Scaling self-supervised representation learning for symbolic piano performance

Louis Bradshaw, Honglu Fan, … Stella Biderman, Simon Colton

Presented at ISMIR 2025

06.01.2026 Clemence Reda

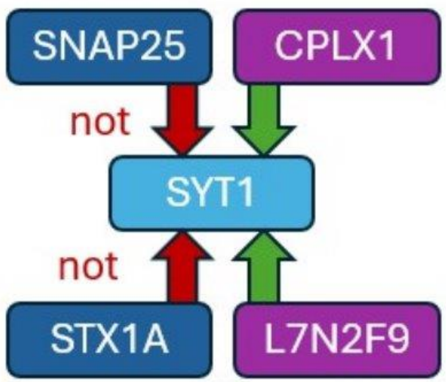
Background

1. Link with biology: dynamics in Boolean networks, canalisation-based representations
2. Symbolic music modeling
3. The SimCLR framework

Link with biology crafting general-purpose embeddings for cellular behavior predicted by Boolean networks

Remember Boolean networks?

Perturbation: overexpression (forced to True) or knockout (False)



Update formula for gene SYT1 at t

not(SNAP25(t) & STX1A(t))
or => SYT1(t+1)
L7N2F9(t) and CPLX1(t)

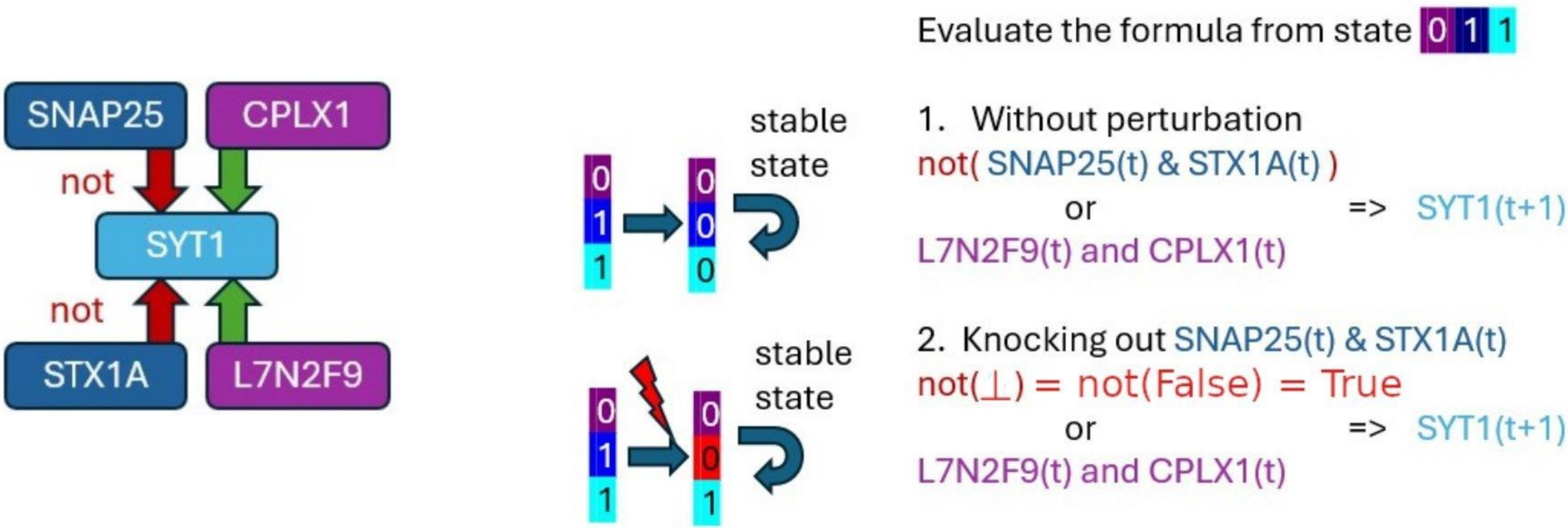
Get analytically / computationally the **stable states**

---

[21] *Kauffman (1969). Journal of Theoretical Biology 22.3, pp. 437–467*
Thomas (1973). Journal of Theoretical Biology 42.3, pp. 563–585

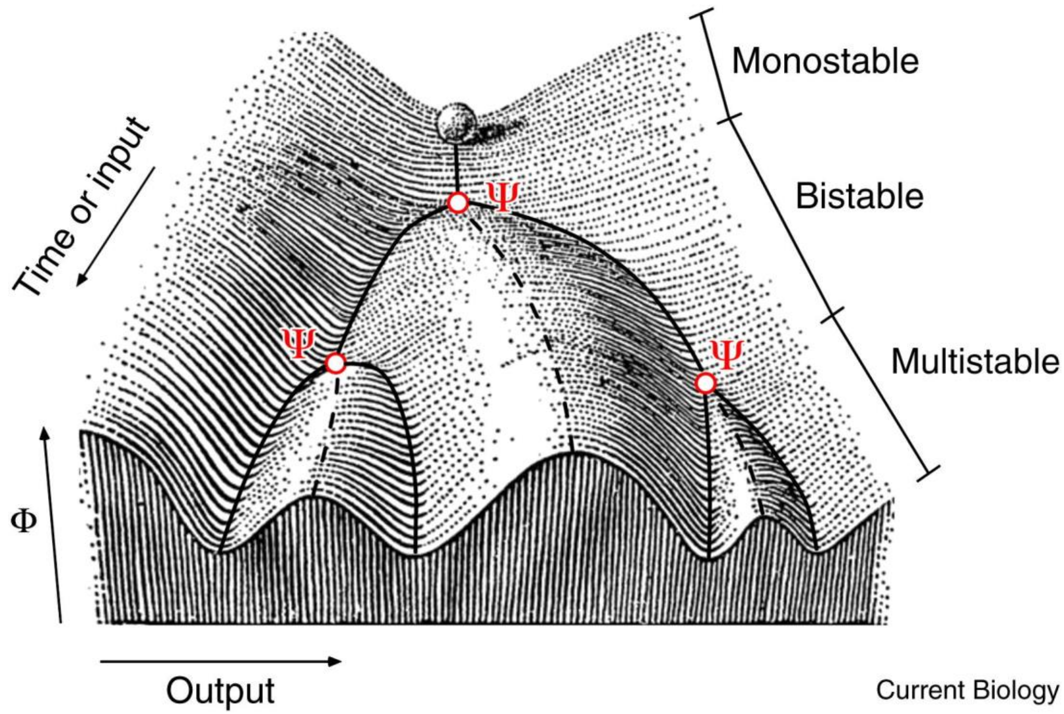
Link with biology crafting general-purpose embeddings for cellular behavior predicted by Boolean networks

Looking at static networks VS cellular behavior



Canalisation crafting general-purpose embeddings for cellular behavior predicted by Boolean networks

Conrad Waddington's (epigenetic) landscape in 1942

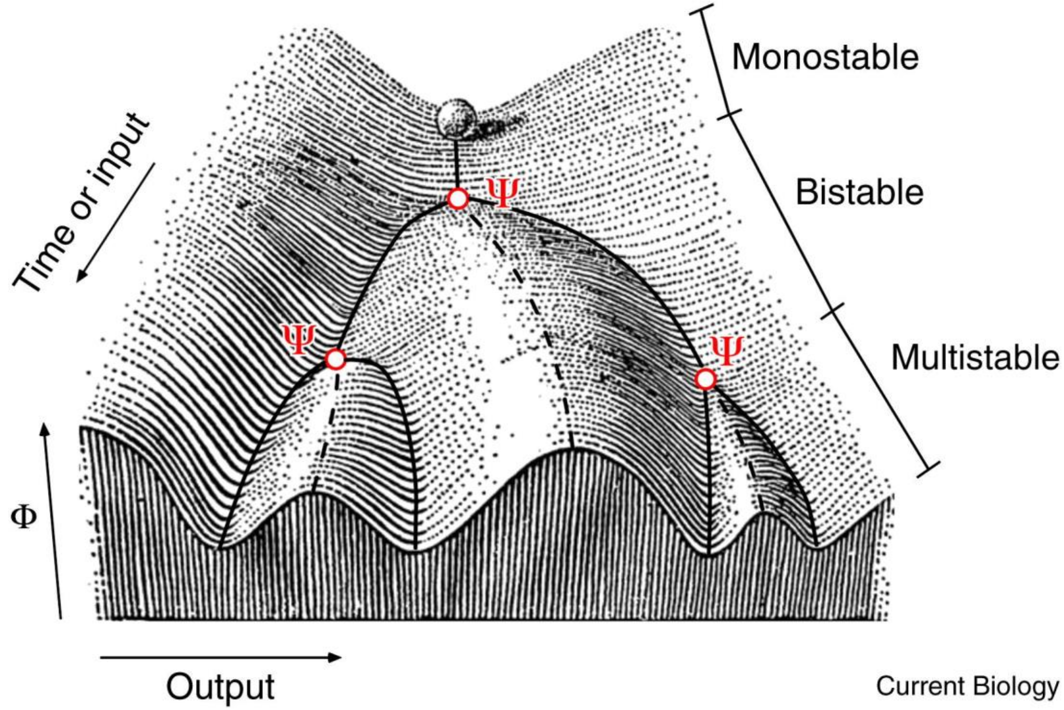


Source from **[1]**

[1] Ferrell. (2012). *Current biology*, *22*(11), R458-R466.

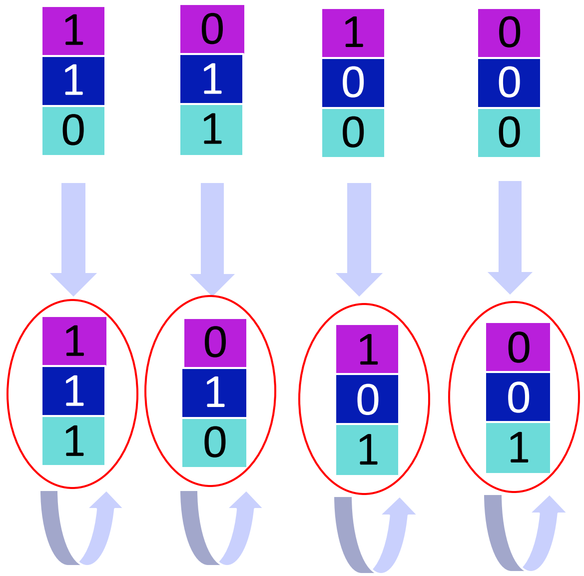
Canalisation crafting general-purpose embeddings for cellular behavior predicted by Boolean networks

Conrad Waddington's (epigenetic) landscape in 1942



Source from [1]

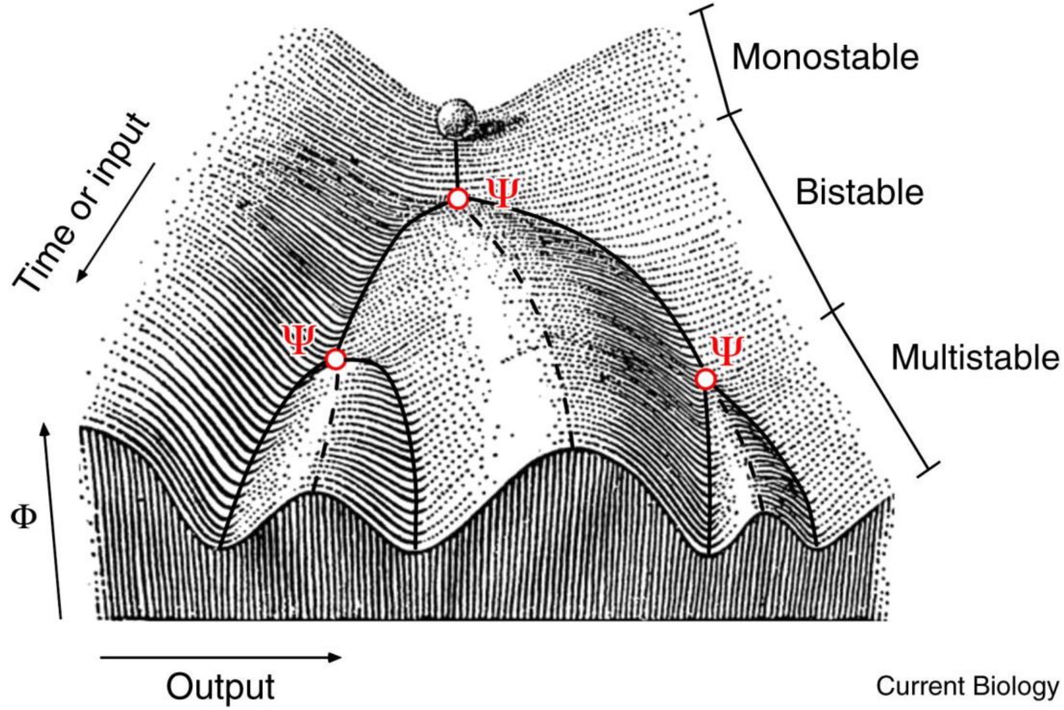
State transition diagram induced by Boolean functions



[1] Ferrell. (2012). *Current biology*, *22*(11), R458-R466.

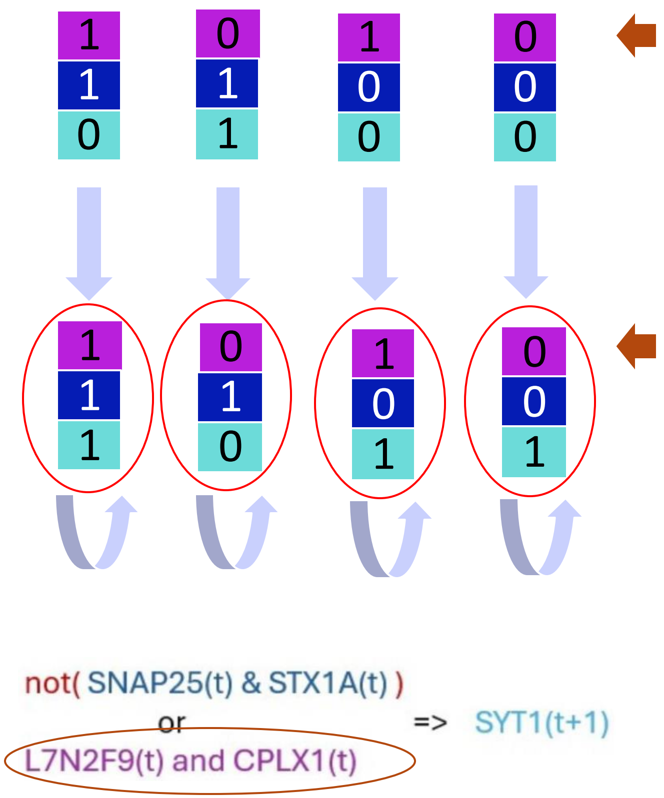
Canalisation crafting general-purpose embeddings for cellular behavior predicted by Boolean networks

Conrad Waddington's (epigenetic) landscape in 1942



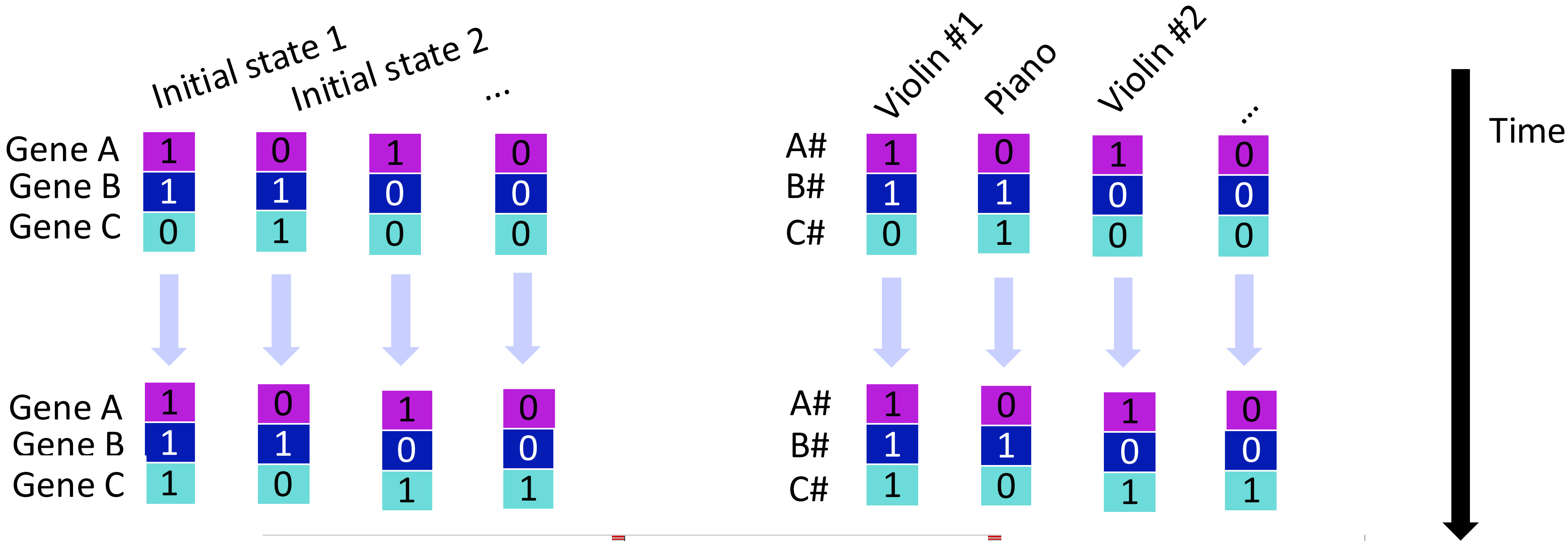
Source from **[1]**

State transition diagram induced by Boolean functions



[1] Ferrell. (2012). *Current biology*, *22*(11), R458-R466.
[2] He & Macauley. (2016). *Physica D: Nonlinear Phenomena*, *314*, 1-8.

Symbolic music modeling Musical Instrument Digital Interface (MIDI) file format




```
MTrk
0:0:0 Meta Text "Clarinet"
0:0:0 ProgCh ch=3 prog=71
0:0:240 On  ch=3 note=g5  vol=55
0:1:0   Off ch=3 note=g5  vol=90
0:1:0   On  ch=3 note=g5  vol=55
0:1:240 Off ch=3 note=g5  vol=90
0:1:240 On  ch=3 note=g5  vol=55
1:0:0   Off ch=3 note=g5  vol=90
1:0:0   On  ch=3 note=eb5 vol=55
2:0:0   Off ch=3 note=eb5 vol=90
2:0:240 On  ch=3 note=f5  vol=55
```

```
MTrk
0:0:0 Meta Text "Viola"
0:0:0   ProgCh ch=11 prog=48
0:0:240 On  ch=11 note=g4  vol=90
0:1:0   Off ch=11 note=g4  vol=90
0:1:0   On  ch=11 note=g4  vol=90
0:1:240 Off ch=11 note=g4  vol=90
0:1:240 On  ch=11 note=g4  vol=90
1:0:0   Off ch=11 note=g4  vol=90
1:0:0   On  ch=11 note=eb4 vol=90
2:0:0   Off ch=11 note=eb4 vol=90
2:0:240 On  ch=11 note=f4  vol=90
```

```
MTrk
0:0:0 Meta Text "Bass"
0:0:0 ProgCh ch=13 prog=48
0:0:240 On  ch=13 note=g2  vol=99
0:1:0   Off ch=13 note=g2  vol=90
0:1:0   On  ch=13 note=g2  vol=99
0:1:240 Off ch=13 note=g2  vol=90
0:1:240 On  ch=13 note=g2  vol=99
1:0:0   Off ch=13 note=g2  vol=90
1:0:0   On  ch=13 note=eb2 vol=99
2:0:0   Off ch=13 note=eb2 vol=90
2:0:240 On  ch=13 note=f2  vol=99
```

The SimCLR framework [1] contrastive learning of visual representations

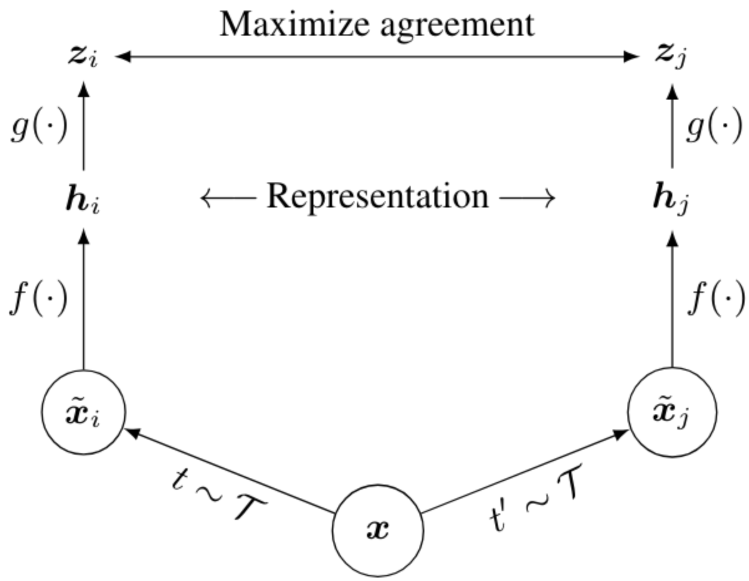
normalized temperature-scaled
cross-entropy loss

Apply nonlinear projection

Apply encoder

Augmented image #2

Augmented image #1



Initial image

Source from **[3]**

[3] Chen, et al. (2020, November). In *International conference on machine learning* (pp. 1597-1607). PMLR.

Content of the paper

"We introduce and open-source Aria, a pretrained autoregressive transformer model trained on transcriptions of solo piano recordings"

Specific issues:

- Refined contrastive learning approach on ~60k hours of music
- Generative capabilities & representation learning with few labels

Prior approaches:

- Transformers for generation, Autoencoders for representations

Content of the paper

"We introduce and open-source Aria, a pretrained autoregressive transformer model trained on transcriptions of solo piano recordings"

1. Tokenization scheme
2. Transformer architecture and preprocessing
3. Contrastive representation learning
4. Experiments: Generative capabilities
5. Experiments: Representation-based tasks

Tokenization scheme more adapted to transformers

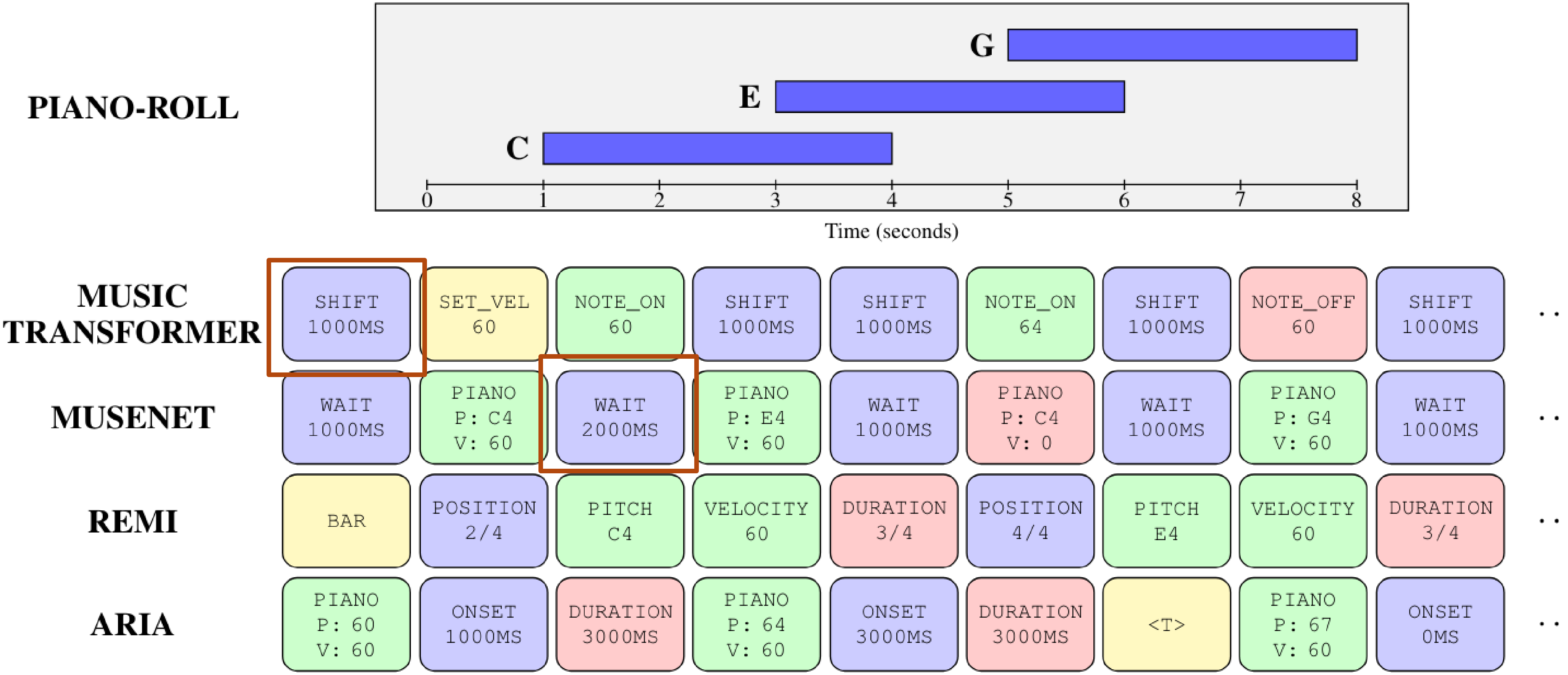


Figure 2 from paper

Tokenization scheme more adapted to transformers

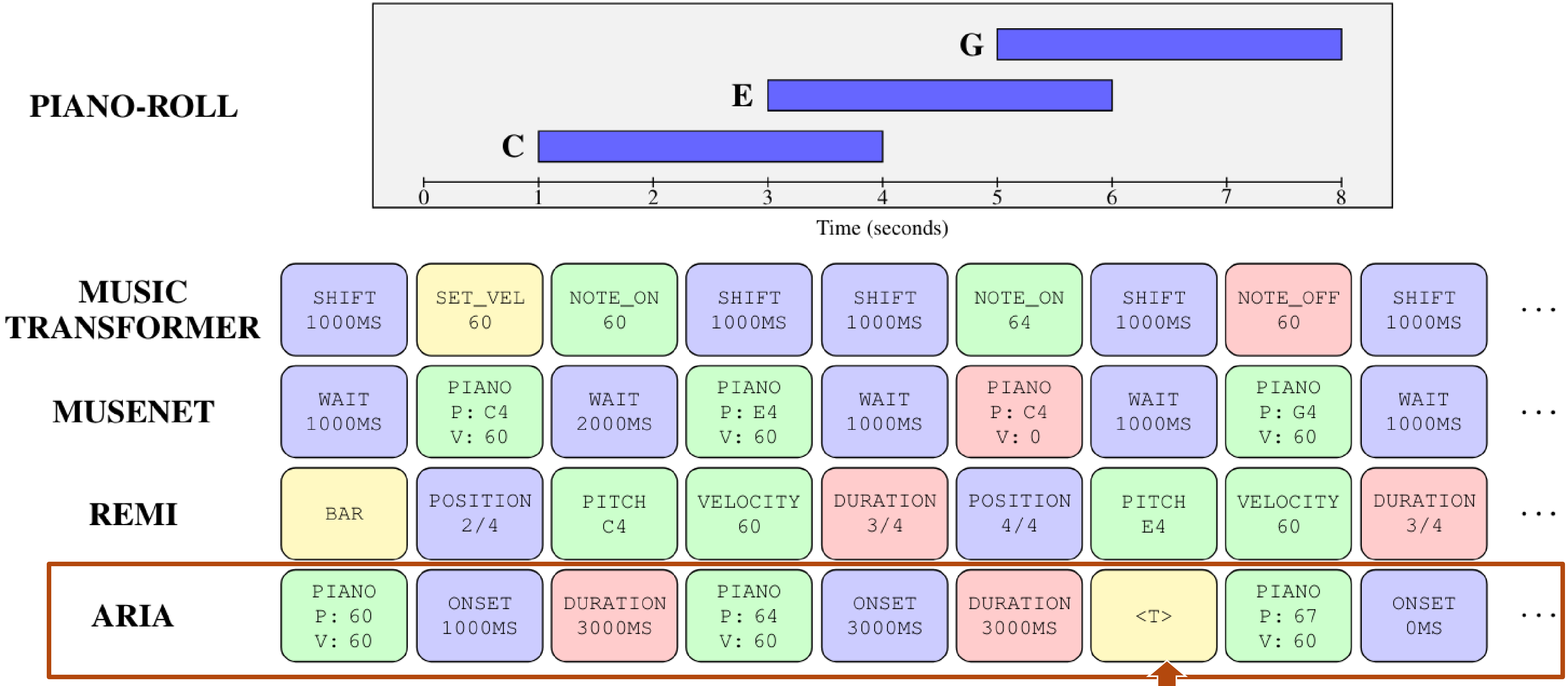


Figure 2 from paper

Transformer architecture and preprocessing

Modified LLaMa 3.2 model ~0.7B parameters, simple multi-head attention and layer normalization

Transformer architecture and preprocessing

Modified LLaMa 3.2 model ~0.7B parameters, simple multi-head attention and layer normalization

Preprocessing ARIA-MIDI data set remove duplicates, "messy" pieces
Context length: ~8k tokens
Data augmentation: random transposition, varying tempo and velocity

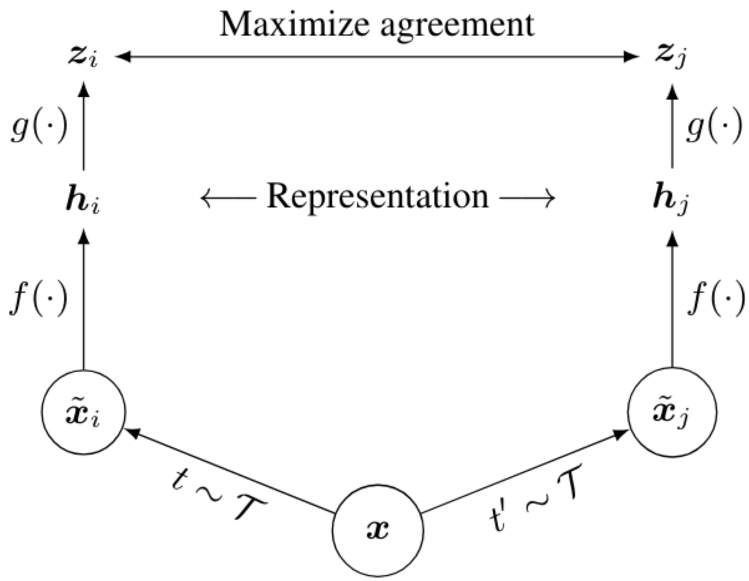
Transformer architecture and preprocessing

Modified LLaMa 3.2 model ~0.7B parameters, simple multi-head attention and layer normalization

Preprocessing ARIA-MIDI data set remove duplicates, "messy" pieces
Context length: ~8k tokens
Data augmentation: *random transposition*, *varying tempo* and *velocity*

Finetuning for generation single epoch on short solo piano prompts (LoRA?)
Add a new token <D> 100 tokens before the end of the prompt

Contrastive representation learning adapted from the SimCLR framework



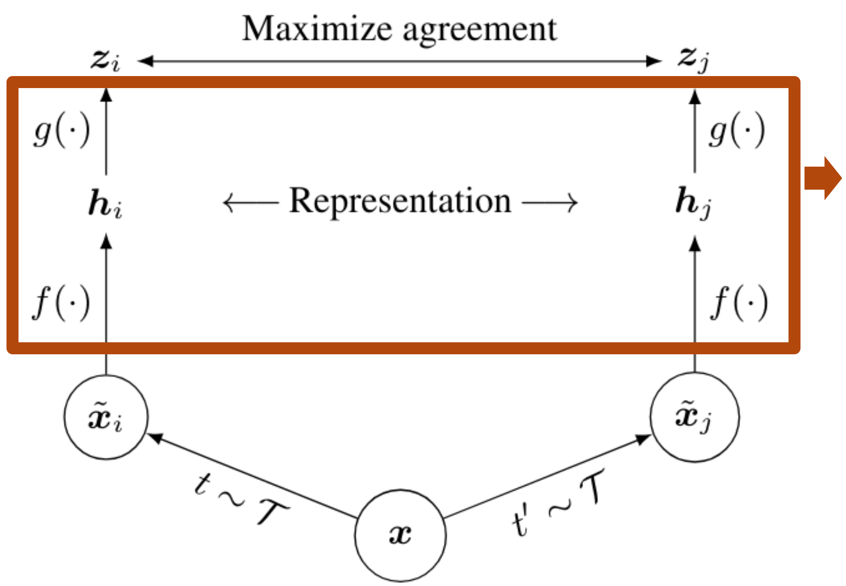
Initial MIDI file

Source from **[3]**

Two *consecutive* slices (300-2k tokens) with data augmentation

[3] Chen, et al. (2020, November). In *International conference on machine learning* (pp. 1597-1607). PMLR.

Contrastive representation learning adapted from the SimCLR framework



Initial MIDI file

Source from [3]

512-sized embeddings

Transformer with embedding head

Two *consecutive* slices (300-2k tokens) with data augmentation

[3] Chen, et al. (2020, November). In *International conference on machine learning* (pp. 1597-1607). PMLR.

Experiments: Generative capabilities

45" continuation of 15" prompts random pairwise A/B human evaluation

Compared Model	**Wins**	**Ties**	**Losses**	**p-value**
AM Transformer	38	0	6	$9.43e\text{-}7$
Suno 3.5	18	9	21	$7.49e\text{-}1$
MusicGen	49	1	0	$3.55e\text{-}15$
Ground Truth	15	9	17	$8.60e\text{-}1$

They say there are 46 participants, yet most lines do not sum to 46...

40 continuations /model (5 subgenres x 8 prompts)

Experiments: Representation-based tasks

"Composer" classification task averaging slice embeddings within a MIDI file

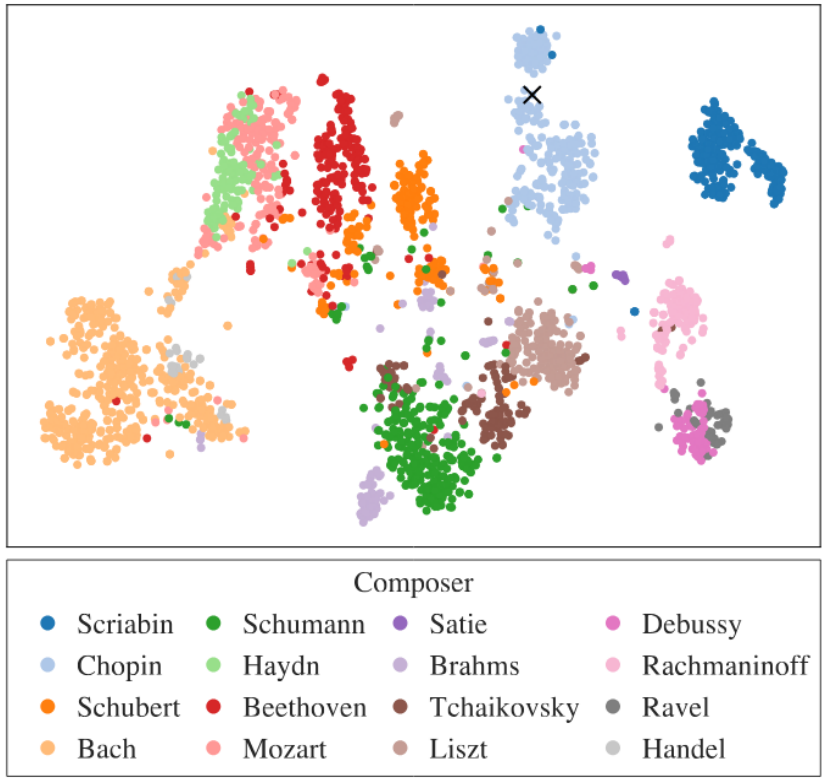


Figure 1 from paper

Experiments: Representation-based tasks

custom classification tasks (Genre, Form, Musical Period, Composer) | benchmarks (Pianist8, VG-MIDI)

Model	Genre		Form		Musical Period		Composer		Pianist8		VG-MIDI	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Main Results												
MERT	83.00	83.00	63.89	63.90	69.50	68.94	69.60	69.30	65.06	65.18	45.45	40.37
M3	85.10	85.10	69.88	70.12	71.20	70.81	71.90	71.72	81.93	81.48	54.55	46.13
CLaMP 3	89.10	89.10	77.79	77.97	80.60	80.20	84.50	84.46	80.72	79.76	45.45	36.53
Aria_{Emb}	92.40	92.40	82.45	82.57	84.70	84.69	90.50	90.49	**91.57**	**92.38**	63.64	63.96
Aria_{Ft}	**93.20**	**93.20**	**87.53**	**87.59**	**86.50**	**86.53**	**96.30**	**96.32**	91.56	92.03	**68.18**	**69.55**

supervised finetuned model with the replacement by a classifier head

Perspectives

1. Comments on the paper
2. Why is it interesting for BioComp?

My comments on the paper

Strengths:

- The method is elegant and adapted to multi-tasking
- Results on embedding learning are interesting
- Open-source: https://github.com/EleutherAI/aria

Weaknesses:

- Results on generative capabilities seem strange to me
- **Only applied to solo piano transcriptions (no multi-track)**
- Could other data augmentations be applied? (e.g., changes to inter-onset intervals)

Your comments?

Why is it interesting for BioComp?

Adaptation of methods applied to imaging to other time-series data

- Subtle dynamical changes
- Required consistency across time (harmonics...)
- Time integration into transformer models
- Restricted number of samples

Under review as a conference paper at ICLR 2026

PIANIST TRANSFORMER: TOWARDS EXPRESSIVE PIANO PERFORMANCE RENDERING VIA SCALABLE SELF-SUPERVISED PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing methods for expressive music performance rendering rely on supervised learning over small labeled datasets, which limits scaling of both data volume and model size, despite the availability of vast unlabeled music, as in vision

[4] https://openreview.net/pdf?id=Kq9MMEynGW

(perhaps a bit of a stretch for Boolean networks)