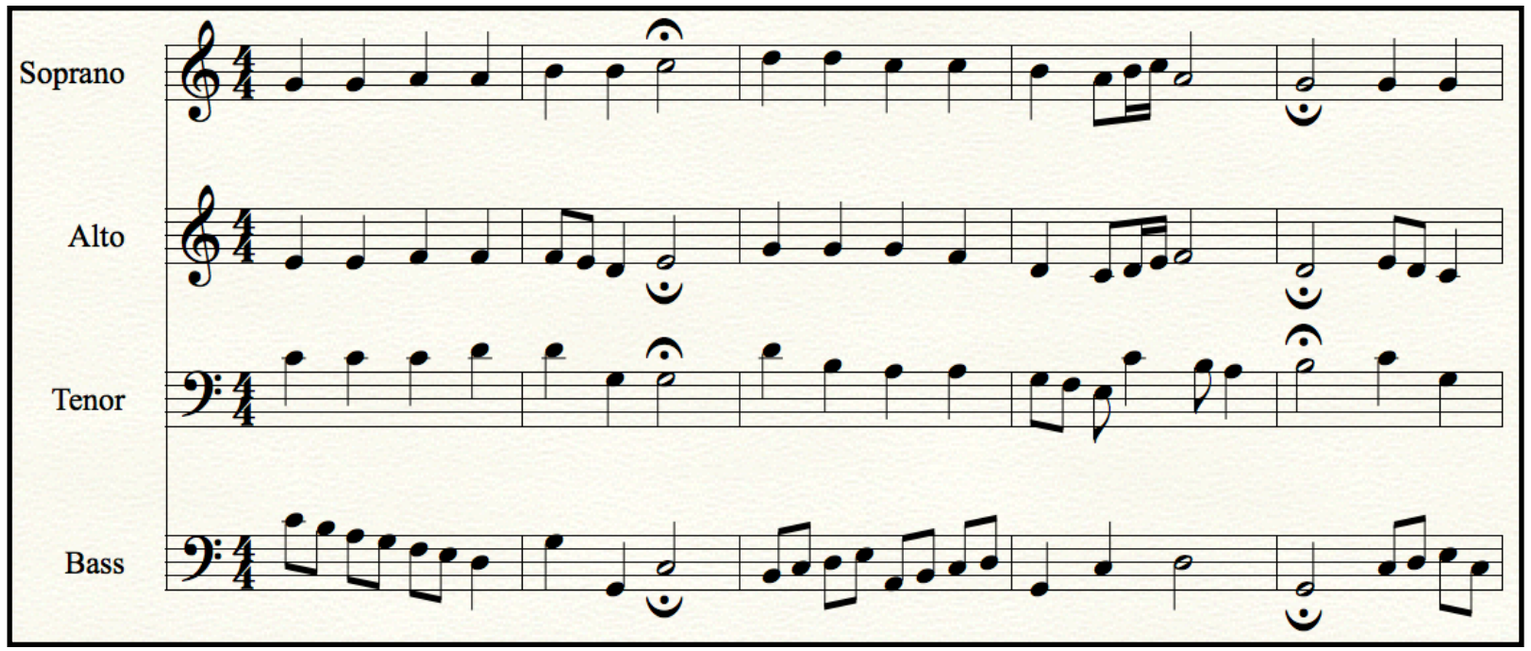
A Music Story



BWV 124

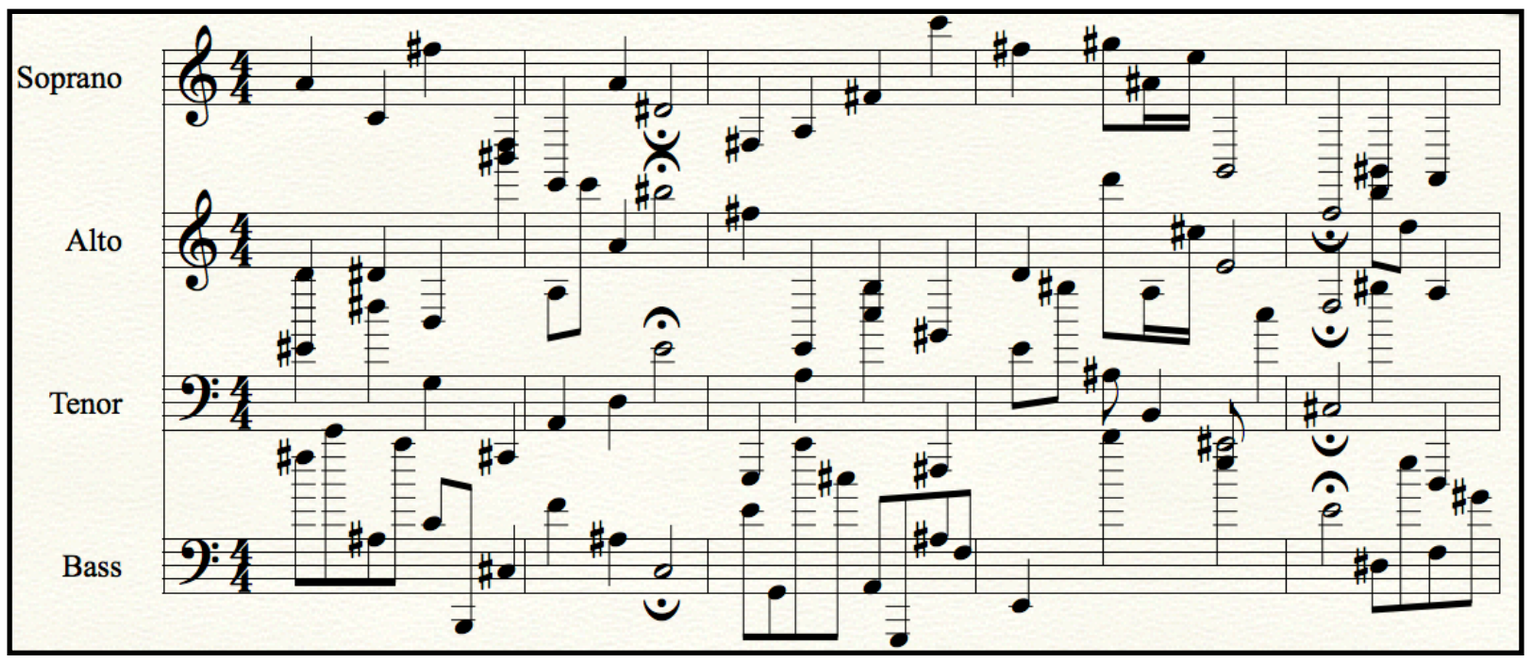


A Music Story

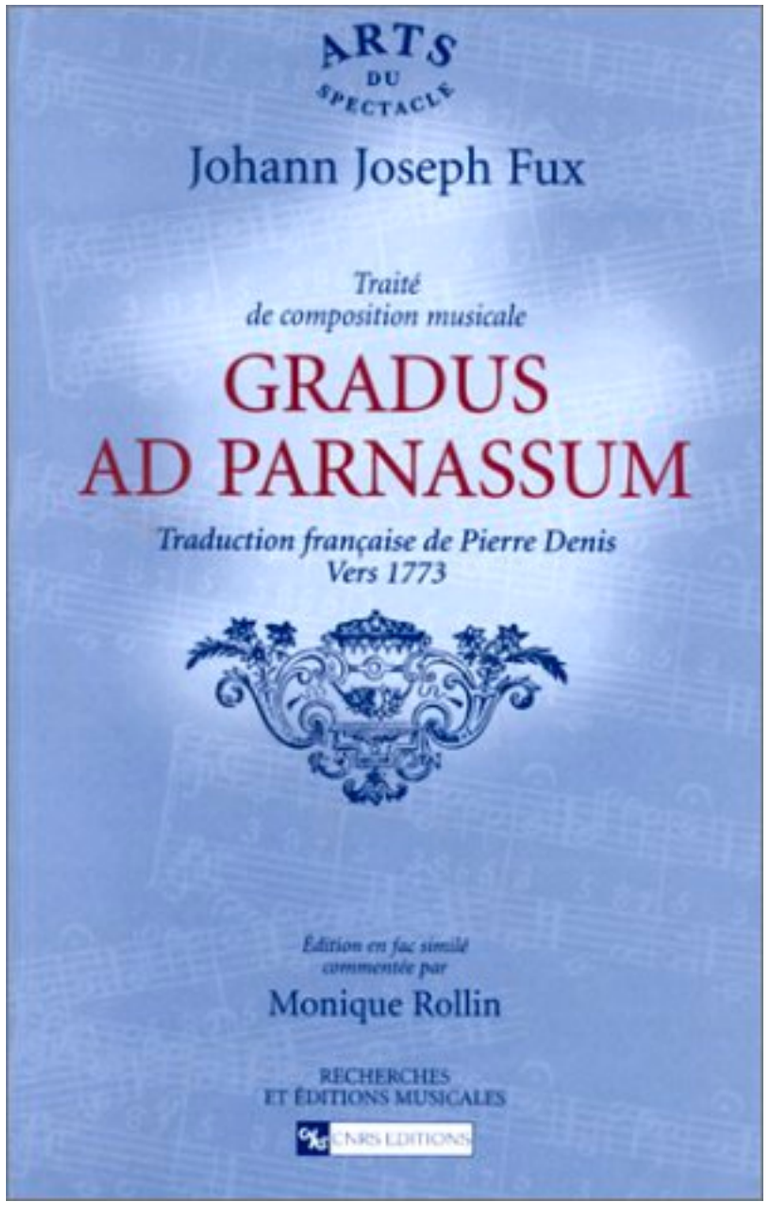


BWV 124

?



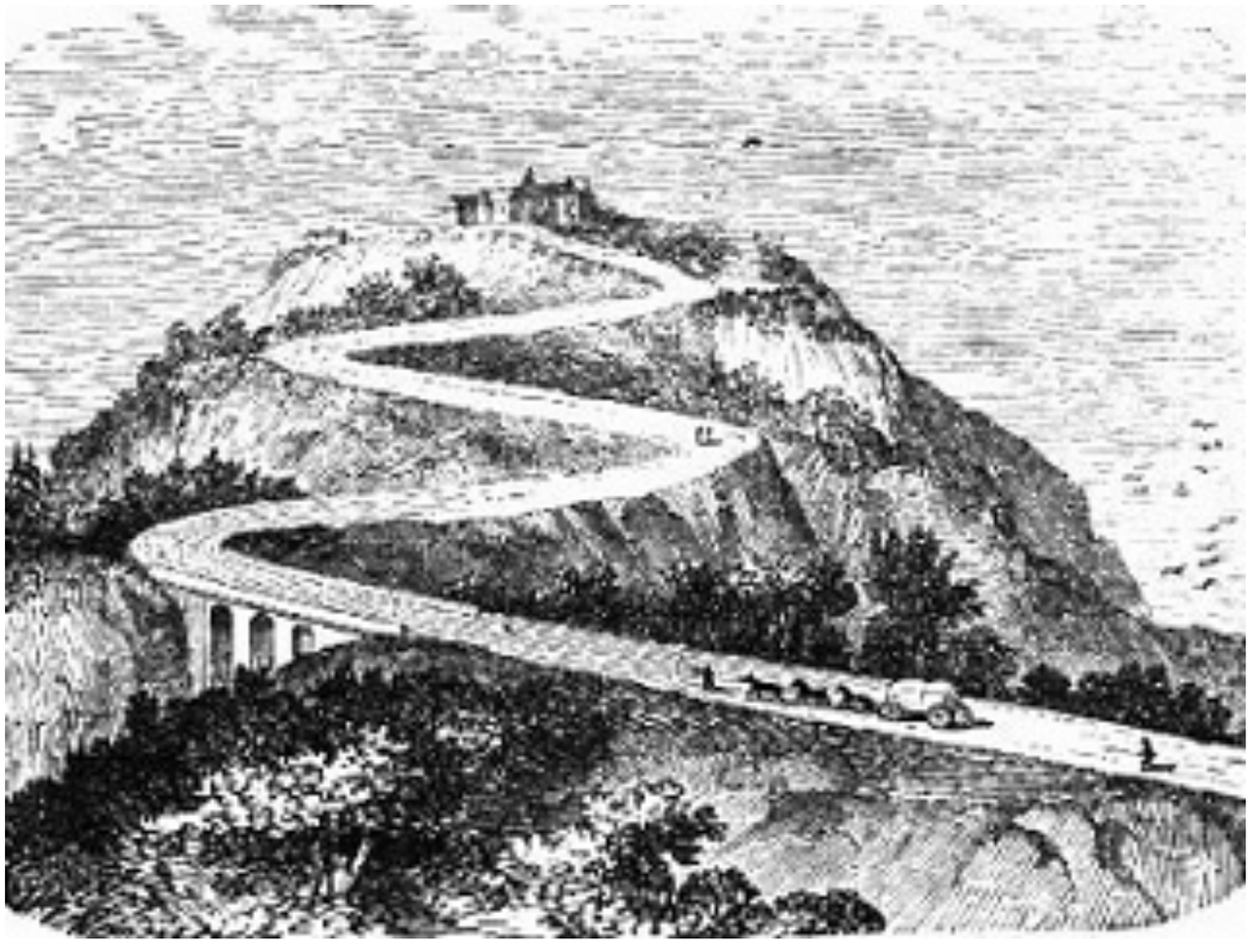
What makes music music?


ARTS DU SPECTACLE
Johann Joseph Fux
Traité
de composition musicale
GRADUS
AD PARNASSUM
Traduction française de Pierre Denis
Vers 1773
Édition en fac similé
commentée par
Monique Rollin
RECHERCHES
ET ÉDITIONS MUSICALES
CNRS EDITIONS

Gradus ad Parnassum

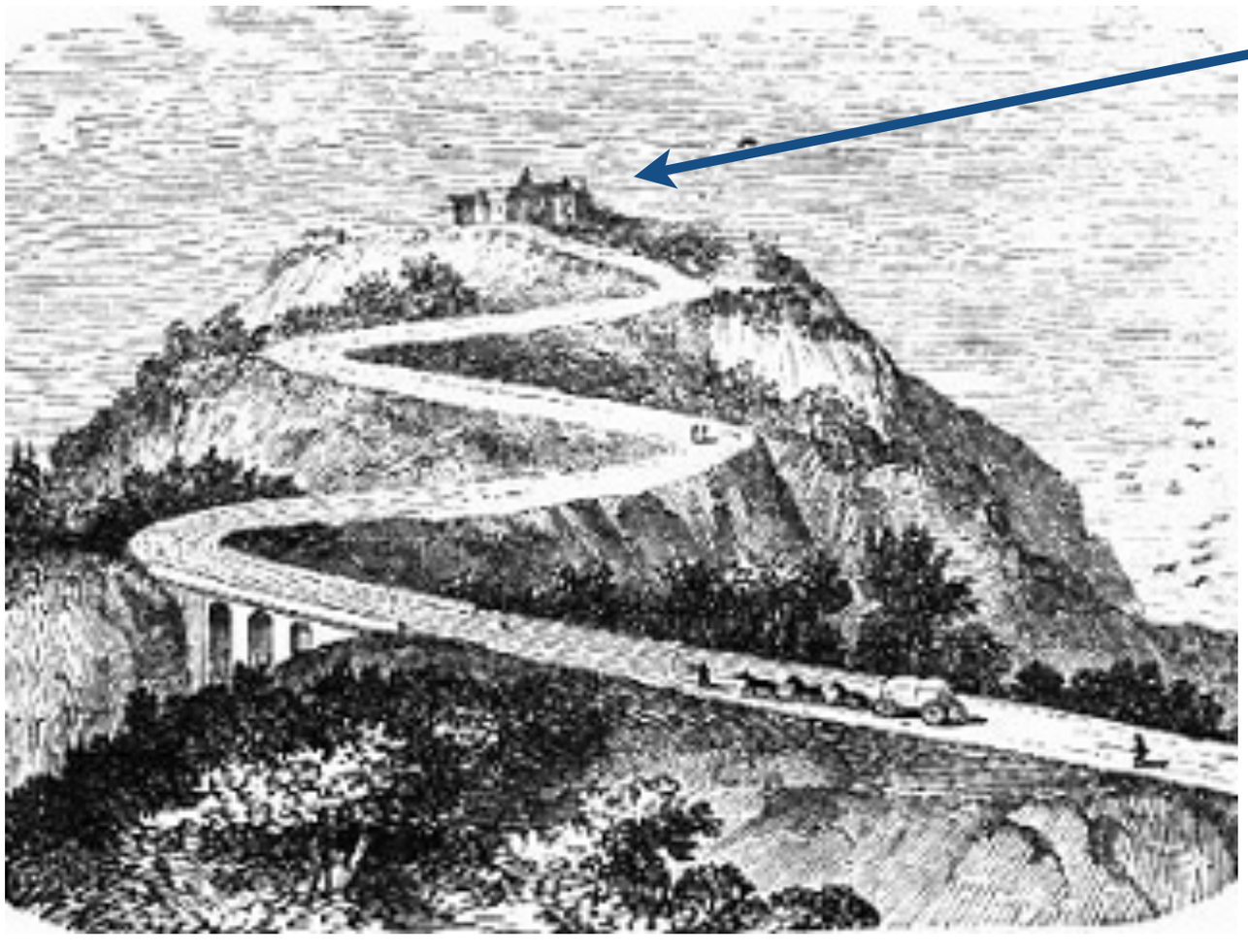
Gradus ad Parnassum

“The path towards Mount Parnassus”



Gradus ad Parnassum

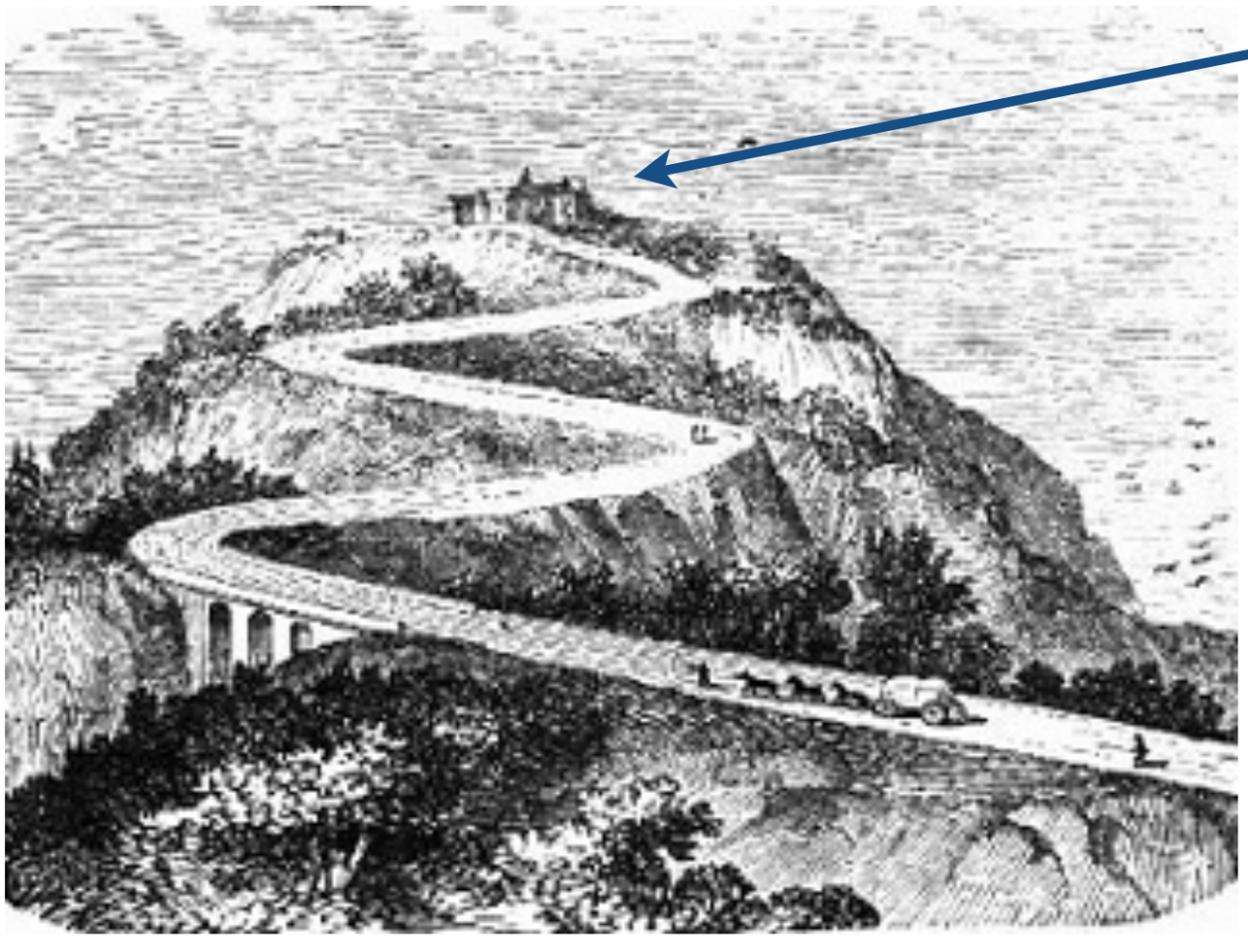
“The path towards Mount Parnassus”



the home of the muses

Gradus ad Parnassum

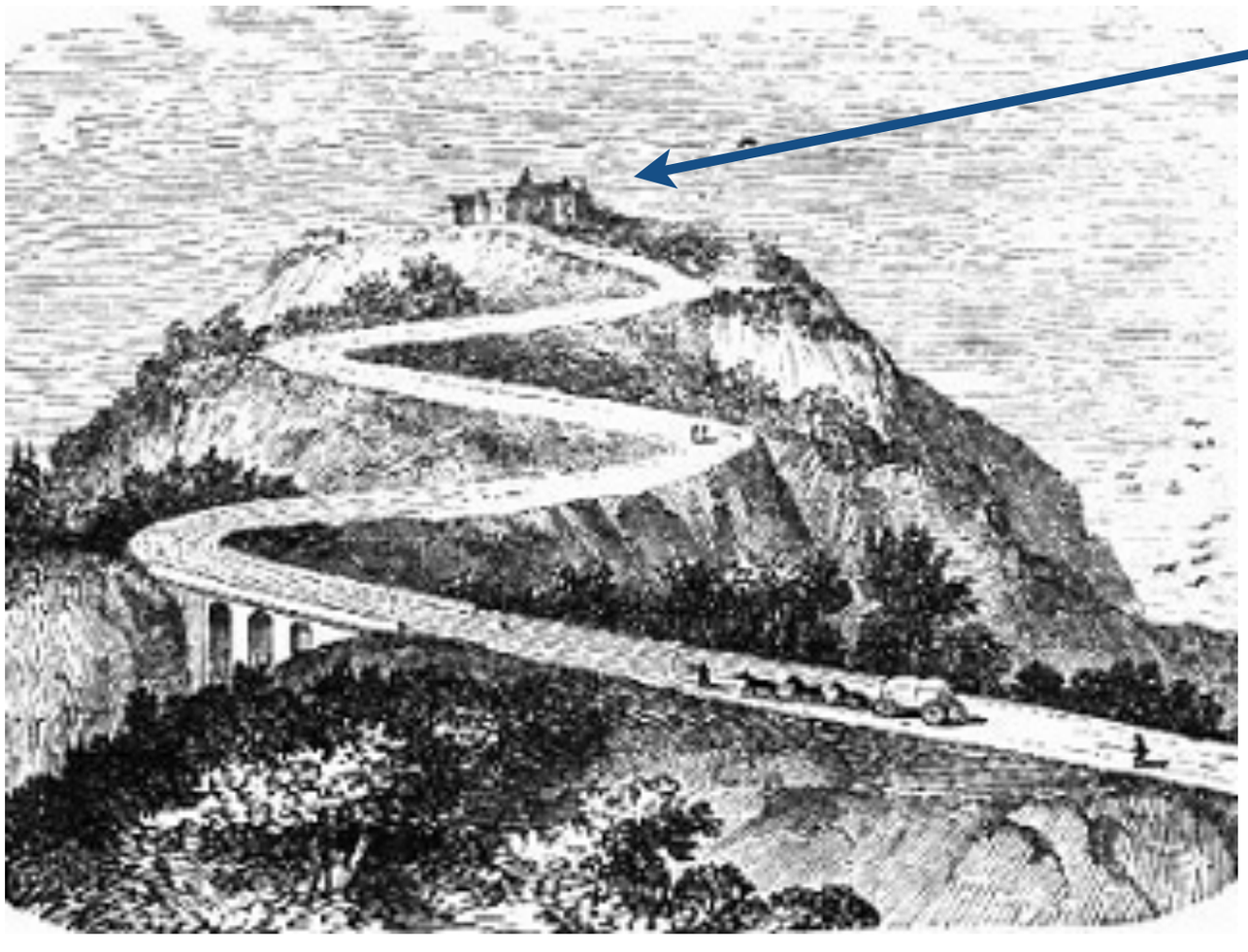
“The path towards Mount Parnassus”



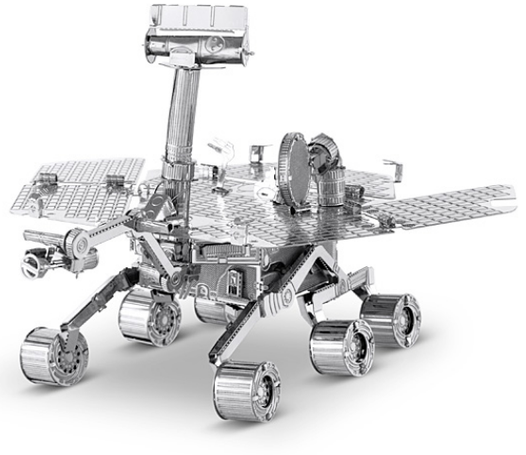
the home of the muses

Gradus ad Parnassum

“The path towards Mount Parnassus”



the home of the muses



MUS-ROVER

Automatic Pathfinder to Mount Parnassus

About

Demo

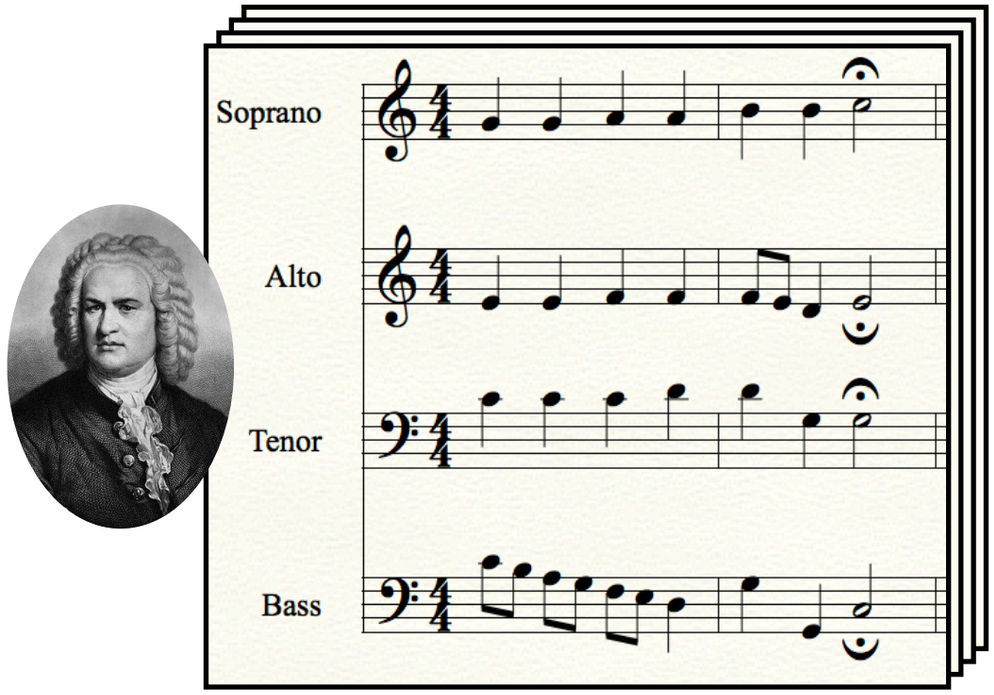
Team

Contact us: mus.rover.2014@gmail.com | University of Illinois at Urbana-Champaign

A New Learning Problem

Automatic Concept Learning

Input



Output

Concepts

Rules

Laws

…

A New Learning Paradigm

New Concept Representation

New Learning Algorithm

Representation: Data Space

Data space: (X, p_X) or (X, p) for short

Representation: Data Space

Data space: (X, p_X) or (X, p) for short

Assume a data point $x \in X$ is an i.i.d. sample drawn from a probability distribution p.

Representation: Data Space

Data space: (X, p_X) or (X, p) for short

Assume a data point $x \in X$ is an i.i.d. sample drawn from a probability distribution p.

However, the data distribution p, or an estimation of it, is **known**.

The goal here is not to estimate p but to **explain** it.

Representation: Chord Space

Considering **chords** from Bach's four-part chorales recorded in sheet music.

Here, a **chord** is any collection of four simultaneously sounding pitches.

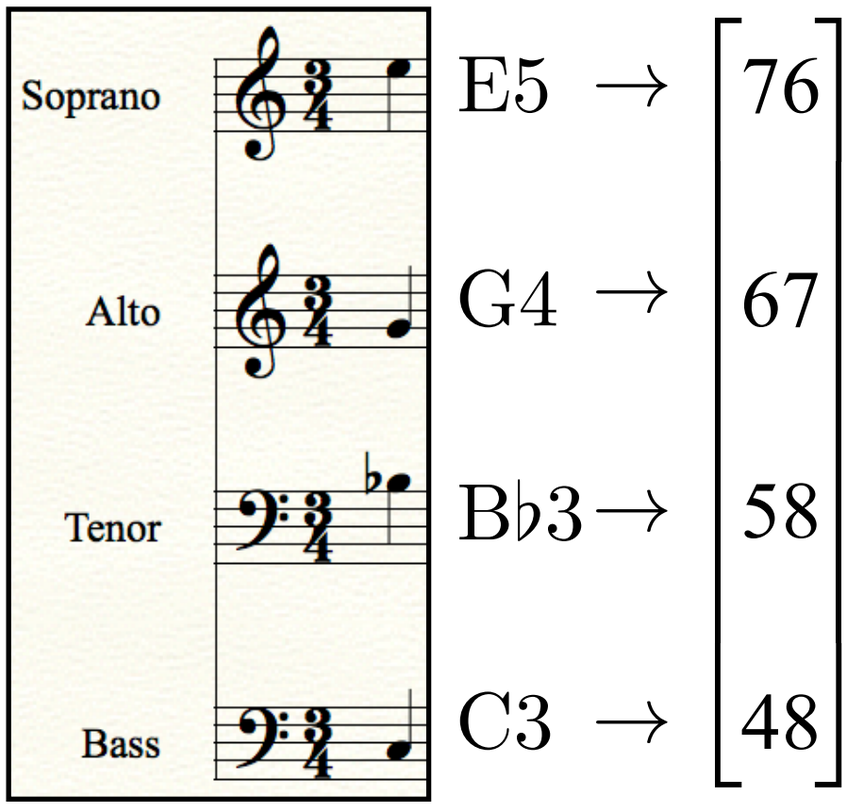
Chord space: $X = \mathbb{Z}^4$

chord: $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \in X$

pitch: $x_i \in \mathbb{Z} \quad (\text{C4} \rightarrow 60)$

voice: $i \in \{1, 2, 3, 4\}$

S A T B



Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

cells (or less formally, clusters)

Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

cells (or less formally, clusters)

An **concept** is a partition cell.

Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

cells (or less formally, clusters)

An **concept** is a partition cell.

A **partition matrix** A is a concise way of representing an abstraction $\mathcal{A}$.

Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

cells (or less formally, clusters)

An **concept** is a partition cell.

A **partition matrix** A is a concise way of representing an abstraction $\mathcal{A}$.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Representation: Abstraction

An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

cells (or less formally, clusters)

An **concept** is a partition cell.

A **partition matrix** A is a concise way of representing an abstraction $\mathcal{A}$.

$$A = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{r} \\ \text{1st cell} \\ \text{2nd cell} \\ \text{3rd cell} \end{array}$$

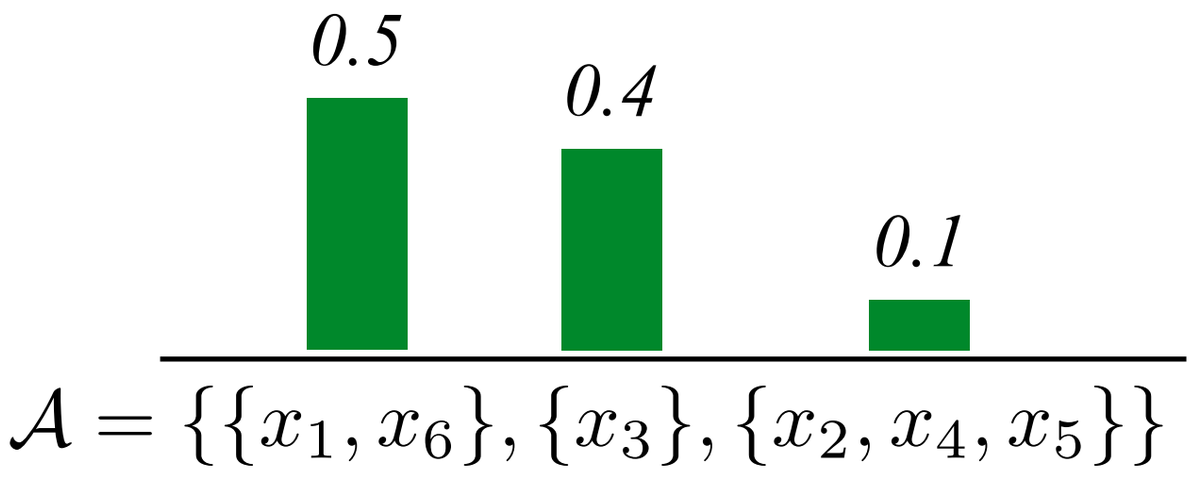
Representation: Probabilistic Rules

A **probabilistic rule** is a pair:

$$(\mathcal{A}, p_{\mathcal{A}})$$

where $\mathcal{A}$ is an abstraction (partition);

$p_{\mathcal{A}}$ is a probability distribution over the abstracted concepts (cells).



Representation Summary

Abstraction: a partition

Concept: a partition cell (cluster)

Probabilistic rule: abstraction & distribution

Representation Summary

Abstraction: a partition

Concept: a partition cell (cluster)

Probabilistic rule: abstraction & distribution

Automatic concept learning: ?

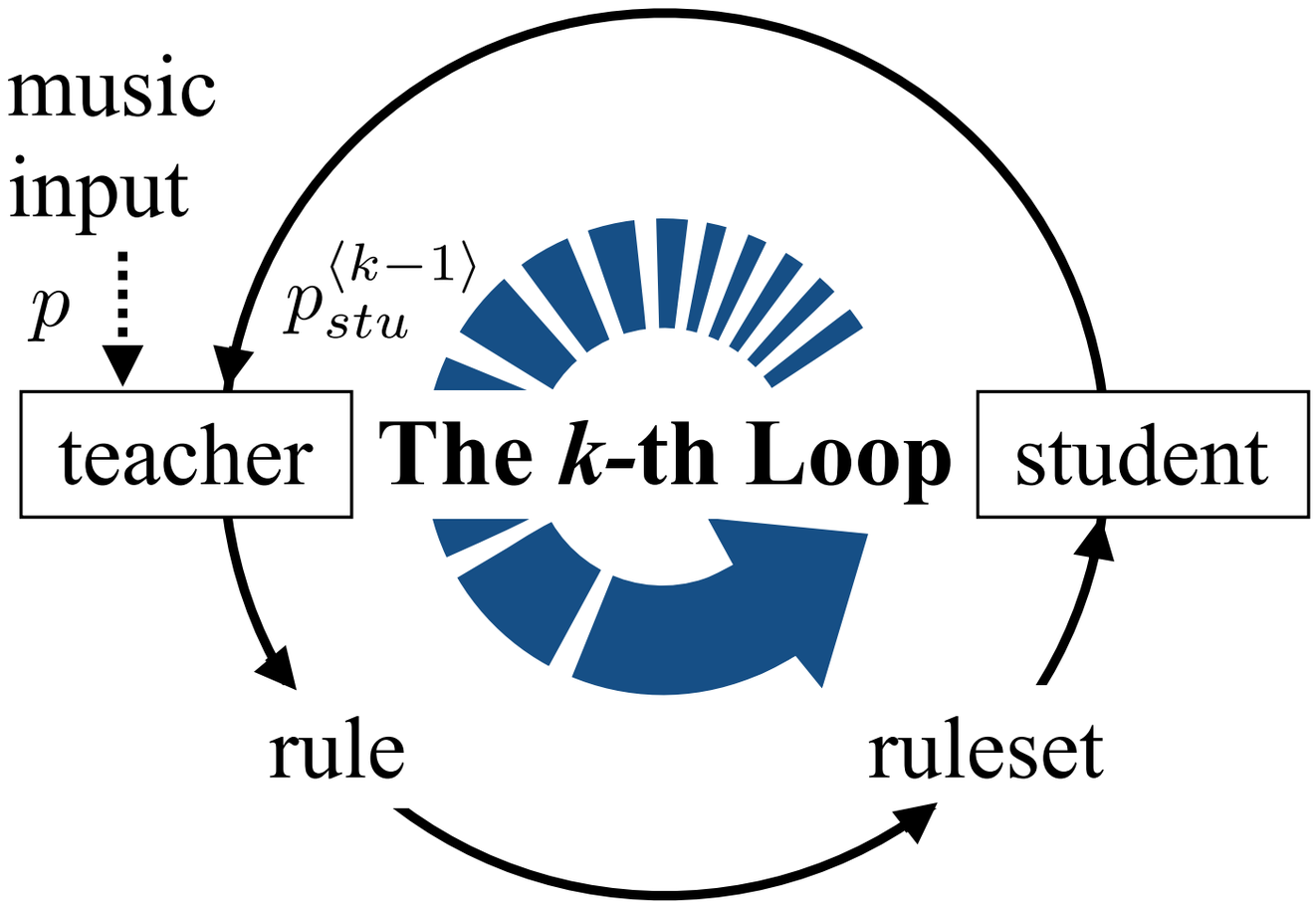
Automatic Concept Learning

is

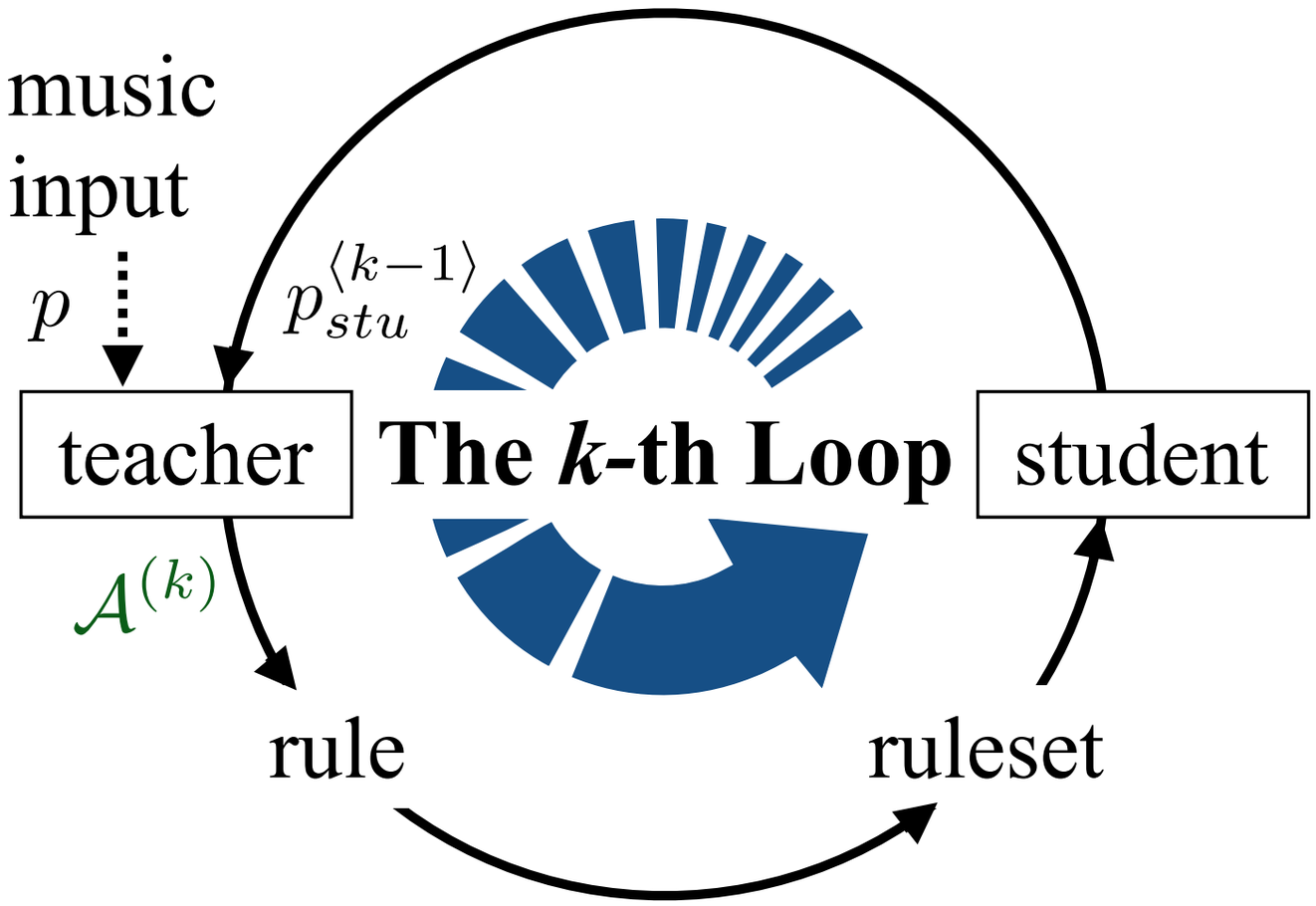
the process of learning probabilistic rules

But how?

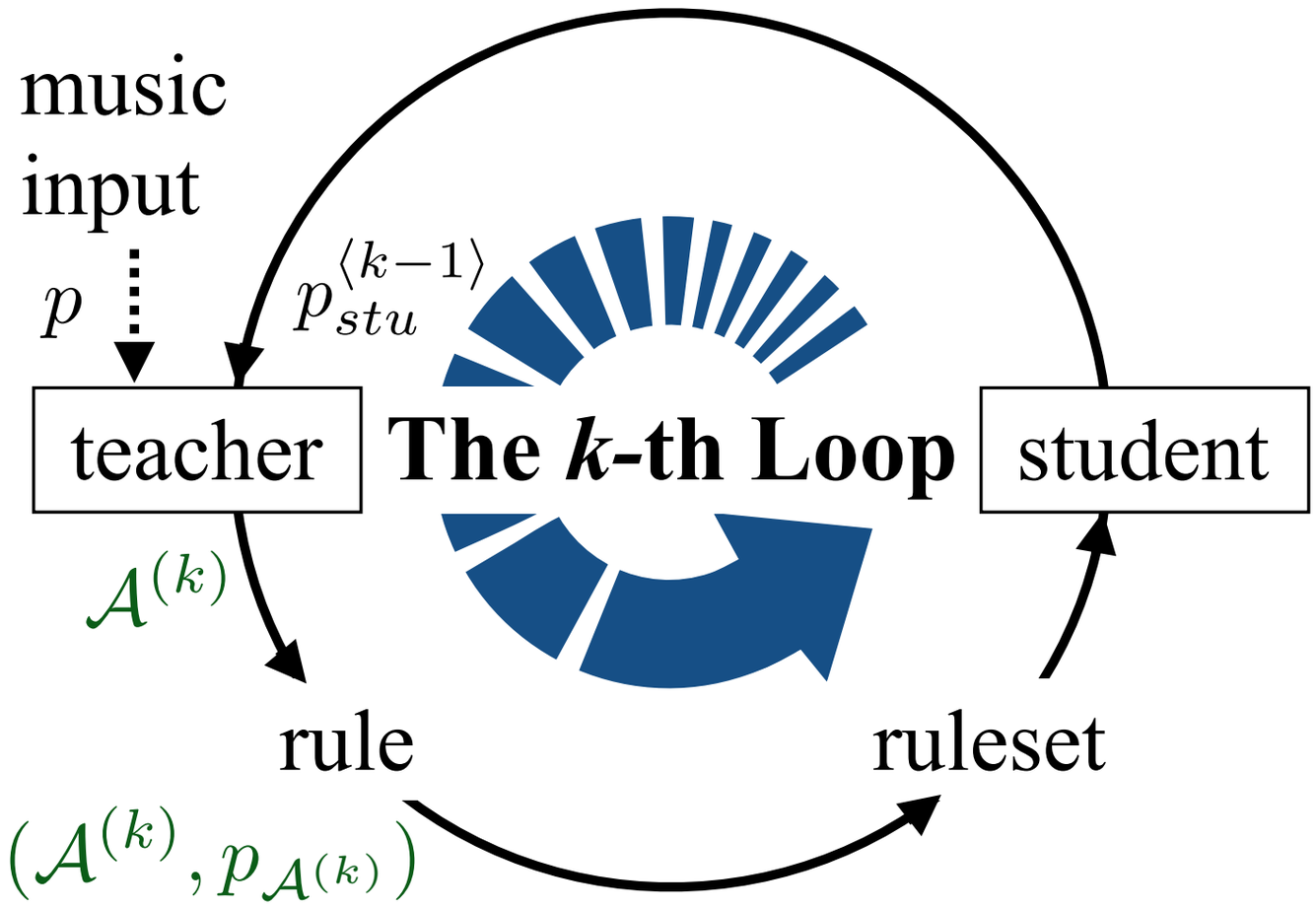
Learning Algorithm



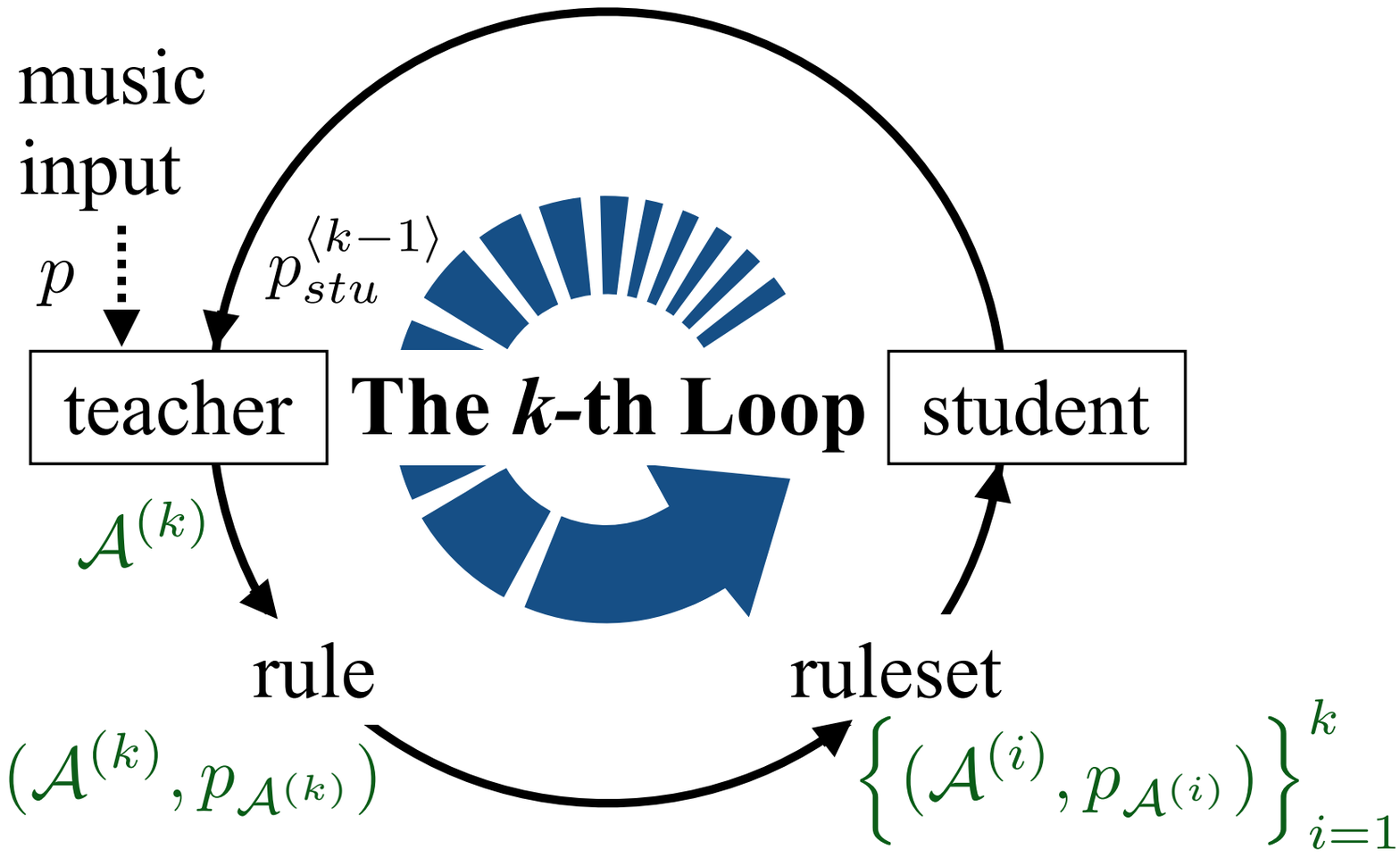
Learning Algorithm



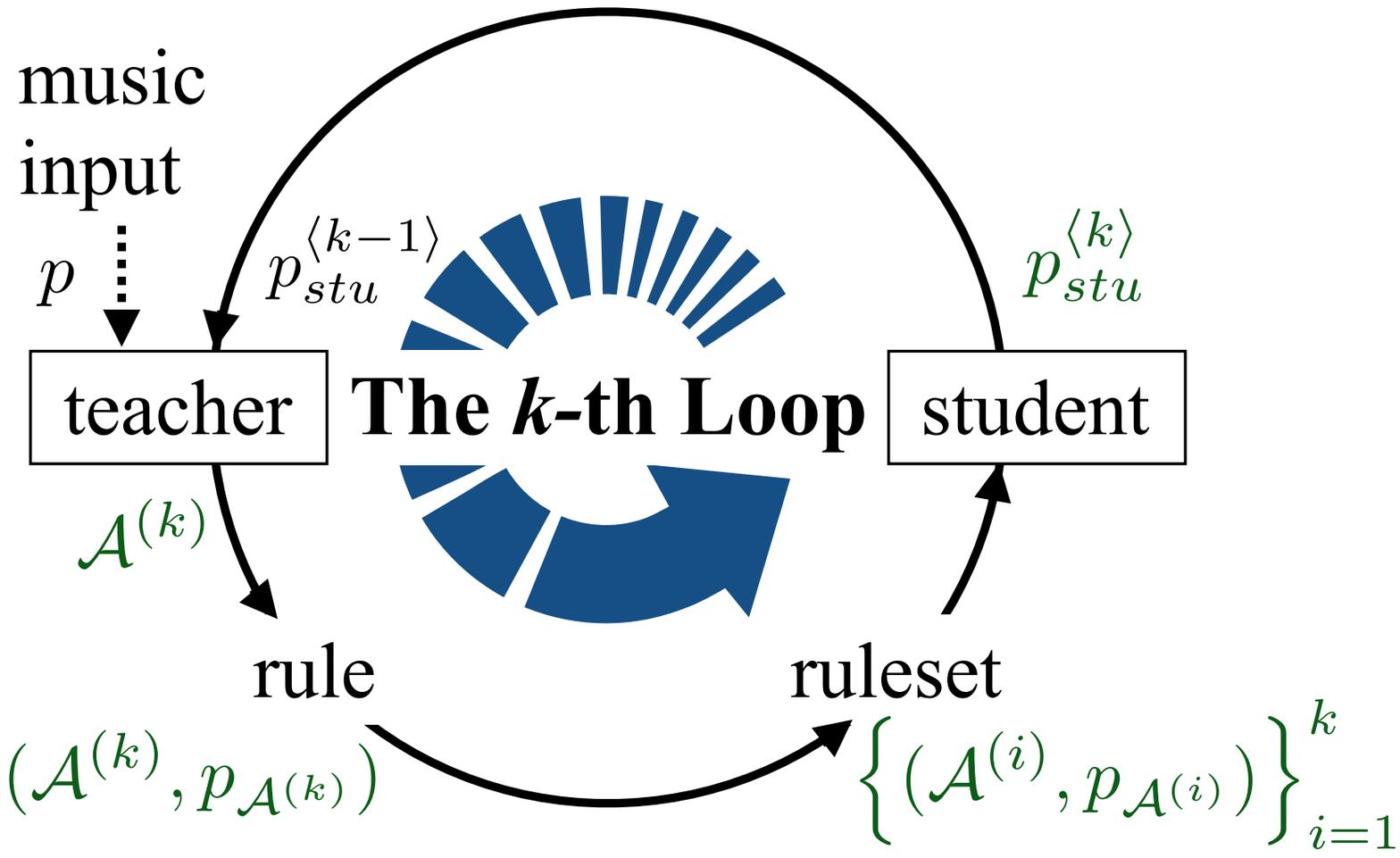
Learning Algorithm



Learning Algorithm



Learning Algorithm



Teacher: a Discriminative Model

Want to find the abstraction which exhibits the largest statistical difference between the student and the input data.

Teacher: a Discriminative Model

The teacher solves an optimization problem:

$$\begin{aligned} \underset{\mathcal{A} \in \mathfrak{P}_X}{\text{maximize}} \quad & D_{KL}\left(p_{\mathcal{A},stu}^{\langle k-1 \rangle} \,\|\, p_{\mathcal{A}}\right) \\ \text{subject to} \quad & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \ldots \end{aligned}$$

Teacher: a Discriminative Model

The teacher solves an optimization problem:

$$\begin{aligned} \underset{\mathcal{A}\in\mathfrak{P}_X}{\text{maximize}} \quad & D_{KL}\left(p^{\langle k-1\rangle}_{\mathcal{A},stu} \,\|\, p_{\mathcal{A}}\right) \\ \text{subject to} \quad & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \ldots \end{aligned}$$

What does $\mathfrak{P}_X$ come from?

Teacher: a Discriminative Model

$\mathfrak{P}_X$: abstraction universe

Teacher: a Discriminative Model

$\mathfrak{P}_X$: abstraction universe

Mathematically, a partition lattice, which is a special type of partially ordered set

Teacher: a Discriminative Model

$\mathfrak{P}_X$: abstraction universe

Mathematically, a partition lattice, which is a special type of partially ordered set

Pictorially, a directed acyclic graph (vertex: partition; edge: coarser than)

Teacher: a Discriminative Model

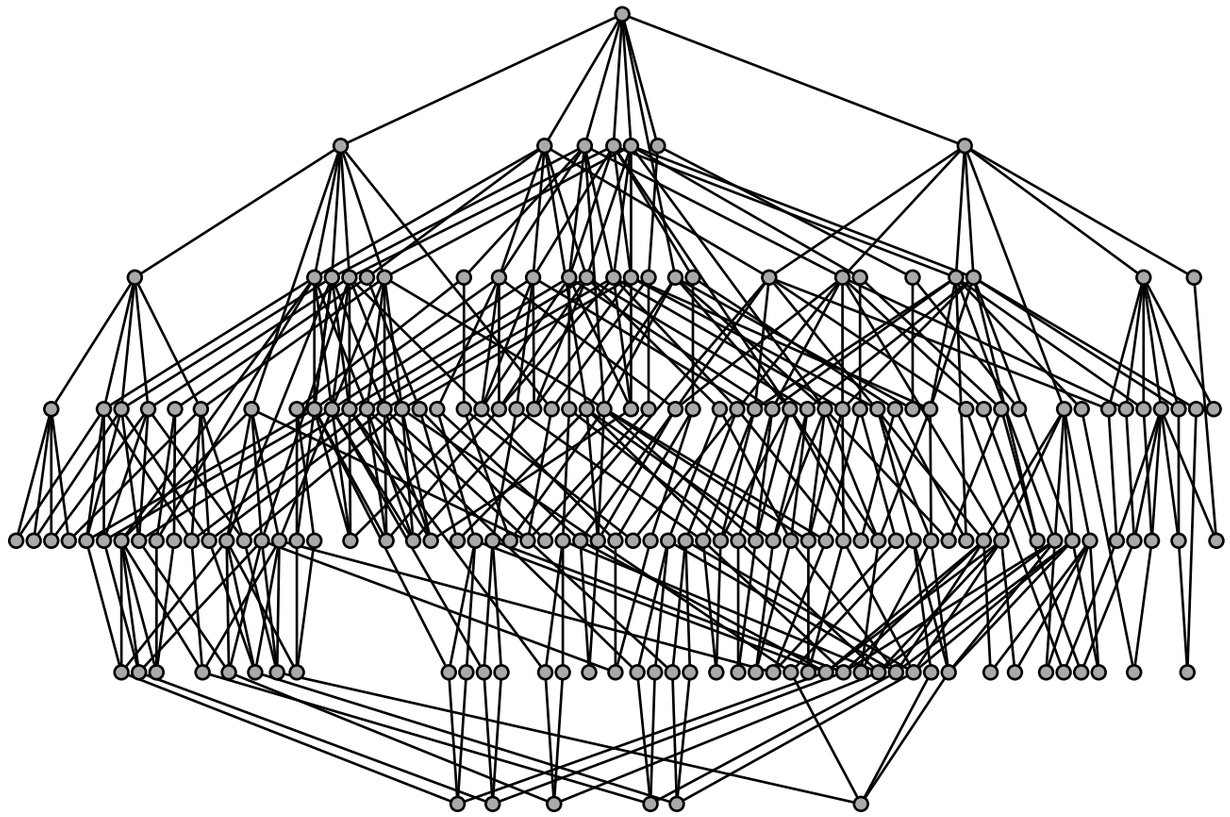
$\mathfrak{P}_X$: abstraction universe

Mathematically, a partition lattice, which is a special type of partially ordered set

Pictorially, a directed acyclic graph (vertex: partition; edge: coarser than)

(more specific)
finer

coarser
(more general)



Teacher: a Discriminative Model

How to construct $\mathcal{A} \in \mathfrak{P}_X$ (abstraction universe) ?

Teacher: a Discriminative Model

How to construct $\mathcal{A} \in \mathfrak{P}_X$ (abstraction universe) ?

- Feature-Induced Partition
- Symmetry-Generated Partition

Teacher: a Discriminative Model

How to construct $\mathcal{A} \in \mathfrak{P}_X$ (abstraction universe) ?

- Feature-Induced Partition
- Symmetry-Generated Partition

In both cases, the underlying mechanism that generates the partition is human-interpretable.

We do not consider arbitrary partitions.

Teacher: a Discriminative Model

How to construct $\mathcal{A} \in \mathfrak{P}_X$ (abstraction universe) ?

- Feature-Induced Partition
- Symmetry-Generated Partition

In both cases, the underlying mechanism that generates the partition is human-interpretable.

We do not consider arbitrary partitions.

Teacher: a Discriminative Model

How to construct $\mathcal{A} \in \mathfrak{P}_X$ (abstraction universe) ?

Feature-Induced Partition

$$\mathcal{A}_\phi = \{\phi^{-1}(\{y\}) \mid y \in \phi(X)\}$$

How to construct $\phi \in \Phi$ (feature universe) ?

$$\phi = d \circ w$$

How to construct $d \in D$ (discriptors) and $w \in W$ (windows) ?

$$d = b_k \circ \cdots \circ b_1, \quad b_i \in B$$

$$w = w_I, \quad I \subseteq \{1, 2, 3, 4\}$$

Teacher: a Discriminative Model

How to construct $\mathcal{A} \in \mathfrak{P}_X$ (abstraction universe) ?

Feature-Induced Partition

$$\mathcal{A}_\phi = \{\phi^{-1}(\{y\}) \mid y \in \phi(X)\}$$

How to construct $\phi \in \Phi$ (feature universe) ?

$$\phi = d \circ w$$

How to construct $d \in D$ (discriptors) and $w \in W$ (windows) ?

$$d = b_k \circ \cdots \circ b_1, \quad b_i \in B$$

$$w = w_I, \quad I \subseteq \{1, 2, 3, 4\}$$

For example: $d = \texttt{mod}_{12} \circ \texttt{diff}, \quad w = w_{\{1,4\}}$

Teacher: a Discriminative Model

The teacher solves an optimization problem:

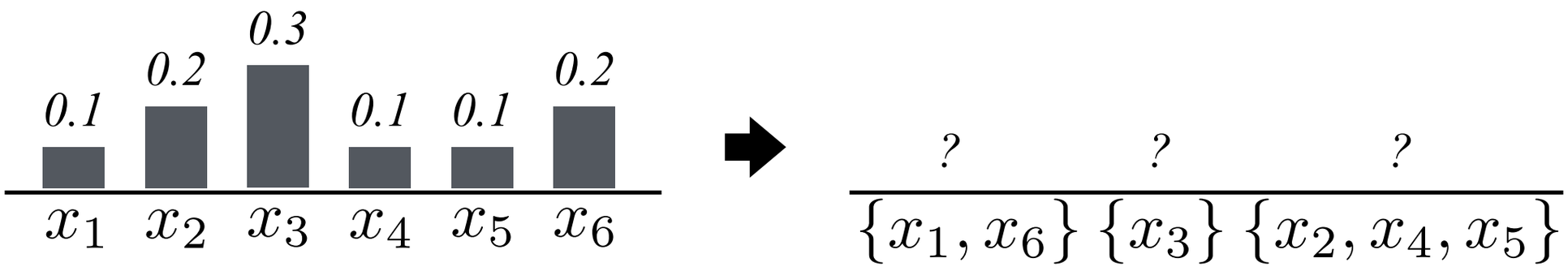
$$\begin{aligned} \underset{\mathcal{A} \in \mathfrak{P}_X}{\text{maximize}} \quad & D_{KL}\left(p_{\mathcal{A},stu}^{\langle k-1 \rangle} \,\|\, p_{\mathcal{A}}\right) \\ \text{subject to} \quad & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \ldots \end{aligned}$$

Teacher: a Discriminative Model

The teacher solves an optimization problem:

$$\begin{aligned} \underset{\mathcal{A}\in\mathfrak{P}_X}{\text{maximize}} \quad & D_{KL}\left(p^{\langle k-1\rangle}_{\mathcal{A},stu} \,\|\, p_{\mathcal{A}}\right) \\ \text{subject to} \quad & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \ldots \end{aligned}$$

Note that: given p, computing $p_{\mathcal{A}}$ is easy:

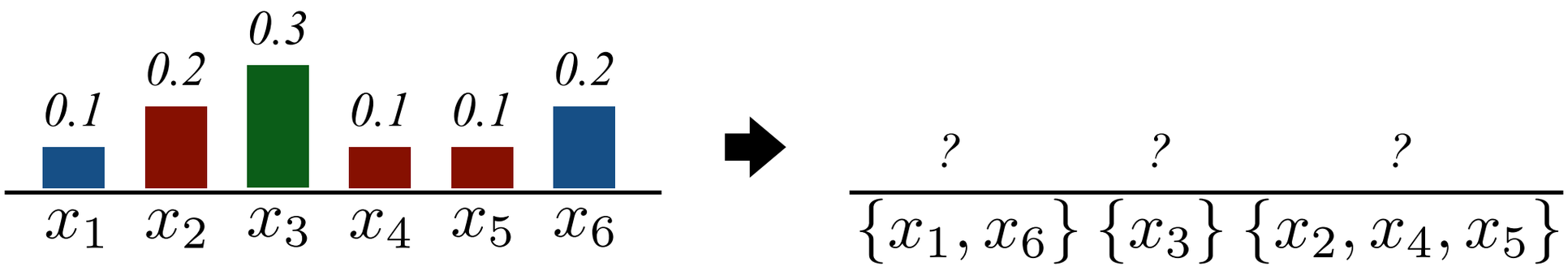


Teacher: a Discriminative Model

The teacher solves an optimization problem:

$$\begin{aligned} \underset{\mathcal{A}\in\mathfrak{P}_X}{\text{maximize}} \quad & D_{KL}\left(p^{\langle k-1\rangle}_{\mathcal{A},stu} \;\|\; p_{\mathcal{A}}\right) \\ \text{subject to} \quad & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \ldots \end{aligned}$$

Note that: given p, computing $p_{\mathcal{A}}$ is easy:

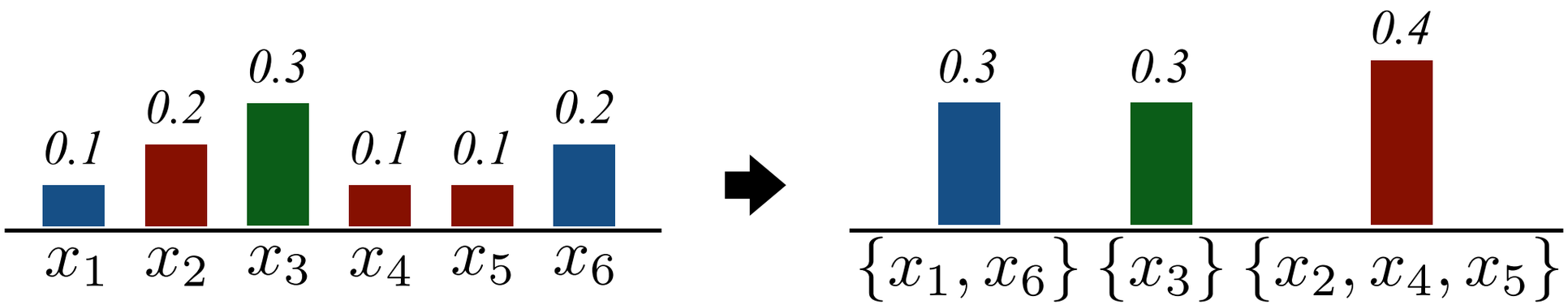


Teacher: a Discriminative Model

The teacher solves an optimization problem:

$$\begin{aligned} \underset{\mathcal{A}\in\mathfrak{P}_X}{\text{maximize}} \quad & D_{KL}\left(p^{\langle k-1\rangle}_{\mathcal{A},stu} \,\|\, p_{\mathcal{A}}\right) \\ \text{subject to} \quad & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \ldots \end{aligned}$$

Note that: given p, computing $p_{\mathcal{A}}$ is easy:



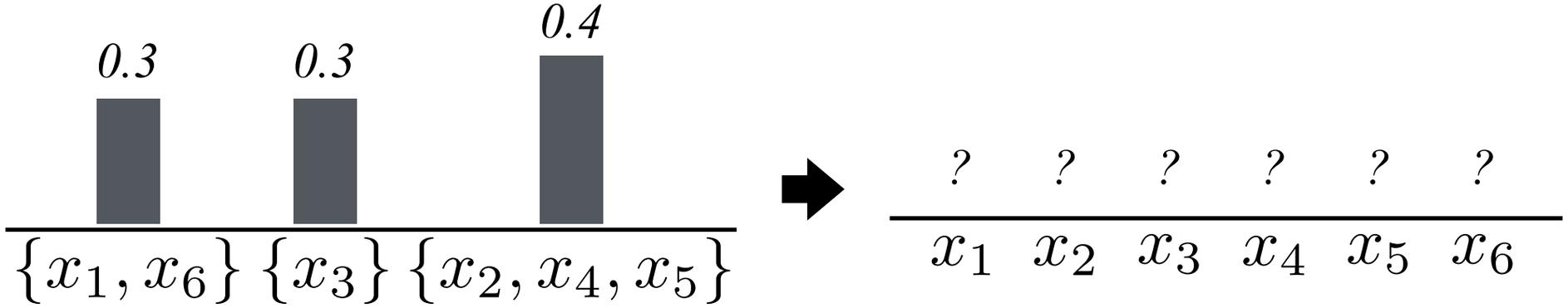
Student: a Generative Model

Apply probabilistic rules, which is known as the rule realization problem.

Student: a Generative Model

Apply probabilistic rules, which is known as the rule realization problem.

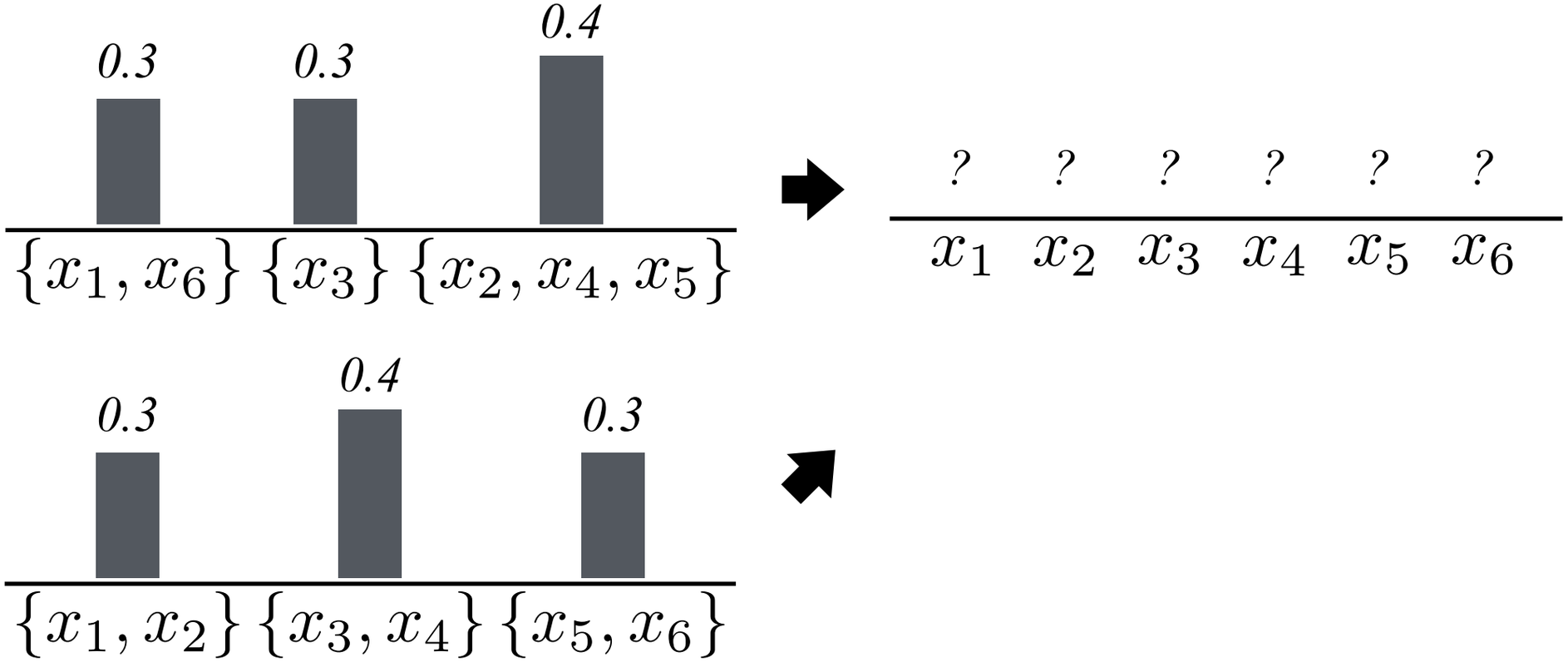
Given $p_{\mathcal{A}}$, compute p:



Student: a Generative Model

Apply probabilistic rules, which is known as the rule realization problem.

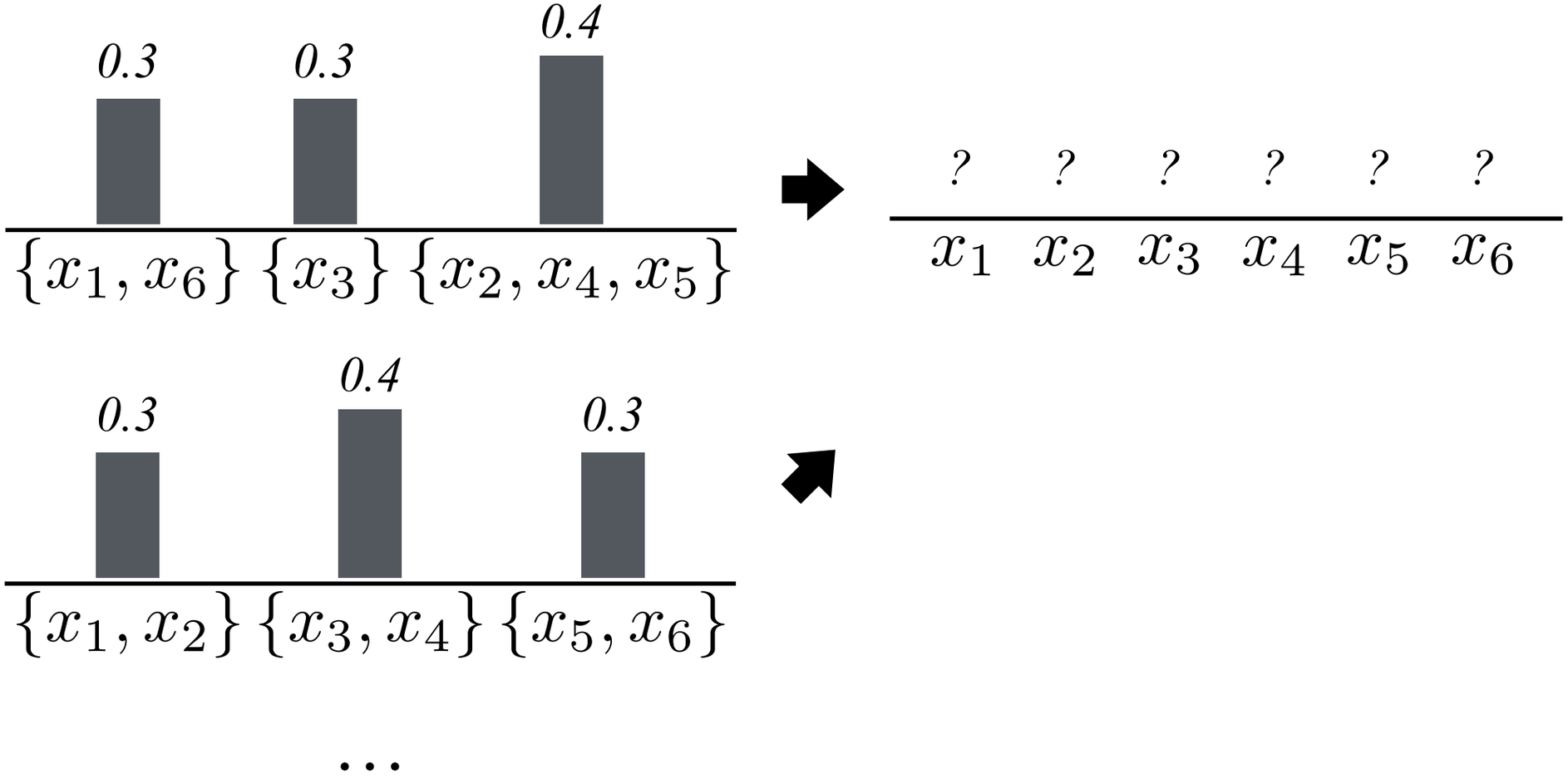
Given $p_{\mathcal{A}}$, compute p:



Student: a Generative Model

Apply probabilistic rules, which is known as the rule realization problem.

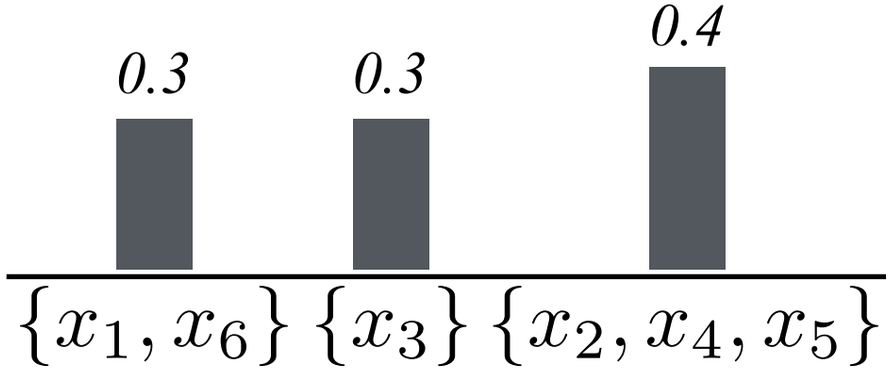
Given $p_{\mathcal{A}}$, compute p:

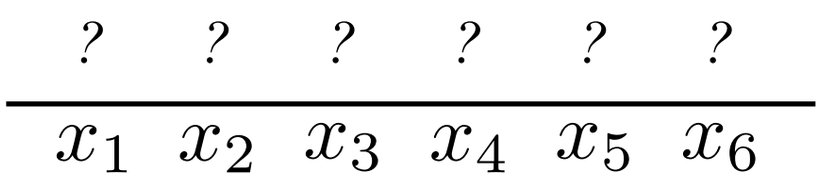


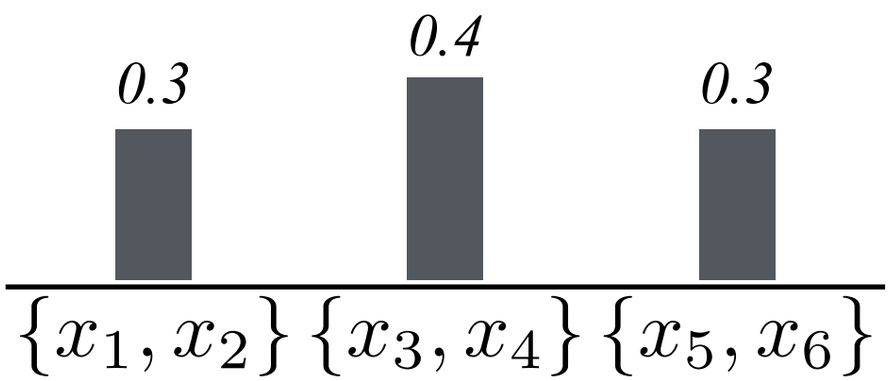
Student: a Generative Model

Apply probabilistic rules, which is known as the rule realization problem.

Given $p_{\mathcal{A}}$, compute p:







...

not necessarily unique

which one do we prefer?

Student: a Generative Model

Want to find the probabilistic model which enables novelty while at the same time satisfies all the rules.

Student: a Generative Model

The student solves another optimization problem:

$$\begin{aligned} \underset{p_{stu}^{\langle k \rangle} \in \Delta_{|X|}}{\text{maximize}} \quad & S_q(p_{stu}^{\langle k \rangle}) := (q-1)^{-1} \left(1 - \|p_{stu}^{\langle k \rangle}\|_q^q\right) \\ \text{subject to} \quad & A^{(i)} p_{stu}^{\langle k \rangle} = p_{\mathcal{A}^{(i)}}, \quad i = 1, \ldots, k \end{aligned}$$

Student: a Generative Model

The student solves another optimization problem:

Tsallis entropy: measures randomness

$$\underset{p_{stu}^{\langle k \rangle} \in \Delta_{|X|}}{\text{maximize}} \quad \overbrace{S_q}(p_{stu}^{\langle k \rangle}) := (q-1)^{-1}\left(1 - \|p_{stu}^{\langle k \rangle}\|_q^q\right)$$

$$\text{subject to} \quad A^{(i)} p_{stu}^{\langle k \rangle} = p_{\mathcal{A}^{(i)}}, \quad i = 1, \ldots, k$$

Student: a Generative Model

The student solves another optimization problem:

Tsallis entropy: measures randomness

$$\underset{p_{stu}^{\langle k \rangle} \in \Delta_{|X|}}{\text{maximize}} \quad S_q(p_{stu}^{\langle k \rangle}) := (q-1)^{-1}\left(1 - \|p_{stu}^{\langle k \rangle}\|_q^q\right)$$

$$\text{subject to} \quad A^{(i)} p_{stu}^{\langle k \rangle} = p_{\mathcal{A}^{(i)}}, \quad i = 1, \ldots, k$$

partition matrix: represents abstraction

Student: a Generative Model

The student solves another optimization problem:

Tsallis entropy: measures randomness

$$\underset{p_{stu}^{\langle k \rangle} \in \Delta_{|X|}}{\text{maximize}} \quad S_q(p_{stu}^{\langle k \rangle}) := (q-1)^{-1}\left(1 - \|p_{stu}^{\langle k \rangle}\|_q^q\right)$$

$$\text{subject to} \quad A^{(i)} p_{stu}^{\langle k \rangle} = p_{\mathcal{A}^{(i)}}, \quad i = 1, \ldots, k$$

partition matrix: represents abstraction

linear equality constraint

Student: a Generative Model

The student solves another optimization problem:

Tsallis entropy: measures randomness

$$\underset{p_{stu}^{\langle k \rangle} \in \Delta_{|X|}}{\text{maximize}} \quad S_q(p_{stu}^{\langle k \rangle}) := (q-1)^{-1}\left(1 - \|p_{stu}^{\langle k \rangle}\|_q^q\right)$$

$$\text{subject to} \quad A^{(i)} p_{stu}^{\langle k \rangle} = p_{\mathcal{A}^{(i)}}, \quad i = 1, \ldots, k$$

partition matrix: represents abstraction

linear equality constraint

$q = 2$: gini impurity function

Student: a Generative Model

The student solves another optimization problem:

Tsallis entropy: measures randomness

$$\underset{p_{stu}^{\langle k \rangle} \in \Delta_{|X|}}{\text{maximize}} \quad S_q(p_{stu}^{\langle k \rangle}) := (q-1)^{-1}\left(1 - \|p_{stu}^{\langle k \rangle}\|_q^q\right)$$

$$\text{subject to} \quad A^{(i)} p_{stu}^{\langle k \rangle} = p_{\mathcal{A}^{(i)}}, \quad i = 1, \ldots, k$$

partition matrix: represents abstraction

linear equality constraint

$q = 2$: gini impurity function

Linear Least-Squares Problem!

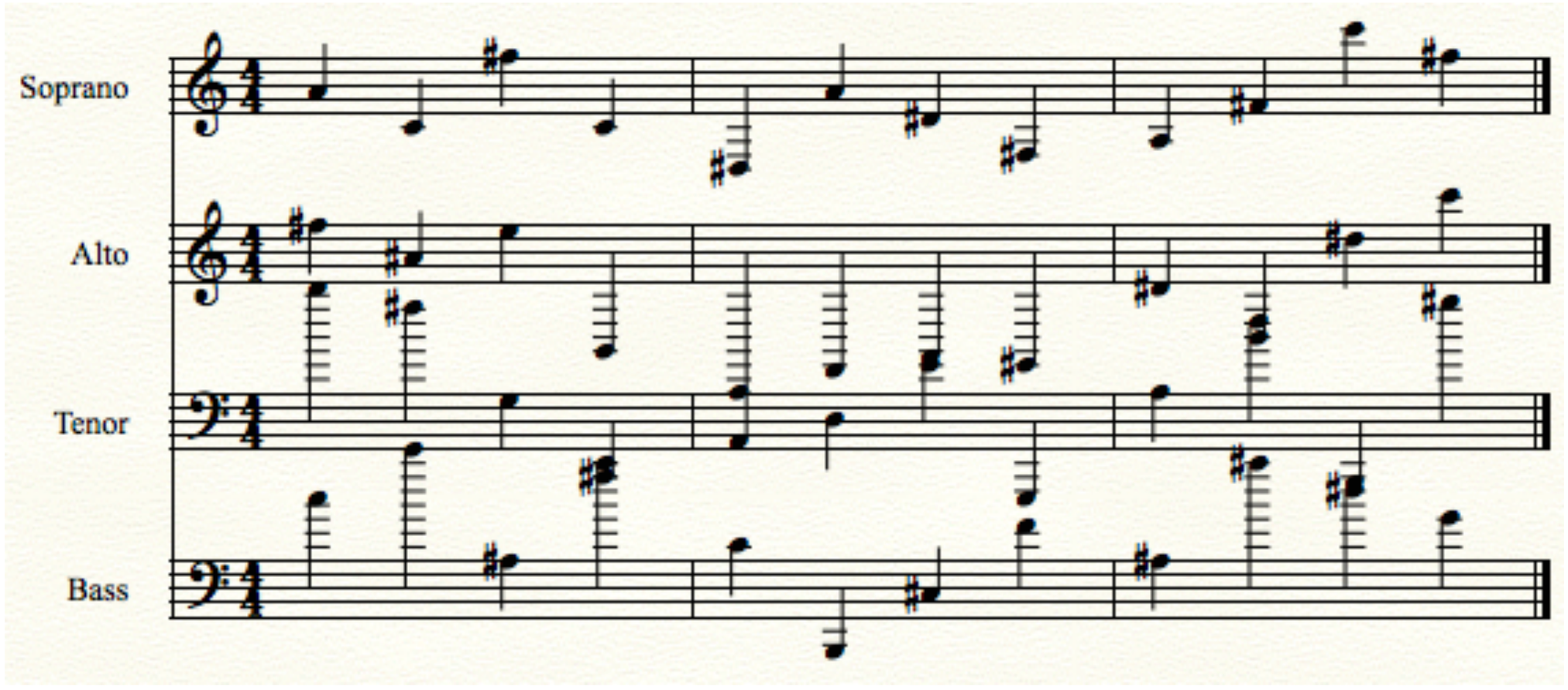
How MUS-ROVER Self-Evolves?

Context-Free Rules (1-gram)

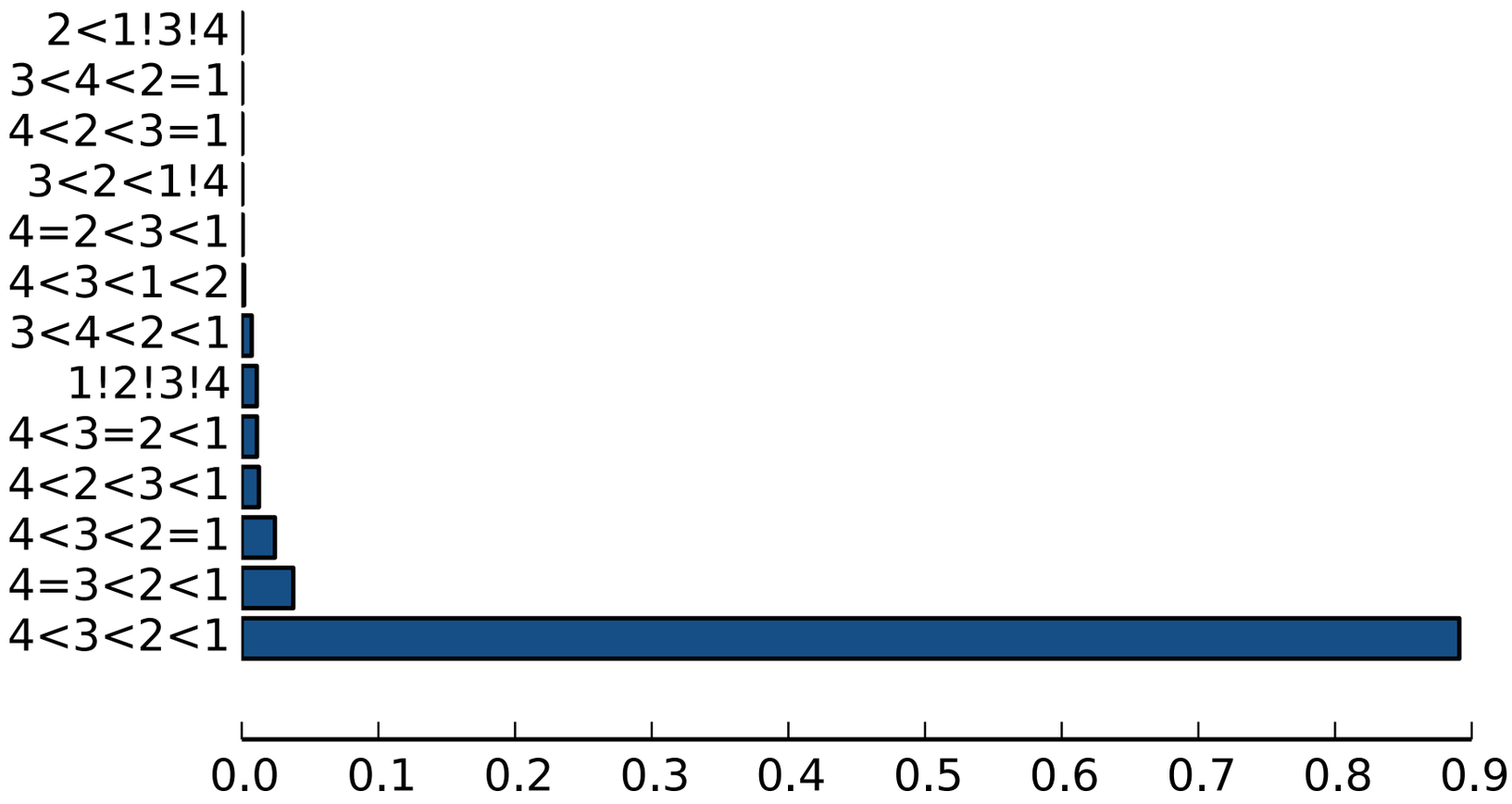
Rules and Outcomes

Fundamentals (1-gram)

Student 0

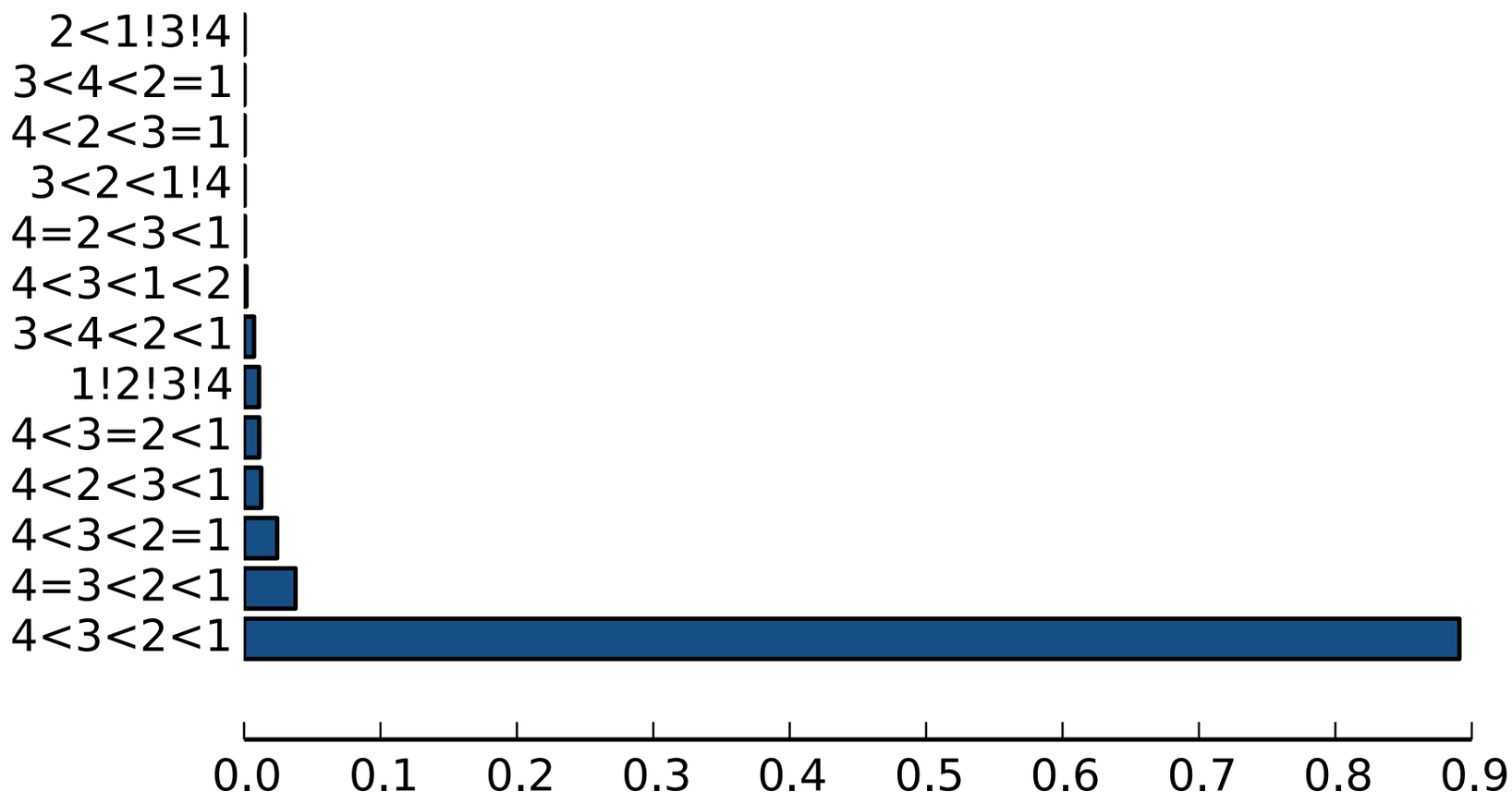


Rule 1: `order` $\circ\, w_{1,2,3,4}$



Fundamentals (1-gram)

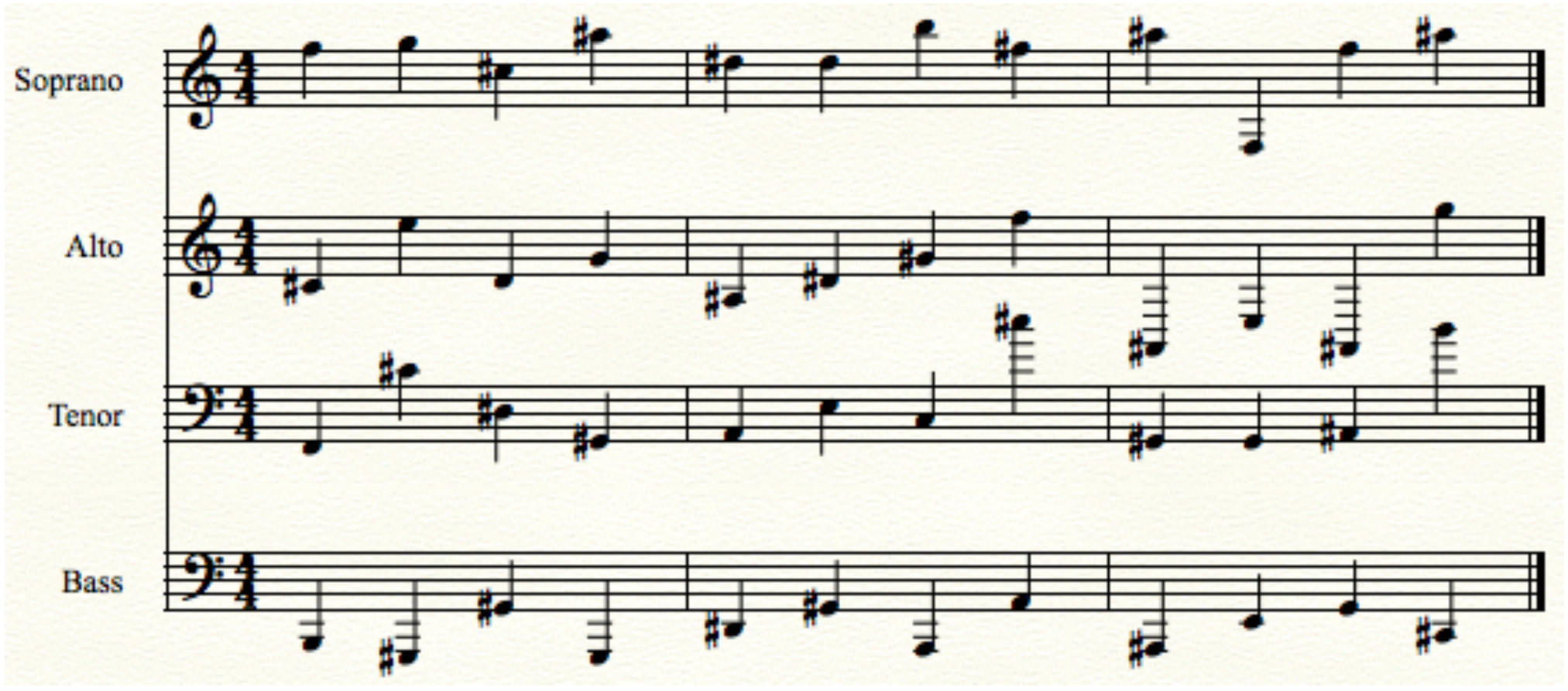
Rule 1: `order` $\circ w_{1,2,3,4}$



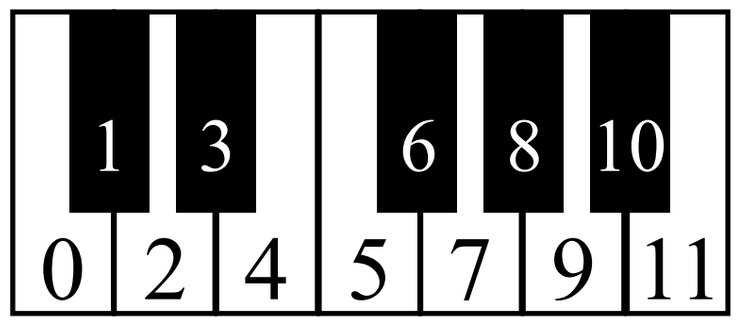
(Spacing) Almost always, the soprano pitch is above the alto, alto above tenor, and tenor above bass.

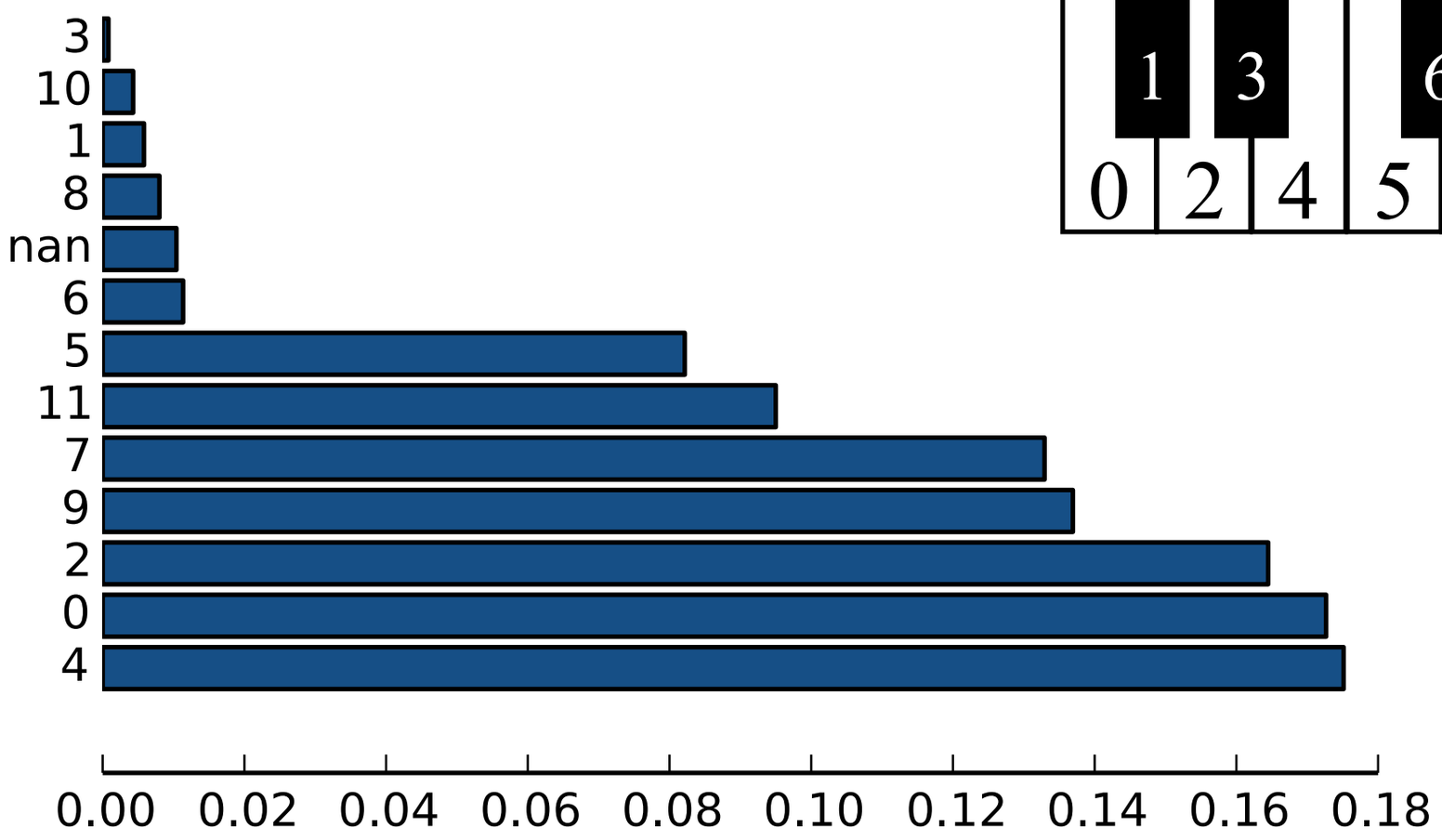
Fundamentals (1-gram)

Student 1

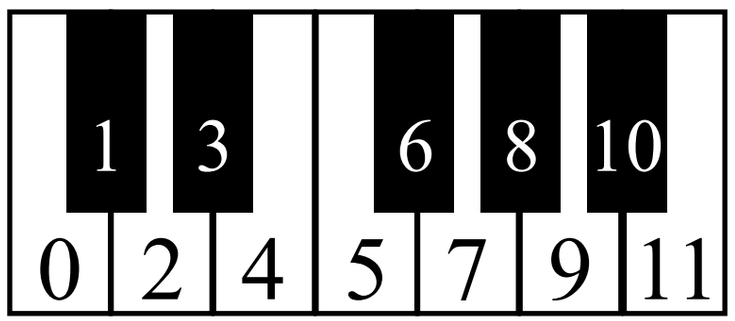


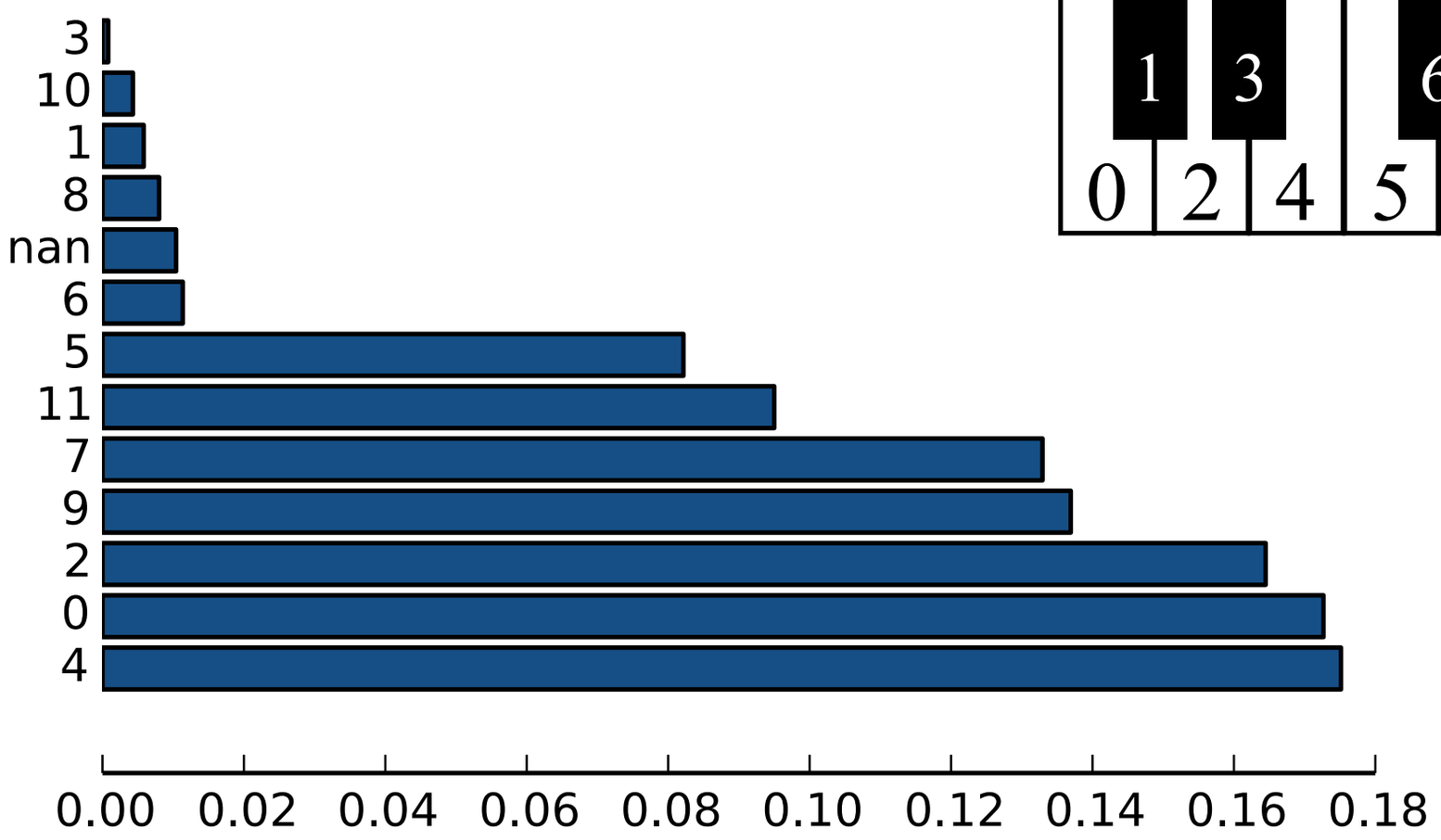
Rule 2: $\texttt{mod}_{12} \circ w_1$





Rule 2: $\texttt{mod}_{12} \circ w_1$

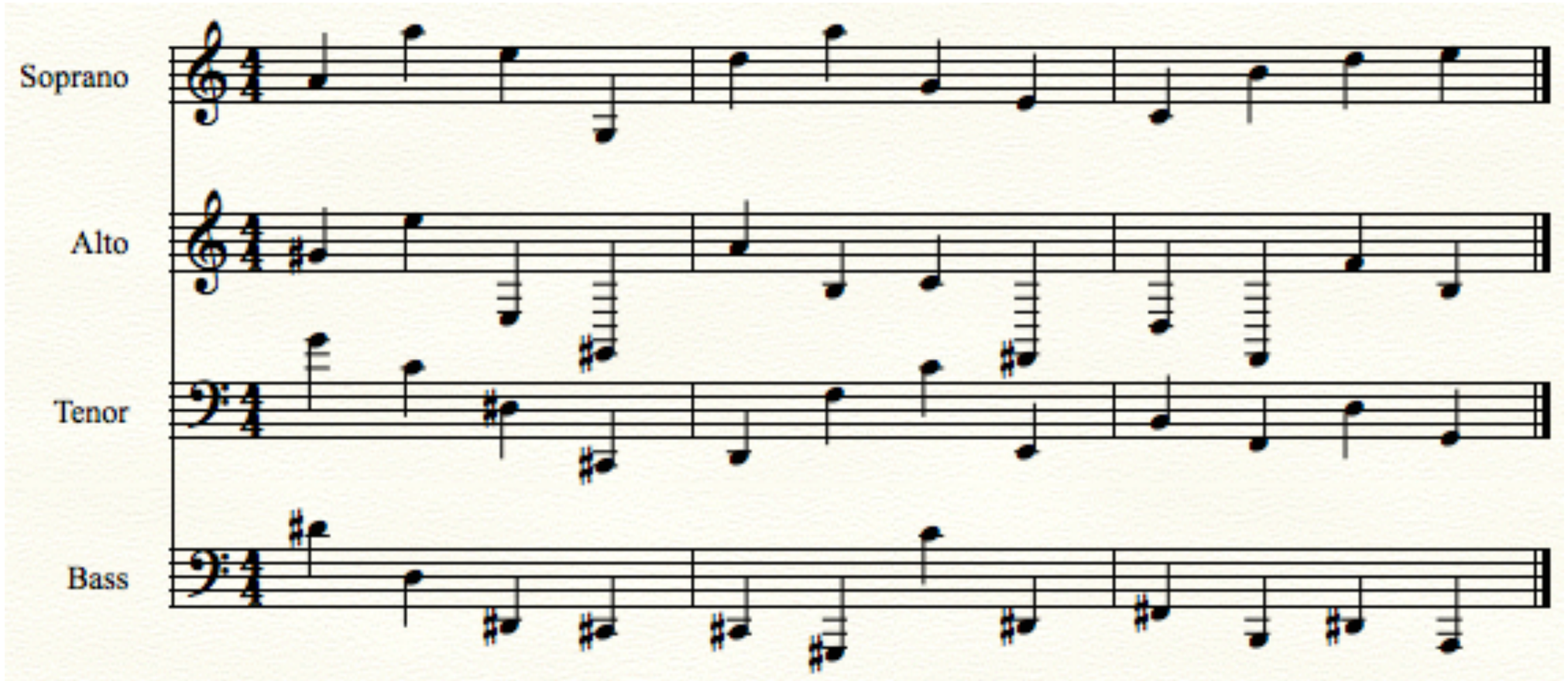




(Scale) The soprano voice is drawn from a diatonic scale with high probability.

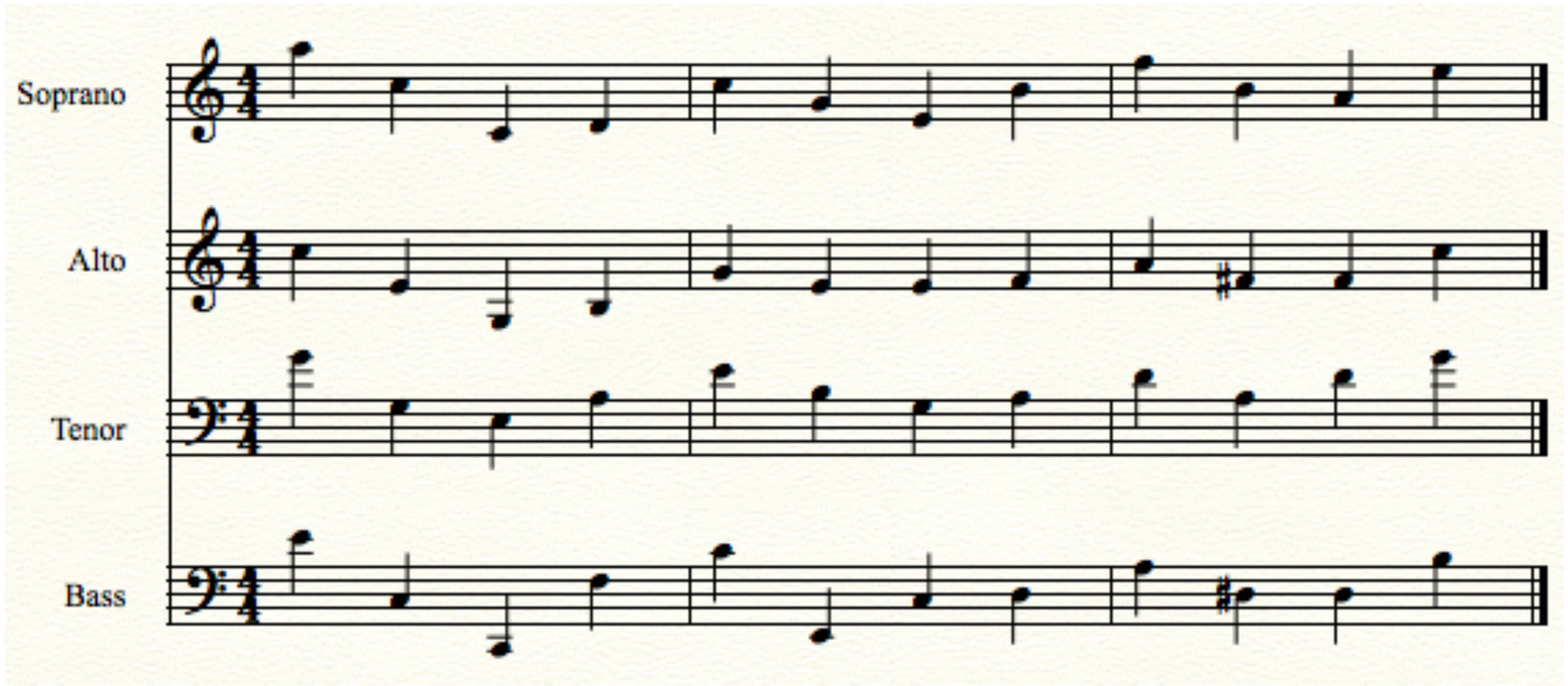
Fundamentals (1-gram)

Student 2



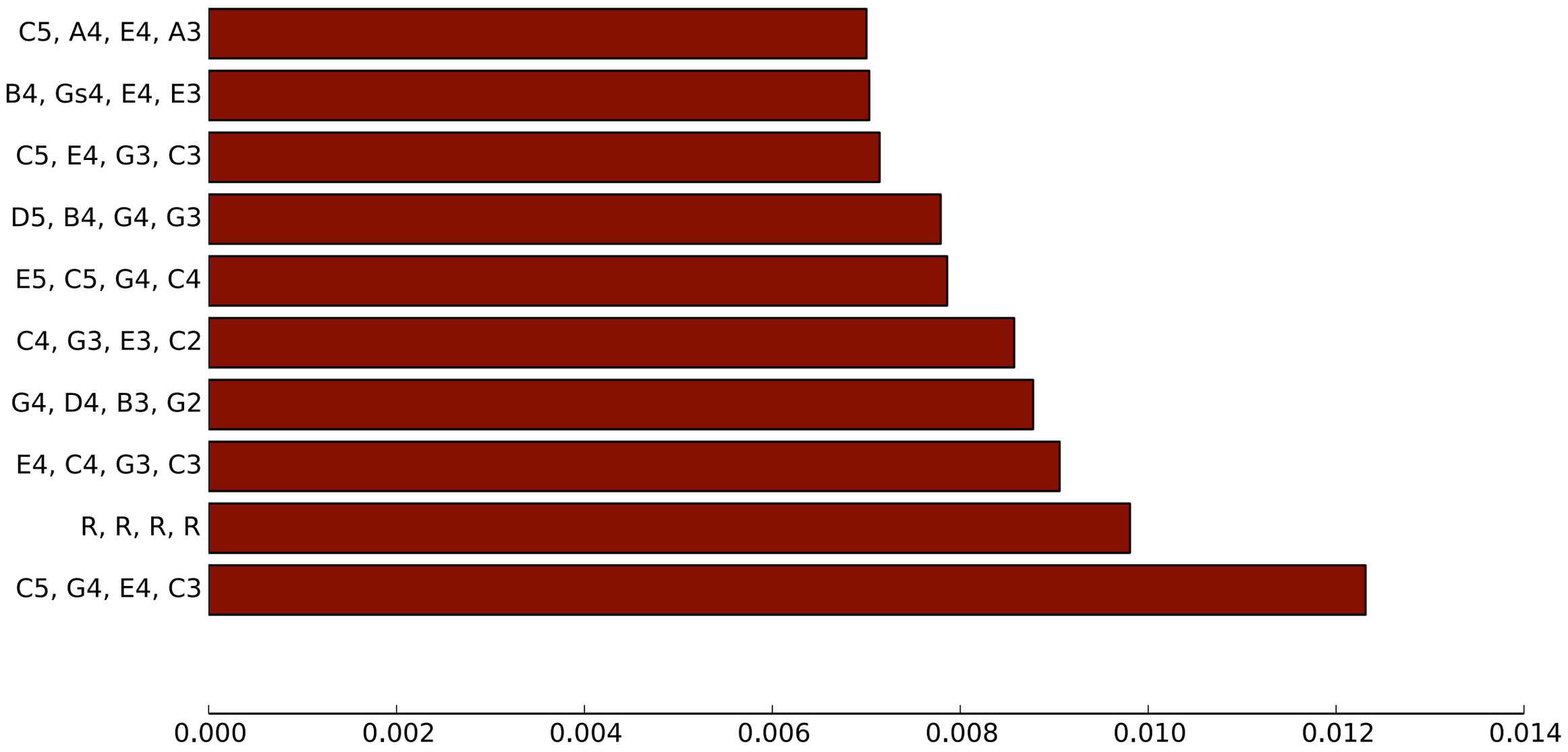
Fundamentals (1-gram)

Student 22



Fundamentals (1-gram)

Student 22



Context-Specific Rules (n-gram)

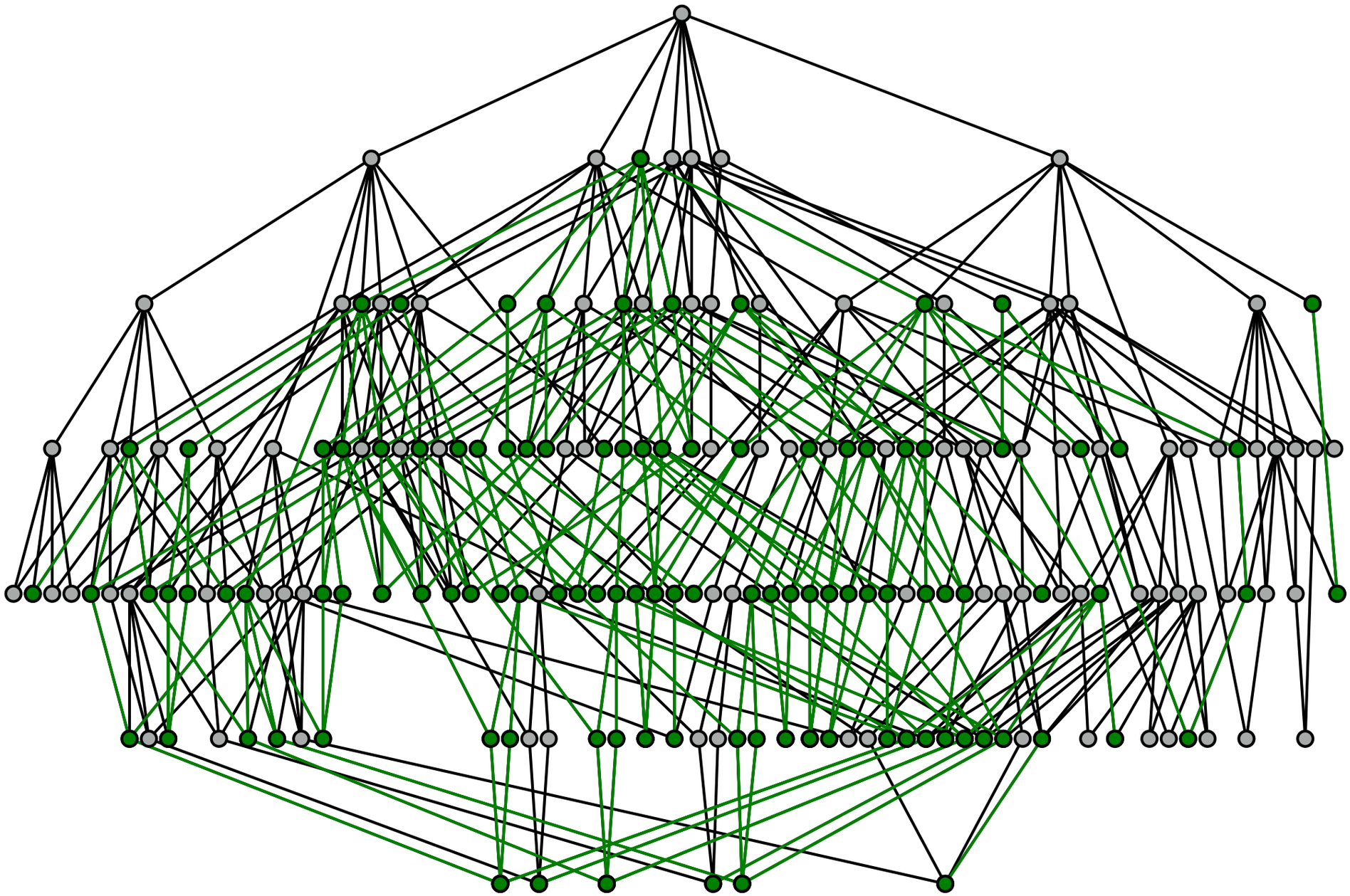
Bach's Music Brain

Part Writing (n-gram): 14th Chord

unlearned

1-gram

Loop 0



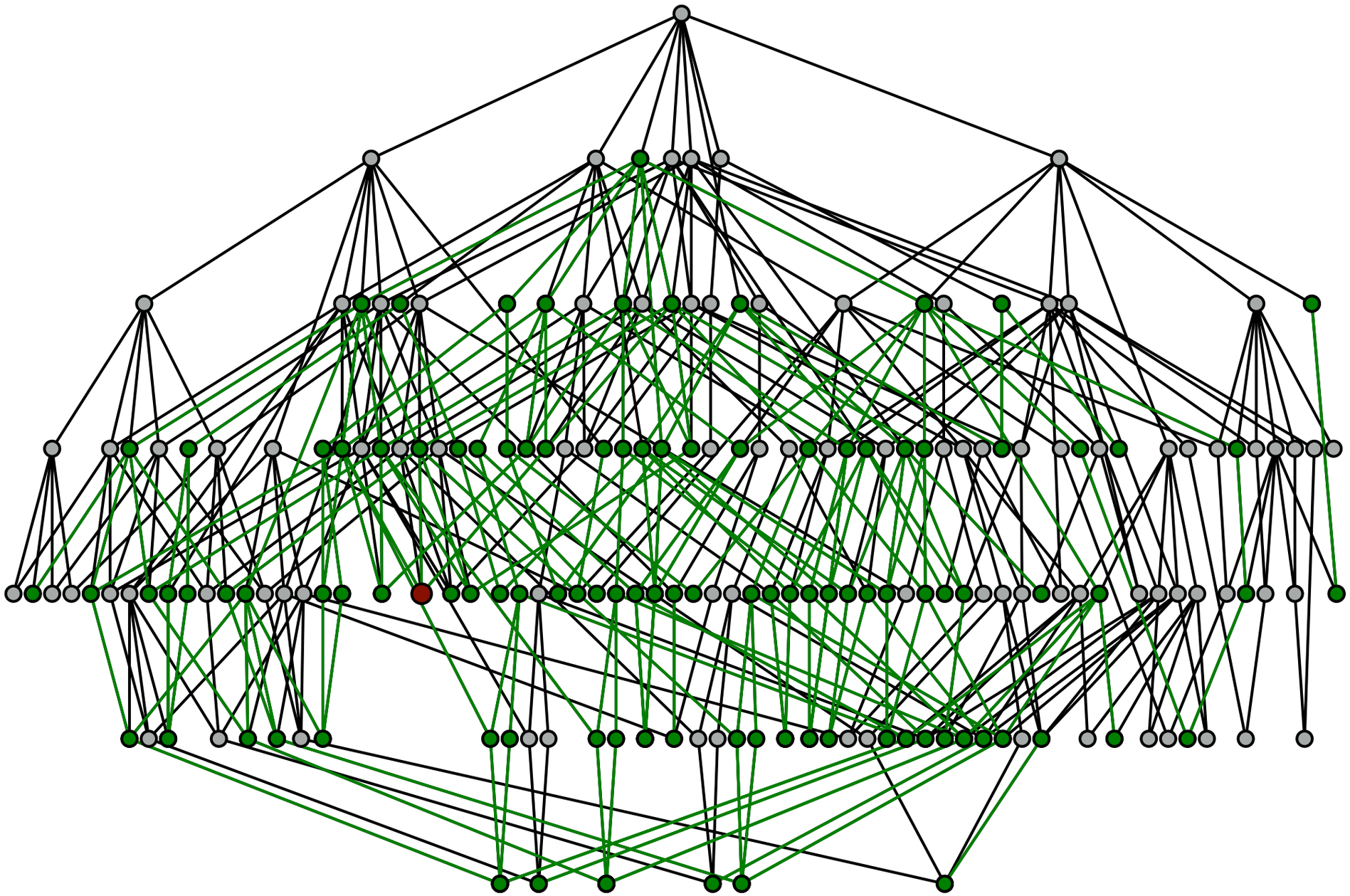
Part Writing (n-gram): 14th Chord

unlearned

1-gram

3-gram

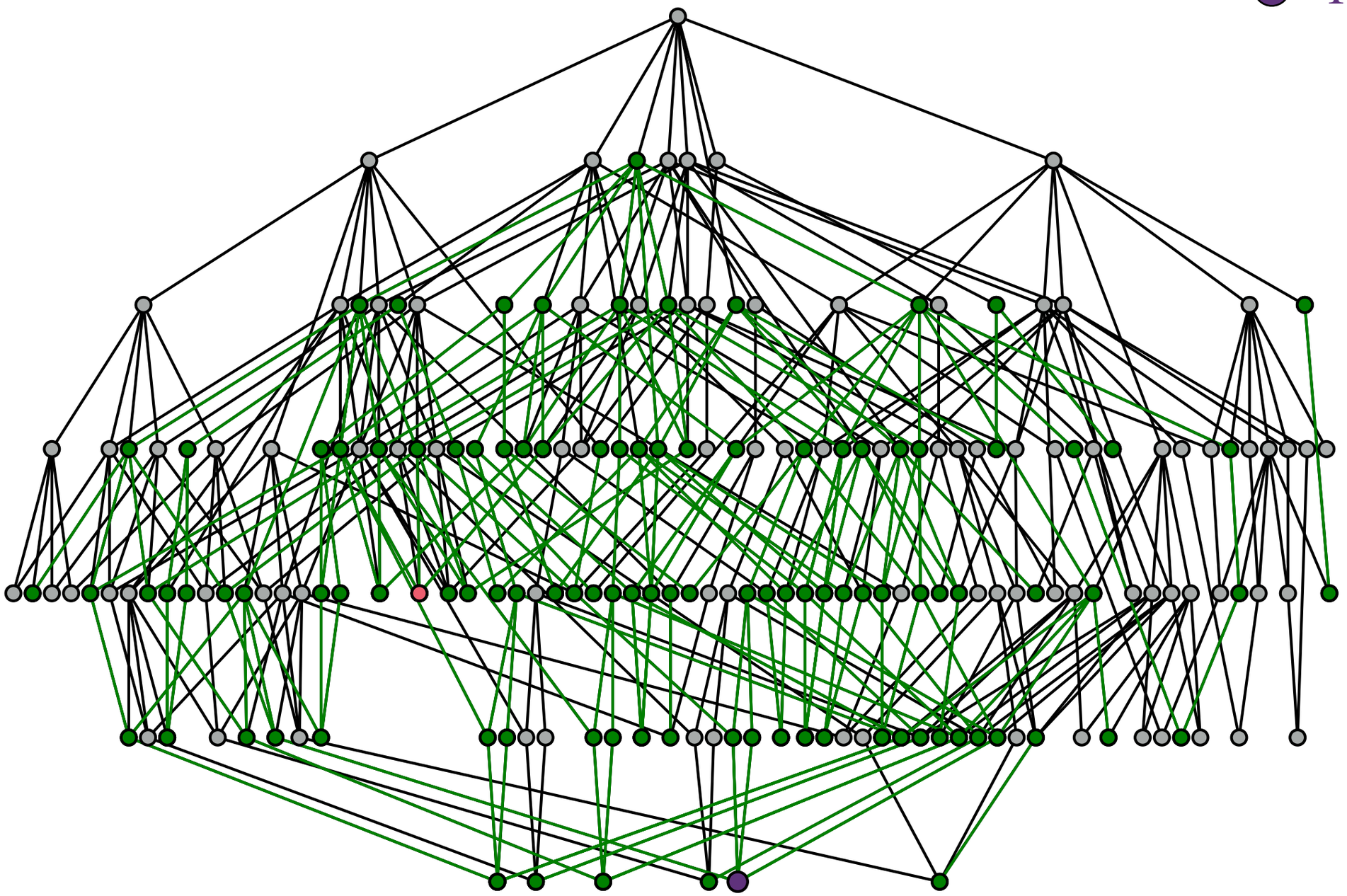
Loop 1



Part Writing (n-gram): 14th Chord

unlearned
1-gram
3-gram
10-gram

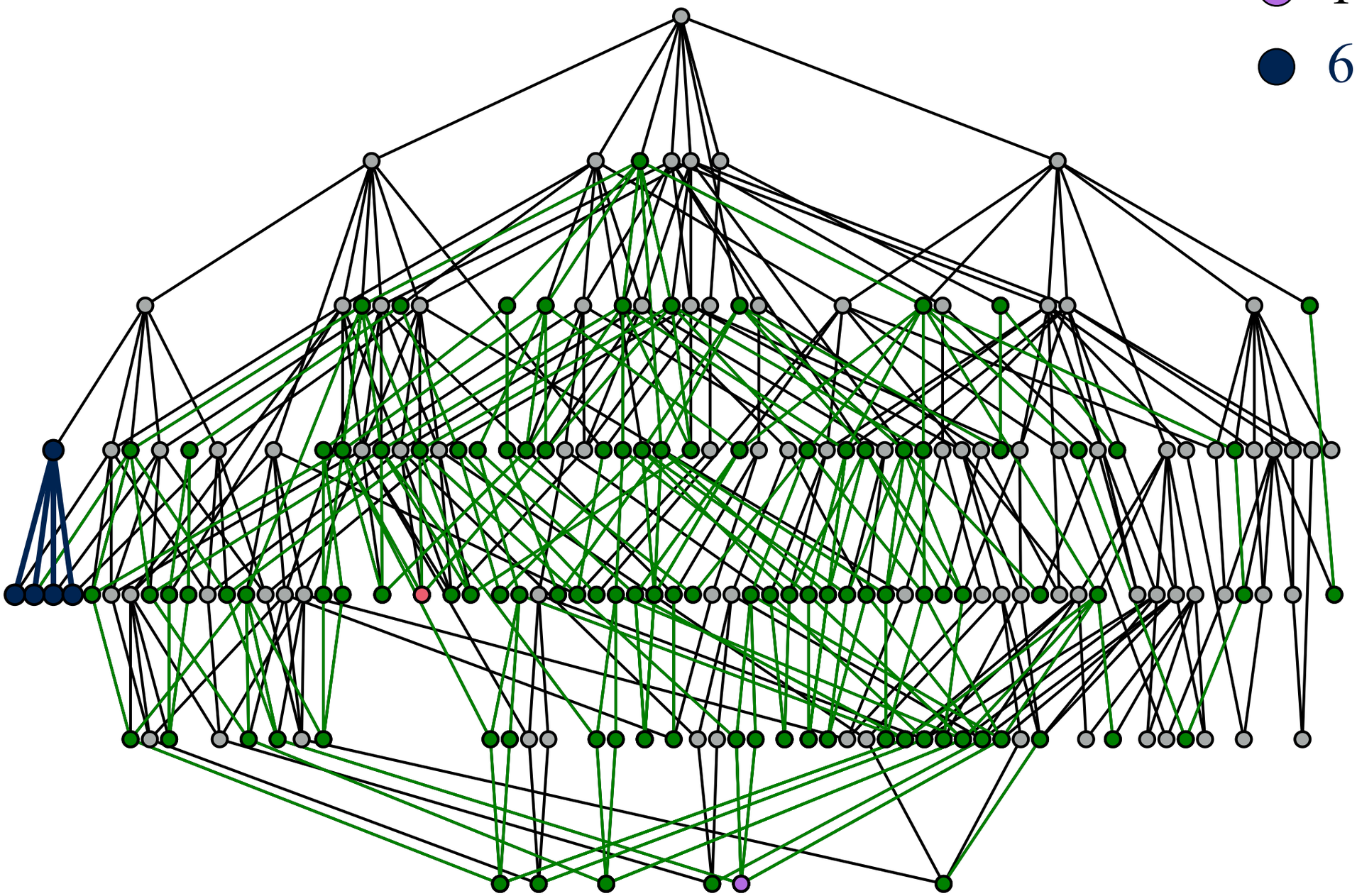
Loop 2



Part Writing (n-gram): 14th Chord

unlearned
1-gram
3-gram
10-gram
6-gram

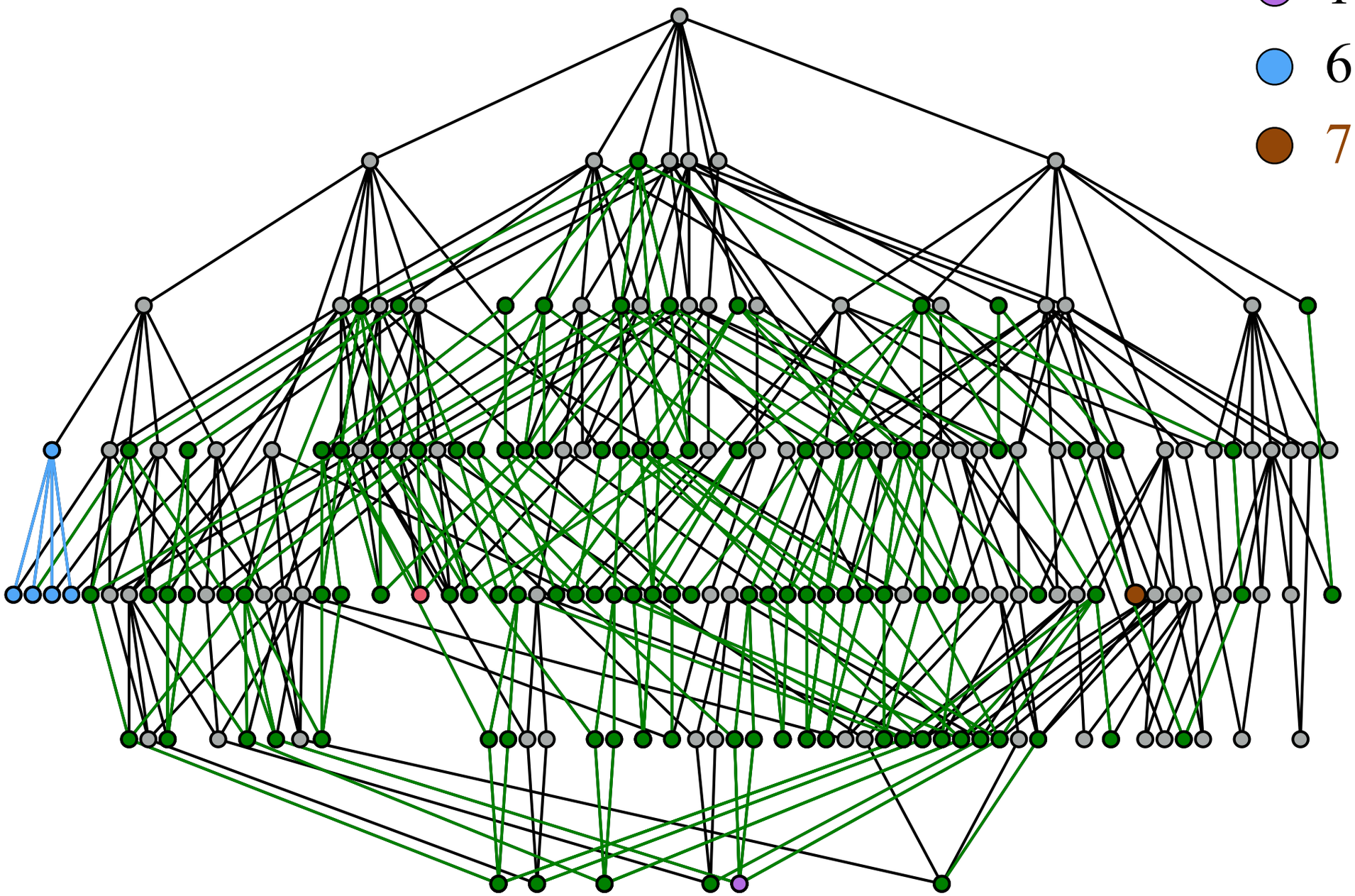
Loop 3



Part Writing (n-gram): 14th Chord

unlearned
1-gram
3-gram
10-gram
6-gram
7-gram

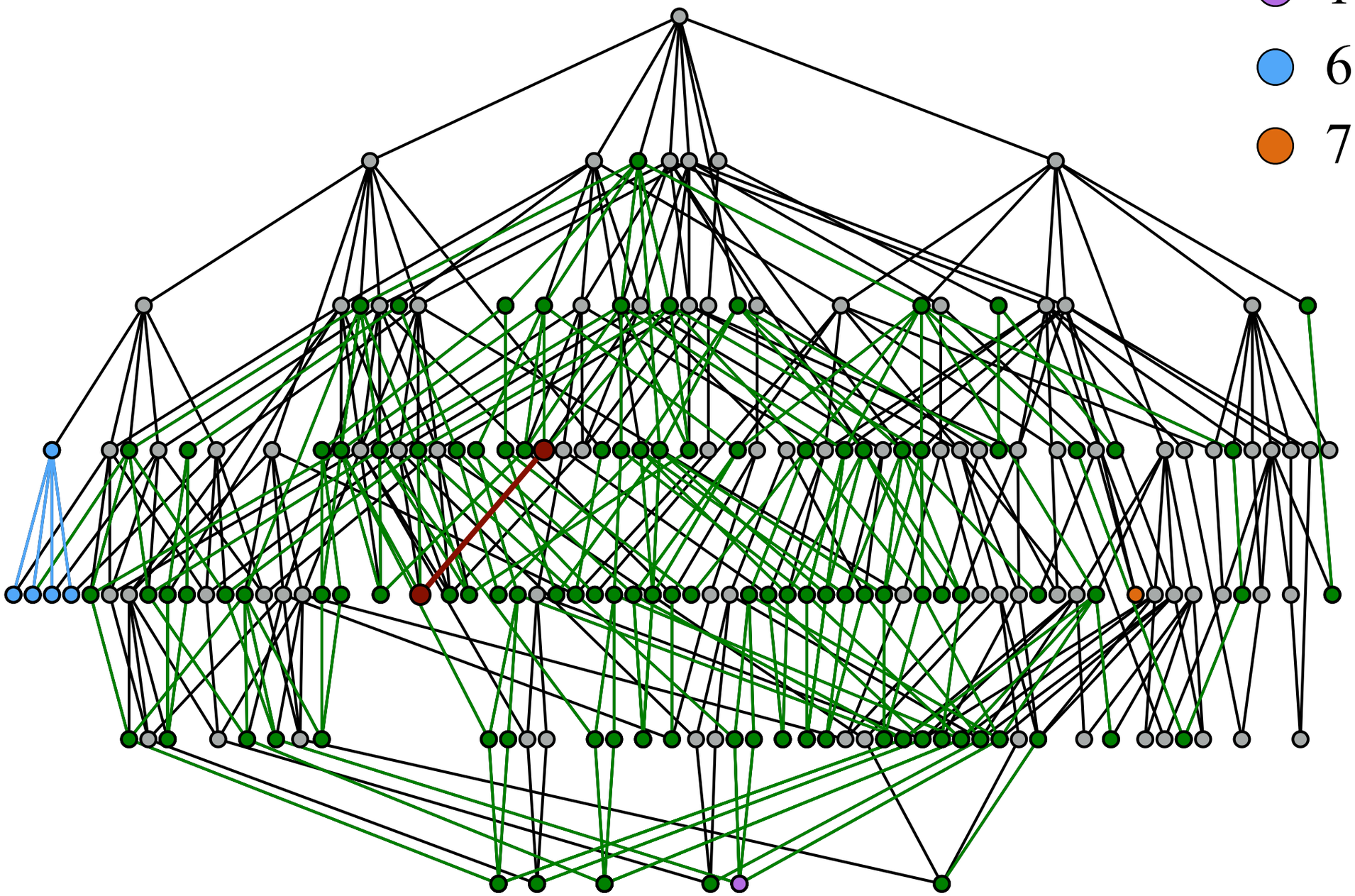
Loop 4



Part Writing (n-gram): 14th Chord

unlearned
1-gram
3-gram
10-gram
6-gram
7-gram

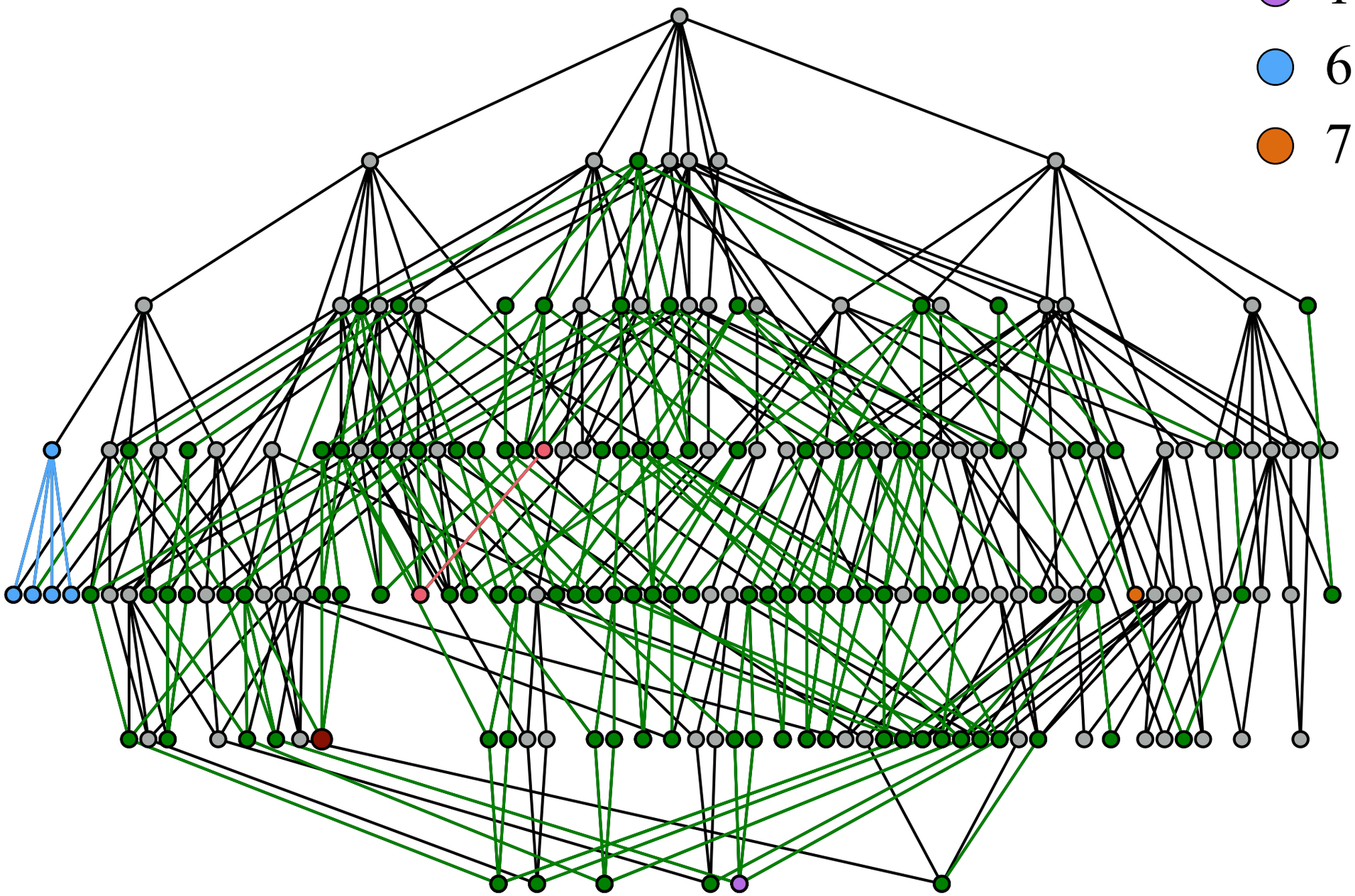
Loop 5



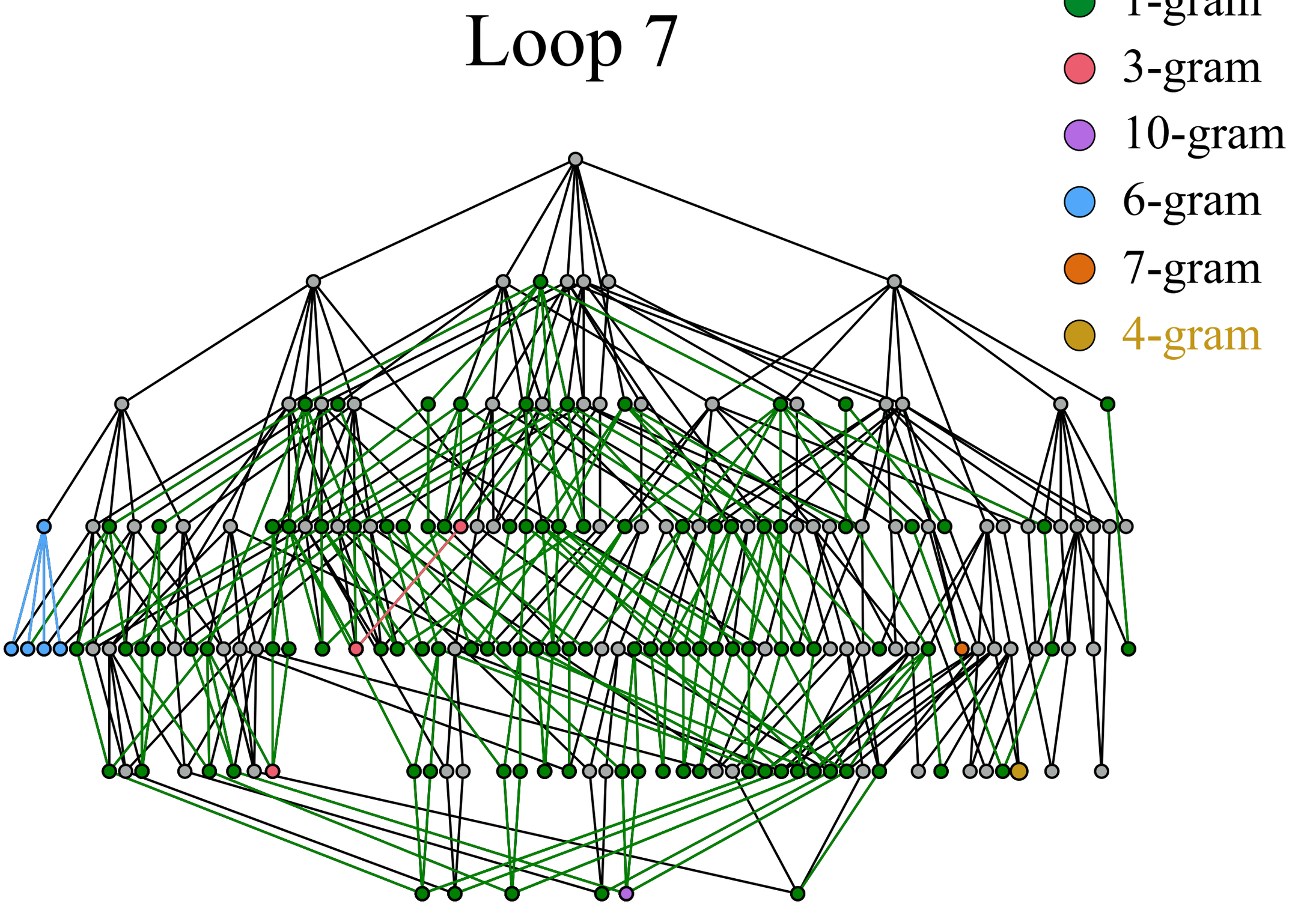
Part Writing (n-gram): 14th Chord

unlearned
1-gram
3-gram
10-gram
6-gram
7-gram

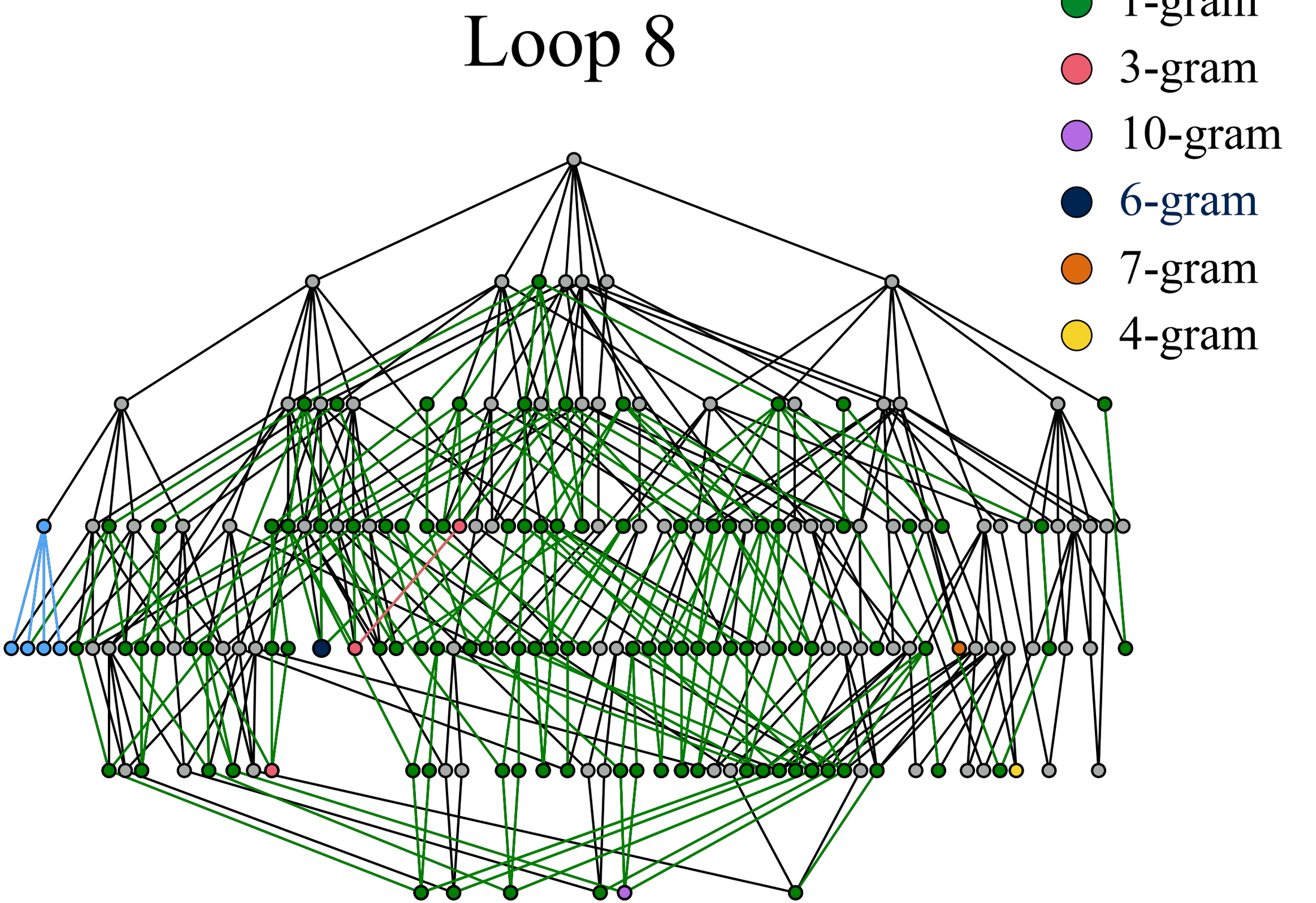
Loop 6



Part Writing (n-gram): 14th Chord

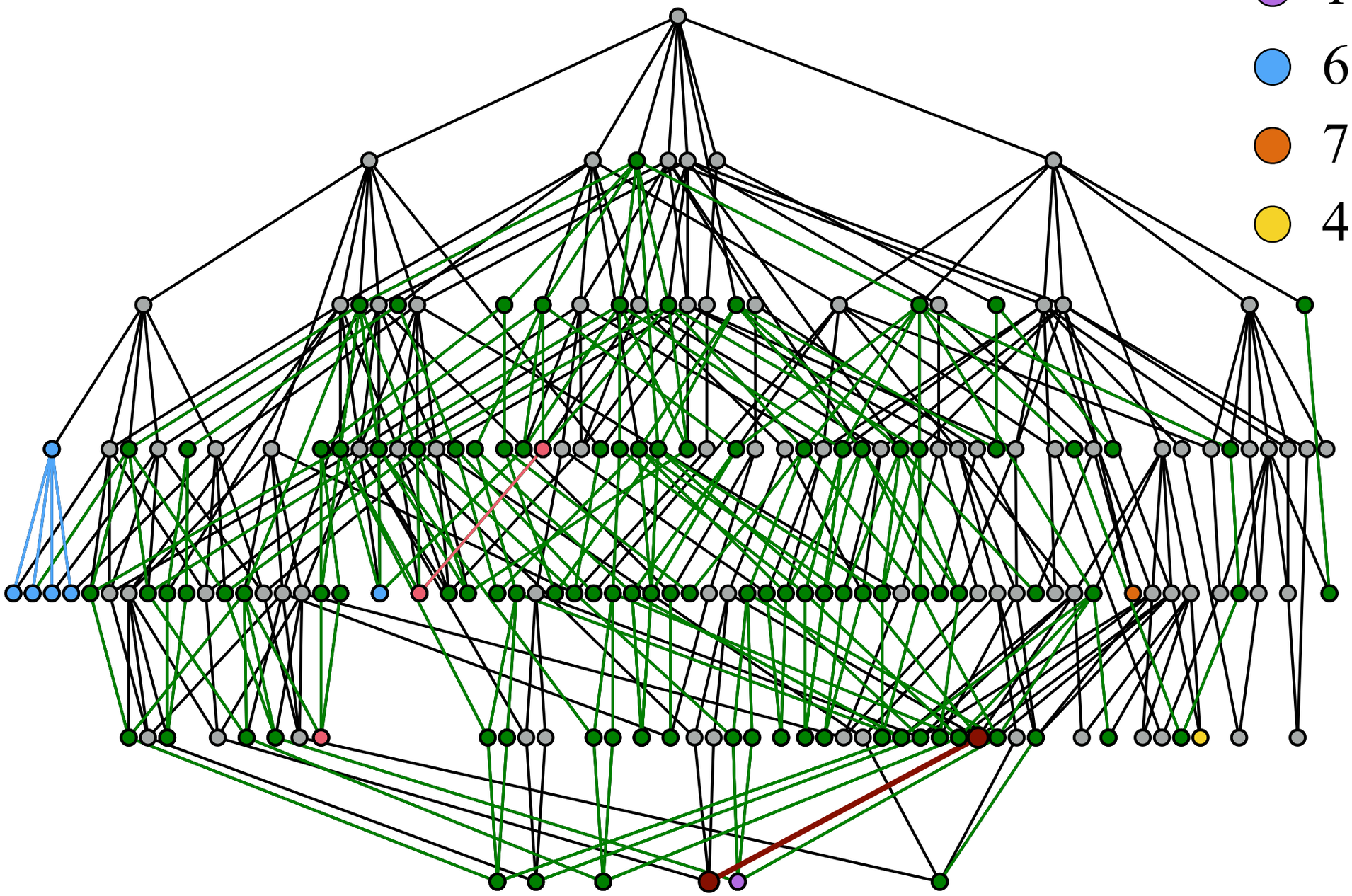


Part Writing (n-gram): 14th Chord



Part Writing (n-gram): 14th Chord

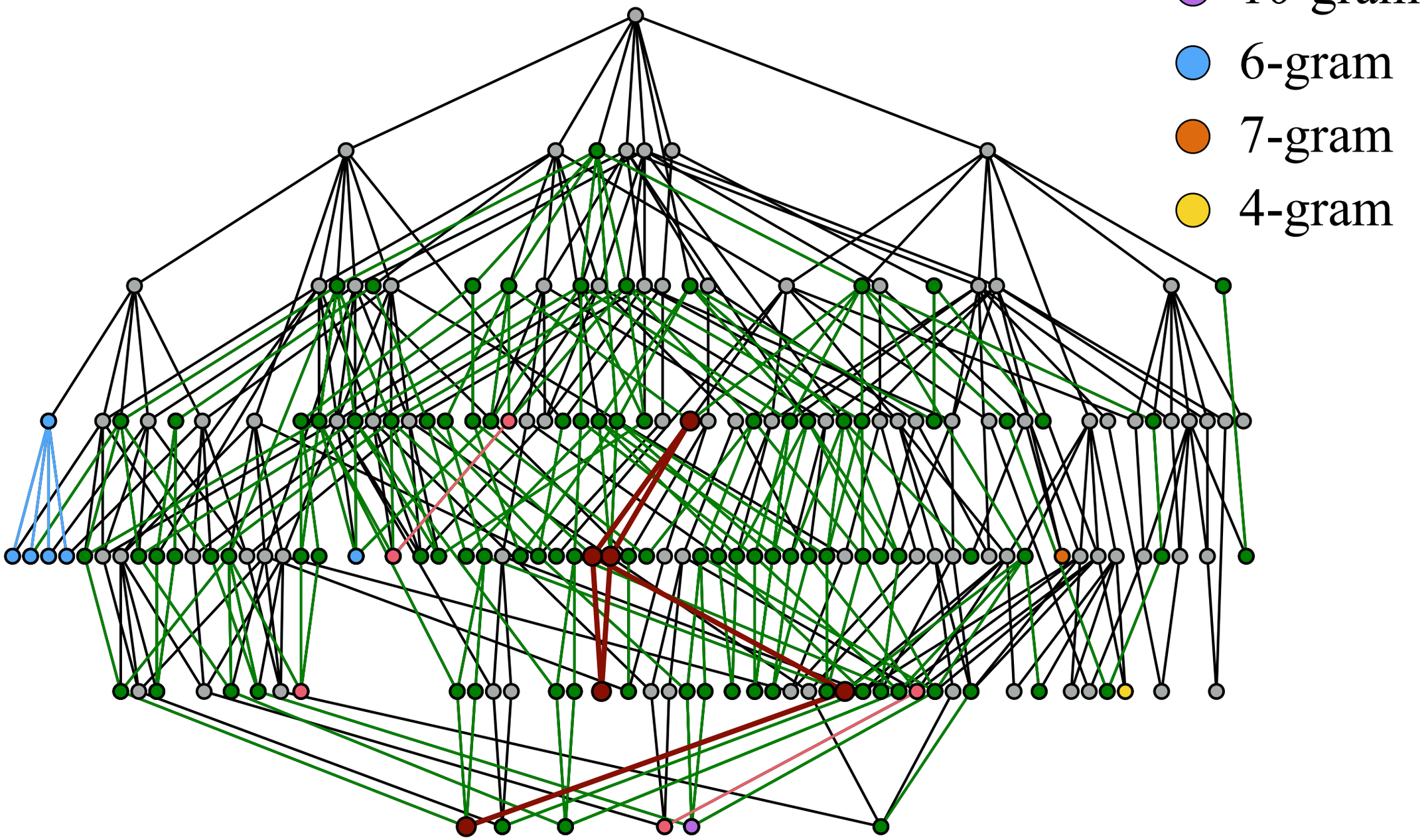
Loop 9



unlearned
1-gram
3-gram
10-gram
6-gram
7-gram
4-gram

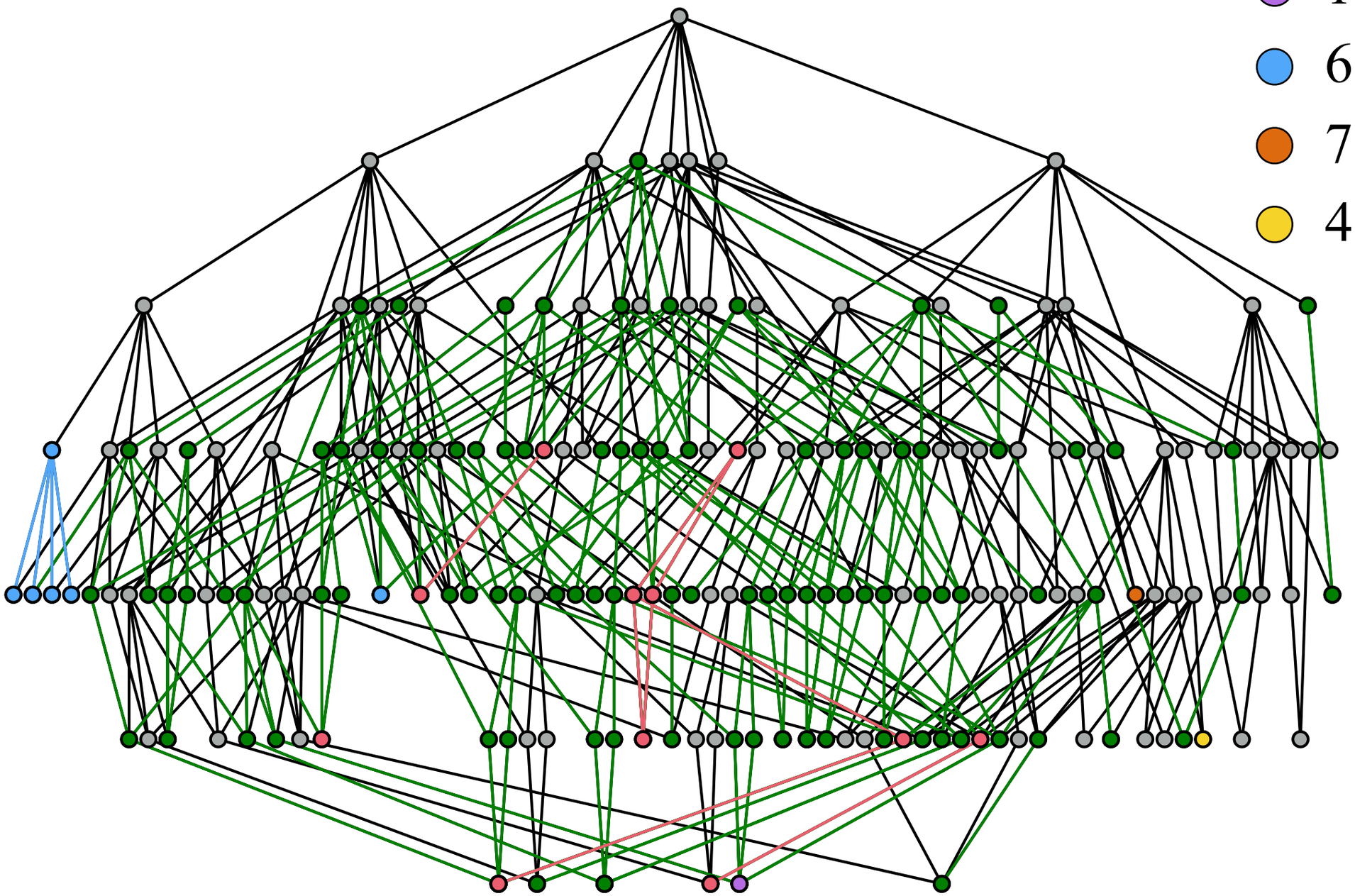
Part Writing (n-gram): 14th Chord

Loop 10



Part Writing (n-gram): 14th Chord

End of Loop 10



unlearned
1-gram
3-gram
10-gram
6-gram
7-gram
4-gram

Generalizing to Other Topic Domains

- Physics
- Cancer biology
- and more

References

- Haizi Yu, Igor Mineyev, and Lav R. Varshney. A Group-Theoretic Approach to Abstraction: Hierarchical, Interpretable, and Task-Free Clustering. *arXiv:1807.11167v1 [cs.LG], 2018*.
- Haizi Yu, Tianxi Li, and Lav R. Varshney. Probabilistic rule realization and selection. In *Proceedings of NIPS 2017*.
- Haizi Yu, and Lav R. Varshney. Towards deep interpretability (MUS-ROVER II): learning hierarchical representations of tonal music. In *Proceedings of ICLR 2017*.
- Haizi Yu, Lav R. Varshney, Guy E. Garnett, and Ranjitha Kumar. MUS-ROVER: A self-learning system for musical compositional rules. In *Proceedings of MUME 2016*.
- Haizi Yu, Lav R. Varshney, Guy E. Garnett, and Ranjitha Kumar. Learning interpretable musical compositional rules and traces. *In ICML 2016 Workshop on Human Interpretability in Machine Learning*.