CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Instituto Milenio
Fundamentos
de los datos

MUSIB: Music Inpainting Benchmark

Mauricio Araneda
Guia: Felipe Bravo
Co-guia: Denis Parra

Introduction

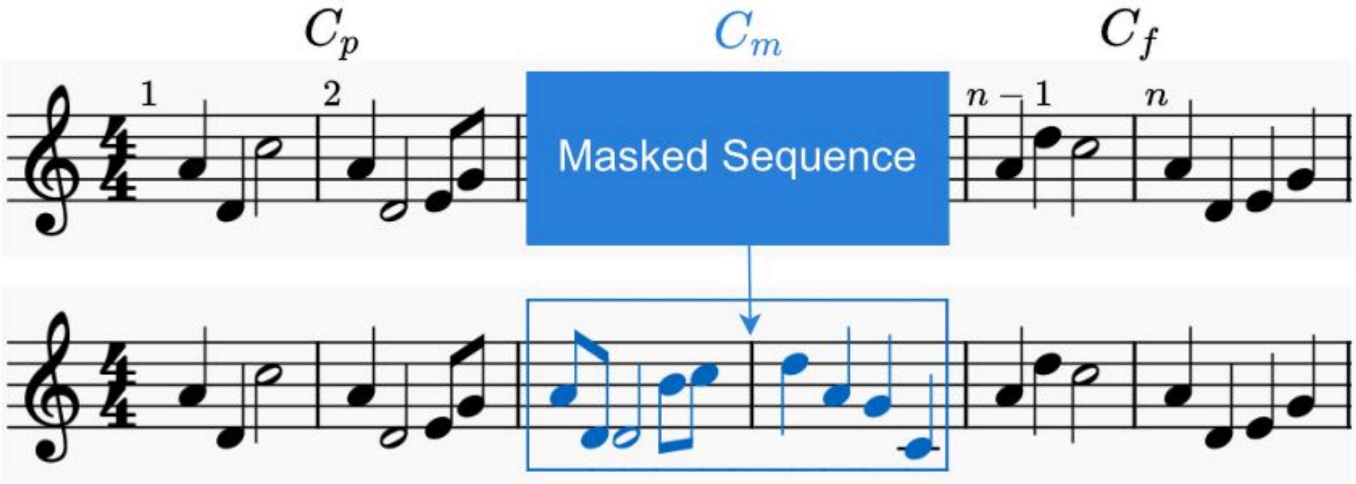
Composing Music is Complex

Musical Score Inpainting

- Sub-task of Automated Music Generation that aims to infill incomplete musical pieces.
- Easy interaction with the user: current ideas they want to join/extend

Musical Score Inpainting

Musical Score Inpainting



Issues with Music Inpainting evaluation

- Proposed methods lack of standardized evaluation setups
- Different data representation, datasets, metrics and baselines
- We dont know the state of the art and if we are making progress

Problem Statement

Evaluation Challenges

- Metrics values differ when changing representations for the exact same data.

Evaluation Challenges

- Metrics values differ when changing representations for the exact same data.
- The sets of metrics for evaluation changes from paper to paper, measuring different features.

Evaluation Challenges

- Metrics values differ when changing representations for the exact same data.
- The sets of metrics for evaluation changes from paper to paper, measuring different features.
- Training and evaluation of models done over different datasets that vary in characteristics such as: format, number of samples, style, notes distribution, etc.

Evaluation Challenges

- Metrics values differ when changing representations for the exact same data.
- The sets of metrics for evaluation changes from paper to paper, measuring different features.
- Training and evaluation of models done over different datasets that vary in characteristics such as: format, number of samples, style, notes distribution, etc.
- The output is generated through a random process.

Hypothesis

Hypothesis

It is possible to find a unifying pattern across several models of musical score inpainting that enables a direct comparison of approaches.

Additionally, we argue that it is possible to extend current evaluation procedures to measure the expected variability of a model.

General Objective

To develop an evaluation framework to properly compare different approaches for musical score inpainting, thus providing solid evidence to define the current progress of this task and its state of the art.

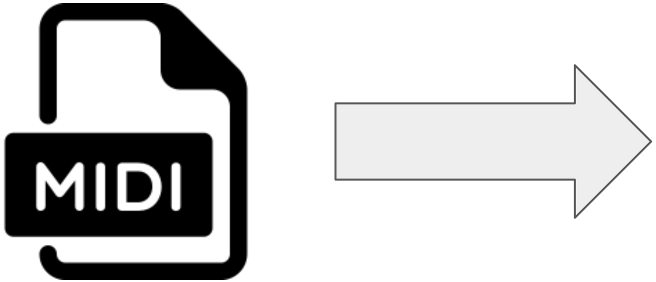
Preliminary Concepts
&
Background

Representation

- Two dimensions:
  - Pitch
  - Rhythm

* There are other dimensions such as dynamics or timbre that we are not discussing here

Representation - MIDI to Vector



Track	Time	Event	Channel	Note	Velocity
2	96	Note_on	0	60	90
2	192	Note_off	0	60	0
2	192	Note_on	0	62	90
2	288	Note_off	0	62	0
2	288	Note_on	0	64	90
2	384	Note_off	0	64	0

Representation - MIDI to Vector

```
Track, Time, Event, Channel, Note, Velocity
 2,   96, Note_on,   0, 60, 90
 2,  192, Note_off,  0, 60,  0
 2,  192, Note_on,   0, 62, 90
 2,  288, Note_off,  0, 62,  0
 2,  288, Note_on,   0, 64, 90
 2,  384, Note_off,  0, 64,  0
```

Representation - MIDI to Vector

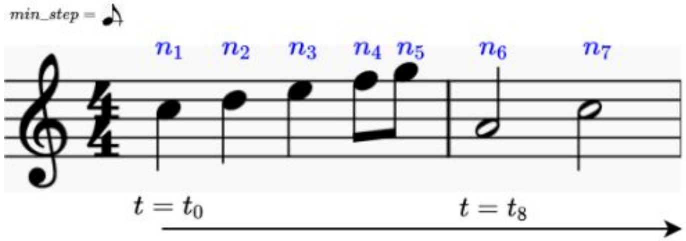
Track	Time	Event	Channel	Note	Velocity
2	96	Note_on	0	60	90
2	192	Note_off	0	60	0
2	192	Note_on	0	62	90
2	288	Note_off	0	62	0
2	288	Note_on	0	64	90
2	384	Note_off	0	64	0

```xml
<score-partwise version="4.0">
  <part-list>
    <score-part id="P1">
      <part-name>Music</part-name>
    </score-part>
  </part-list>
  <part id="P1">
    <measure number="1">
      <attributes>
        <divisions>1</divisions>
        <key>
          <fifths>0</fifths>
        </key>
        <time>
          <beats>4</beats>
          <beat-type>4</beat-type>
        </time>
        <clef>
          <sign>G</sign>
          <line>2</line>
        </clef>
      </attributes>
      <note>
        <pitch>
          <step>C</step>
          <octave>4</octave>
        </pitch>
        <duration>4</duration>
        <type>whole</type>
      </note>
    </measure>
  </part>
</score-partwise>
```

Representation - Time Discretization

- Since time is a continuous space, we need to discretize each note start time and its duration to fit on a time grid.
- The choice of how much resolution we want for this grid is arbitrary:
    - A common approach is to consider a 4 4 measure to have 16 time-steps, equally spaced.
    - This means that the minimal step is fixed to be a sixteenth note (semi corchea).

Representation - Note Sequence



$$x = [C_4, \_, D_4, \_, E_4, \_, F_4, G_4, A_3, \_, \_, \_, C_4, \_, \_, \_]$$

MUSIB

Motivation

- Several research communities have highlighted the need for stronger standards on evaluation and reproducibility.
- Most evaluations of musical inpainting models do not share representation of data, metrics, datasets, or baselines.
- We need to replicate the results of these approaches under standardized conditions for fair comparison and express them in the same metrics

What is MUSIB

- 4 models
- 2 datasets
- 7 metrics

Current Evaluation Conditions

Model	Representation	Dataset	Metrics
VLI	REMI-16	AILabs1k7	H1, H4, GS
SketchNet	NoteSeq-24	IrishFolk	NLL, pAcc, rAcc
InpaintNet	NoteSeq-24	IrishFolk	NLL
A-RNN	NoteSeq-16	JSBChorales	Accuracy, JS Div

Experimental Setup

Experiment

- The evaluation is done over extract of songs that we call contexts
- Each context size is fixed to be 16-measures
  - Past and Future are 6-measures long
  - Middle is 4-measures long
- Each measure is discretized by 16 or 24 time steps depending on the model implementation.
- Split of 8/1/1 ratio for train/val/test.
- Early Stopping with a patience of 5 epochs.

Models

Anticipation-RNN

- Based on RNNs
- Use of Unary Constraints

Music InpaintNet

- Based on a combination of VAE and RNNs
- Use of latent space

Music SketchNet

- Based on a combination of VAE and RNN
- Separate Encoding for pitch and rhythm

Variable Length Piano Infilling

- Based on XLNET
- Encodes notes as word tokens for a pre-trained language model.

Datasets

Datasets

- JSB Chorales and IrishFolkSong datasets.
- We prioritized these datasets due to:
    - They have been used to train several musical inpainting models.
    - Represent different musical styles.
    - Important differences in size.

Pipeline

- We filtered:
  - Invalid files (i.e., no instruments or zero-length)
  - Repeated files (files with the same hash)
  - Files shorter than 16-measures long.

Dataset Sizes

Dataset	Songs	Contexts
JSB Chorales	171	2360
IrishFolkSong	17358	324556

Metrics

Note Metrics

- One-on-one comparisons between the generated data and the expected true data.
- Agnostic to representation of data.

Note Metrics

♩ = 4 time steps

True data

Pred data

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	Pitch	Duration
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	Pitch	Duration
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	Pitch	Duration
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	Pitch	Duration
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	TP	FP	FN
Position	2	1	1

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

Note Metrics

♩ = 4 time steps

True data

Pred data

	TP	FP	FN
Position	2	1	1

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

$$pos_{precision} = \frac{2}{2+1} = 0.67$$

$$pos_{recall} = \frac{2}{2+1} = 0.67$$

$$pos_{f1} = 0.67$$

Note Metrics

♩ = 4 time steps

True data

Pred data

	TP	FP	FN
Position	2	1	1
Pitch	2	0	–
Duration	1	1	–

	Position	*Pitch*	*Duration*
$n_{1\ true}$	0	60	4
$n_{2\ true}$	4	62	4
$n_{3\ true}$	8	64	4

	Position	*Pitch*	*Duration*
$n_{1\ pred}$	0	60	4
$n_{2\ pred}$	8	64	2
$n_{3\ pred}$	10	65	2

$$pos_{precision} = \frac{2}{2+1} = 0.67$$

$$pos_{recall} = \frac{2}{2+1} = 0.67$$

$$pos_{f1} = 0.67$$

$$pitch_{acc} = \frac{2}{2+0} = 1$$

$$rhythm_{acc} = \frac{1}{1+1} = 0.5$$

Divergence Metrics

Although note metrics are useful for one-on-one comparison, there are cases in music generation where the attributes can not be directly compared since there are multiple correct options.

This variability in music is common and even desirable. However, there is a lack of methods to measure the correct variability of these attributes in generated data.

Divergence Metrics

How do we verify that a given musical attribute in a set of predicted songs is within the correct range of variability?

We argue that we need to look at the distribution of this attribute in true data and measure how close it is to the one in generated data.

By measuring this closeness between distributions we relax the condition of correctness to accept multiple valid answers.

Divergence Metrics

$Y_{true}^{(0)} =$ 

Divergence Metrics

$Y_{true}^{(0)} =$  $\xrightarrow{f(\cdot)}$

Divergence Metrics

$Y^{(0)}_{true} =$  $\xrightarrow{f(\cdot)} f_{Y^{(0)}_{true}} \in [0, 1]$

Divergence Metrics

$Y_{true}^{(0)} =$  $\xrightarrow{f(\cdot)} f_{Y_{true}^{(0)}} \in [0, 1]$

$Y_{true}^{(1)} =$  $\xrightarrow{f(\cdot)} f_{Y_{true}^{(1)}} \in [0, 1]$

Divergence Metrics

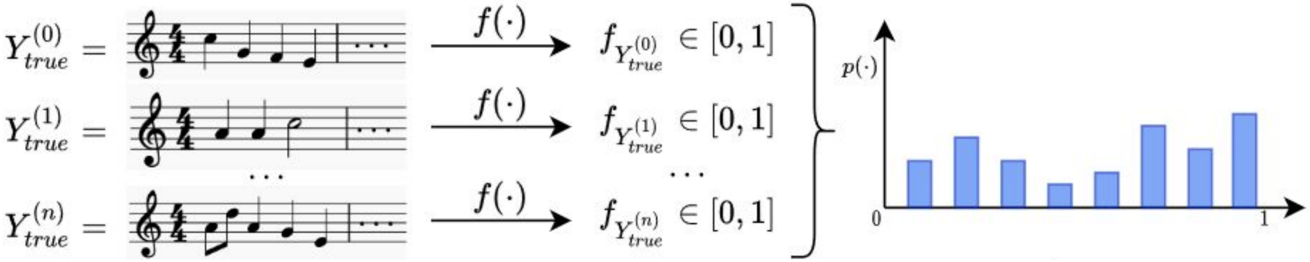
$Y^{(0)}_{true} =$  $\xrightarrow{f(\cdot)} f_{Y^{(0)}_{true}} \in [0, 1]$

$Y^{(1)}_{true} =$  $\xrightarrow{f(\cdot)} f_{Y^{(1)}_{true}} \in [0, 1]$

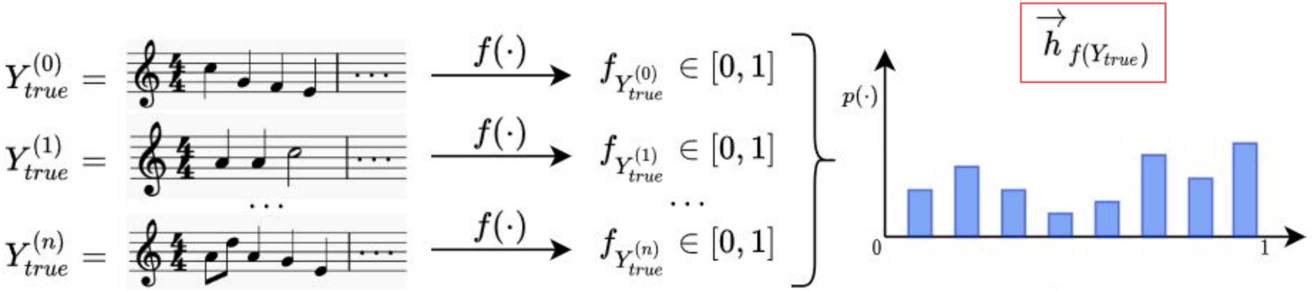
$\cdots \qquad \cdots$

$Y^{(n)}_{true} =$  $\xrightarrow{f(\cdot)} f_{Y^{(n)}_{true}} \in [0, 1]$

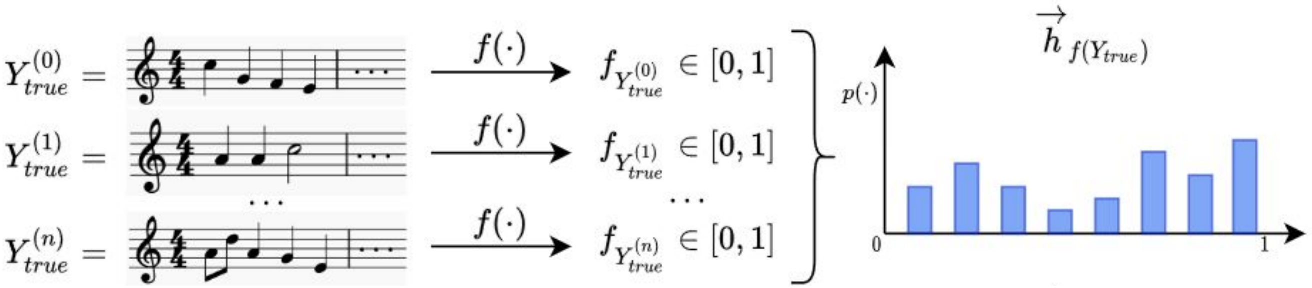
Divergence Metrics



Divergence Metrics



Divergence Metrics



Divergence Metrics

$Y^{(0)}_{pred} =$  $\xrightarrow{f(\cdot)} f_{Y^{(0)}_{pred}} \in [0, 1]$

Divergence Metrics

$Y^{(0)}_{pred} =$  $\xrightarrow{f(\cdot)} f_{Y^{(0)}_{pred}} \in [0, 1]$

$Y^{(1)}_{pred} =$  $\xrightarrow{f(\cdot)} f_{Y^{(1)}_{pred}} \in [0, 1]$

Divergence Metrics

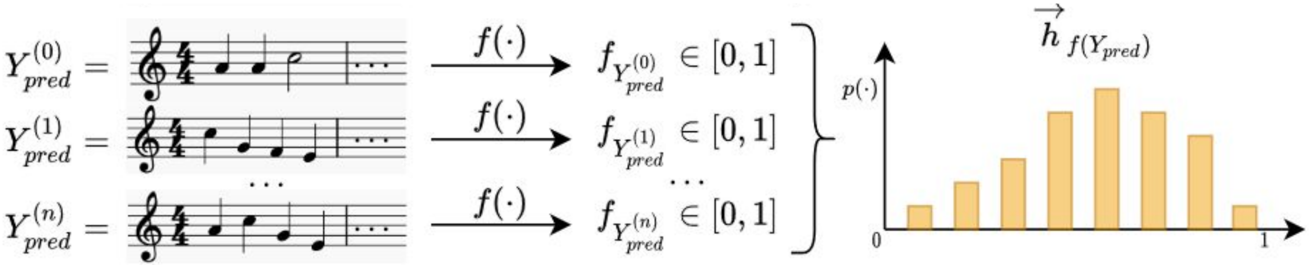
$Y_{pred}^{(0)} =$  $\xrightarrow{f(\cdot)} f_{Y_{pred}^{(0)}} \in [0, 1]$

$Y_{pred}^{(1)} =$  $\xrightarrow{f(\cdot)} f_{Y_{pred}^{(1)}} \in [0, 1]$

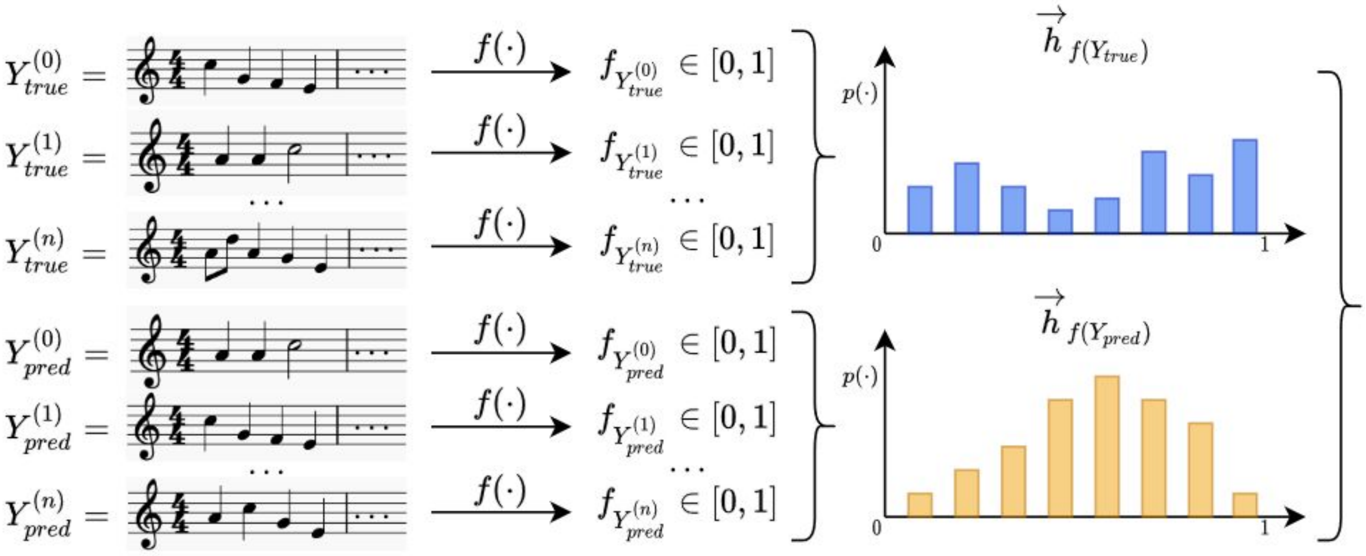
$\cdots \qquad \cdots$

$Y_{pred}^{(n)} =$  $\xrightarrow{f(\cdot)} f_{Y_{pred}^{(n)}} \in [0, 1]$

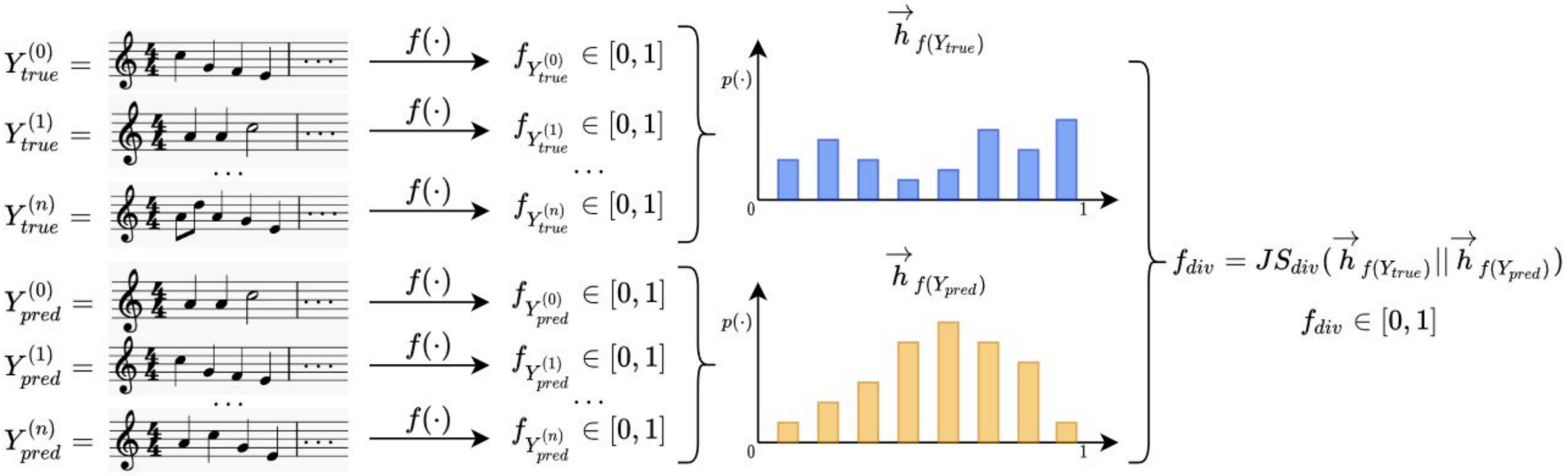
Divergence Metrics



Divergence Metrics



Divergence Metrics



Divergence Metrics

$$f_{div}(Y_{true}||Y_{pred}) = JS_{div}(\overrightarrow{h}_{f(Y_{true})}||\overrightarrow{h}_{f(Y_{pred})})$$

Divergence Metrics

In our work we propose three divergence metrics:

- Silence Divergence

Divergence Metrics

In our work we propose three divergence metrics:

- Silence Divergence
- Pitch Class Divergence

Divergence Metrics

In our work we propose three divergence metrics:

- Silence Divergence
- Pitch Class Divergence
- Groove Similarity Divergence

Results

Results - IrishFolk

IrishFolk Dataset (≈ 300K samples)

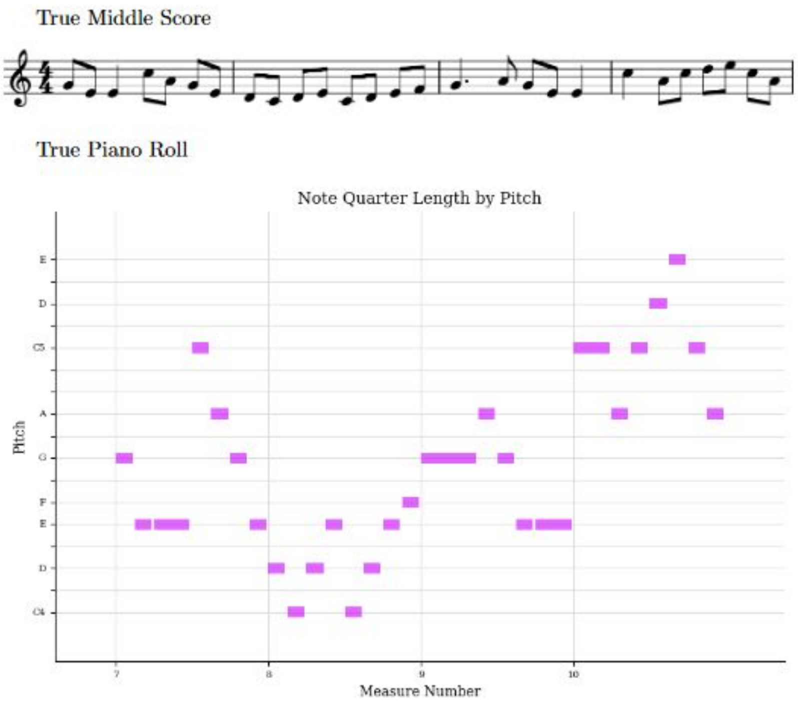
Model	$NLL \downarrow$	$pos_{F1} \uparrow$	$pAcc \uparrow$	$rAcc \uparrow$	$S_{div} \downarrow$	$H_{div} \downarrow$	$GS_{div} \downarrow$
Anticipation-RNN	0.453 (*0.662)	0.930	0.657	0.860	0.017	0.060	0.007
InpaintNet	0.487 (*0.662)	0.860	0.517	0.750	0.013	0.174	0.024
SketchNet	0.539 (*0.516)	0.914	0.560	0.868	**0.005**	0.134	0.009
VLI	0.059	**0.968**	**0.911**	**0.965**	0.015	**0.010**	**0.006**

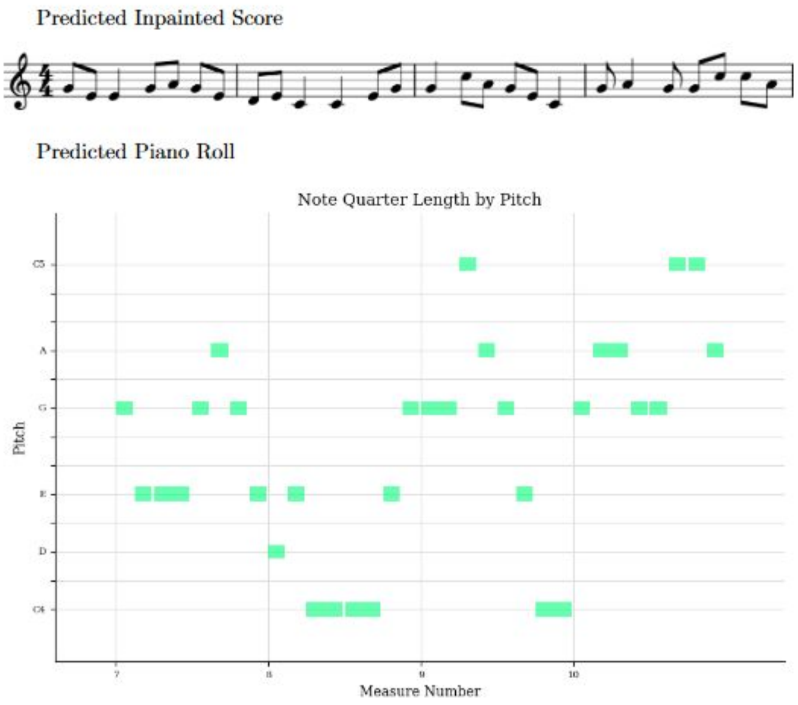
Results - JSB Chorales

JSB Chorales Dataset ($\approx$ 2.4K samples)

Model	$NLL \downarrow$	$pos_{F1} \uparrow$	$pAcc \uparrow$	$rAcc \uparrow$	$S_{div} \downarrow$	$H_{div} \downarrow$	$GS_{div} \downarrow$
Anticipation-RNN	0.459	0.832	0.243	0.682	0.240	0.525	0.232
InpaintNet	0.327	**0.852**	**0.505**	**0.788**	**0.059**	0.411	**0.153**
SketchNet	0.605	0.833	0.272	0.708	0.079	0.529	0.228
VLI	1.053	0.827	0.283	0.747	0.087	**0.286**	0.306

Results - IrishFolk

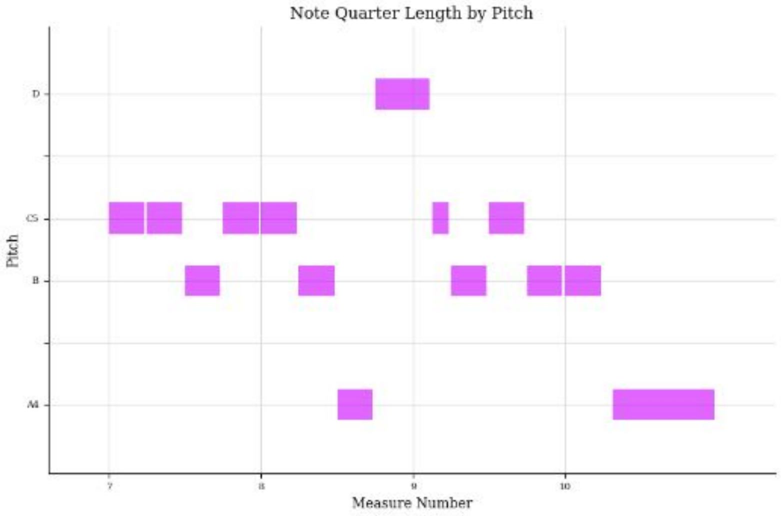




Results - JSB Chorales

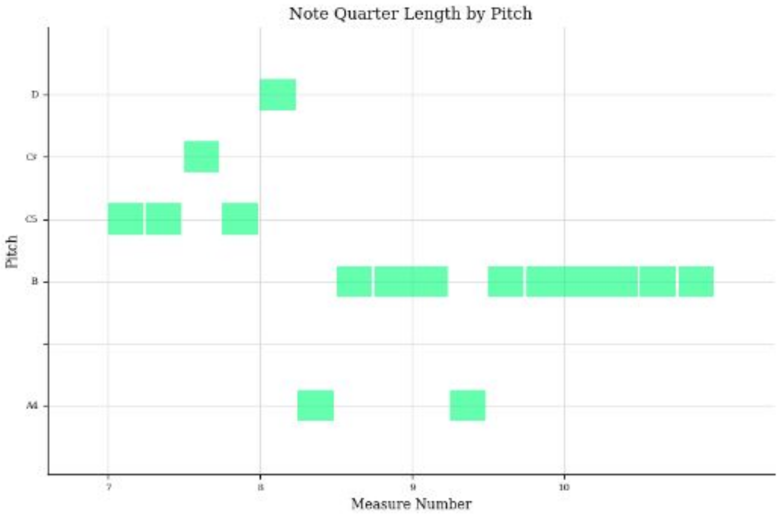
True Middle Score

True Piano Roll



Predicted Inpainted Score

Predicted Piano Roll



Conclusions

Conclusions

- We proposed MUSIB, a new standardization framework and benchmark for musical score inpainting evaluation.

Conclusions

- We proposed MUSIB, a new standardization framework and benchmark for musical score inpainting evaluation.
- We compiled, standardized and extended metrics to measure meaningful musical attributes.

Future Work

- Evaluation over:
  - Polyphonic music inpainting models

Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task

Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task
  - Data augmentation strategies

Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task
  - Data augmentation strategies
- Subjective evaluation with listeners to correlate with our proposed metrics.

Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task
  - Data augmentation strategies
- Subjective evaluation with listeners to correlate with our proposed metrics.
- Definition of new divergence metrics to capture new features such as:

Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task
  - Data augmentation strategies
- Subjective evaluation with listeners to correlate with our proposed metrics.
- Definition of new divergence metrics to capture new features such as:
  - Amount of repetition in a sequence

Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task
  - Data augmentation strategies
- Subjective evaluation with listeners to correlate with our proposed metrics.
- Definition of new divergence metrics to capture new features such as:
  - Amount of repetition in a sequence
  - Amount of polyphony

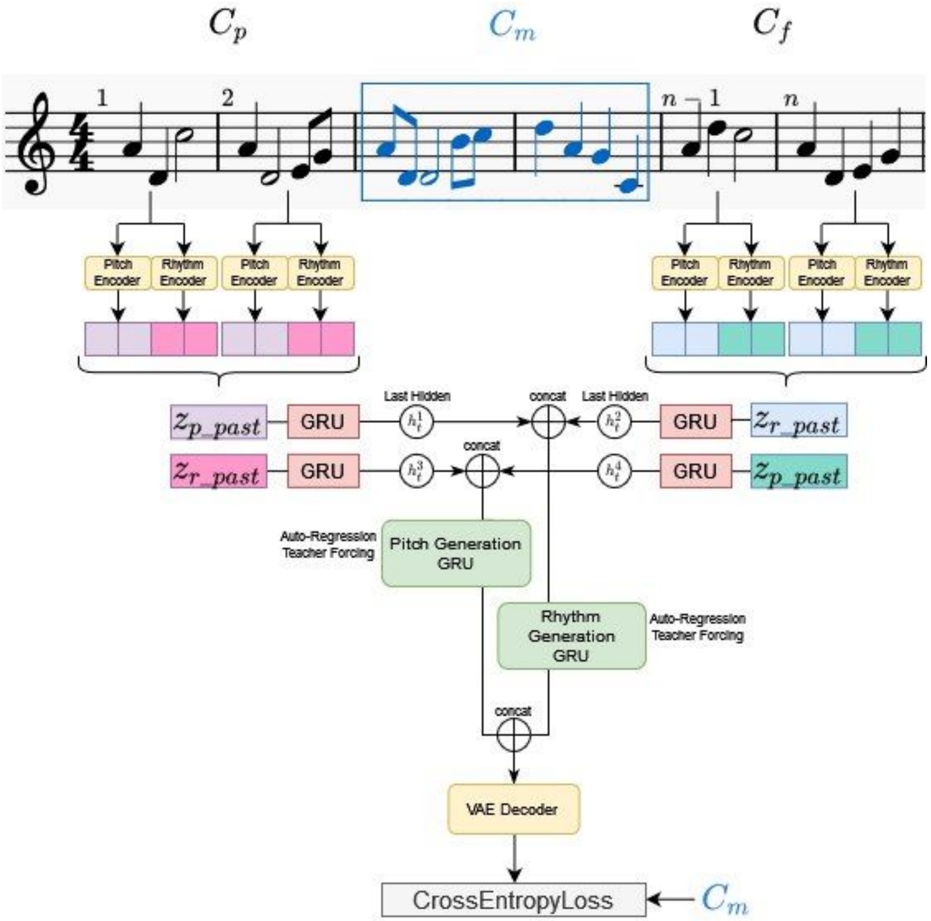
Future Work

- Evaluation over:
  - Polyphonic music inpainting models
  - Variable length infilling task
  - Data augmentation strategies
- Subjective evaluation with listeners to correlate with our proposed metrics.
- Definition of new divergence metrics to capture new features such as:
  - Amount of repetition in a sequence
  - Amount of polyphony
  - Etc

Thanks :)

Appendix

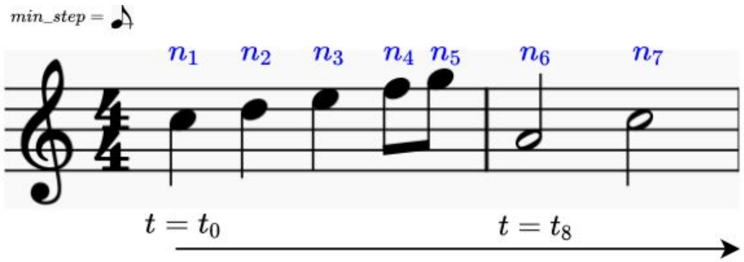
SketchNet



C_p
C_m
C_f
1
2
$n-1$
n
Pitch Encoder
Rhythm Encoder
Pitch Encoder
Rhythm Encoder
Pitch Encoder
Rhythm Encoder
Pitch Encoder
Rhythm Encoder
Last Hidden
concat
Last Hidden
z_{p_past}
GRU
h_t^1
h_t^2
GRU
z_{r_past}
concat
z_{r_past}
GRU
h_t^3
h_t^4
GRU
z_{p_past}
Auto-Regression Teacher Forcing
Pitch Generation GRU
Rhythm Generation GRU
Auto-Regression Teacher Forcing
concat
VAE Decoder
CrossEntropyLoss
C_m

Data Representation

(a)



(b) $x = [C_4, \_, D_4, \_, E_4, \_, F_4, G_4, A_3, \_, \_, \_, C_4, \_, \_, \_]$

(c)

$$
\begin{aligned}
x &= (x_{pitch}, x_{rhythm}) \\
x_{pitch} &= [C_4, D_4, E_4, F_4, G_4, A_3, C_4] \\
x_{rhythm} &= [1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0]
\end{aligned}
$$

(d) $x = [\,[n_1, n_2, n_3, n_4, n_5], [n_6, n_7]\,]$

n	Tempo	Bar Start	Position	Pitch	Velocity	Duration
n_1	120	1	0 | 8	C_4	90	♩
n_2	120	0	2 | 8	D_4	90	♩
n_3	120	0	4 | 8	E_4	90	♩
n_4	120	0	6 | 8	F_4	90	♪
n_5	120	0	7 | 8	G_4	90	♪
n_6	120	1	0 | 8	A_3	90	𝅗𝅥
n_7	120	0	4 | 8	C_4	90	𝅗𝅥

Silence Density

$$S(x) = \frac{1}{T}\sum_{t=0}^{T} \mathbb{1}_{n\_notes(x_t)=0}$$

Silence Divergence

$$S_{div}(Y_{true}||Y_{pred}) = JS_{div}(\overrightarrow{h}_{S(Y_{true})}||\overrightarrow{h}_{S(Y_{pred})})$$

Pitch Class Entropy

$$\sum_{i=0}^{n_1} \sum_{j=0}^{n_2} |\mathcal{H}_{m_i} - \mathcal{H}_{m_j}|$$

Pitch Class Entropy

$$\frac{1}{n_1 n_2} \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} |\mathcal{H}_{m_i} - \mathcal{H}_{m_j}|$$

Pitch Class Entropy

$$H(x) = \frac{1}{n_1 n_2} \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} |\mathcal{H}_{m_i} - \mathcal{H}_{m_j}|$$

Pitch Class Divergence

$$H_{div}(Y_{true}||Y_{pred}) = JS_{div}(\overrightarrow{h}_{H(Y_{true})}||\overrightarrow{h}_{H(Y_{pred})})$$

Groove Pattern Similarity

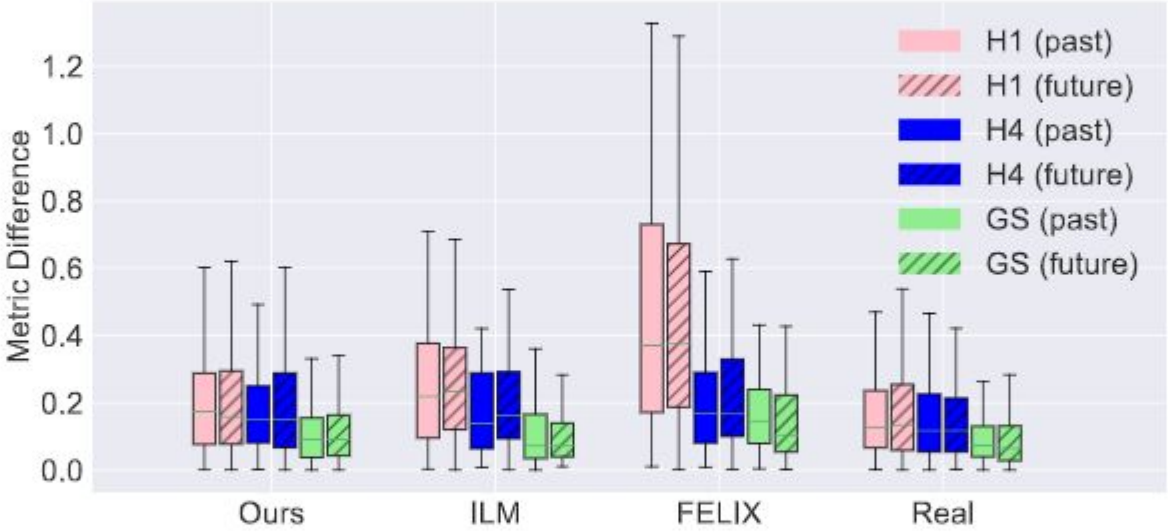
$$\mathcal{GS}(\overrightarrow{g}^a, \overrightarrow{g}^b) = 1 - \frac{1}{T}\sum_{t=0}^{T-1} XOR(g_t^a, g_t^b)$$

Groove Similarity Divergence

$$GS_{div}(Y_{true}||Y_{pred}) = JS_{div}(\overrightarrow{h}_{GS(Y_{true})}||\overrightarrow{h}_{GS(Y_{pred})})$$

Divergence Metrics

- Inspired by VLI evaluation methodology.
  - Comparison of distributions, although it is visually.
  - We want to formalize this intuition numerically.



Pitch Accuracy

$$y_{true} = [C_4, \_, D_4, \_, E_4, \_, F_4, \_]$$

Pitch Accuracy

$$y_{true} = [C_4, \_, D_4, \_, E_4, \_, F_4, \_]$$

$$y_0 = [C_4, \_, D_4, \_, E_4, \_, F\#_4, \_]$$

Pitch Accuracy

$y_{true} = [C_4, \_, D_4, \_, E_4, \_, F_4, \_]$

$y_0 = [C_4, \_, D_4, \_, E_4, \_, F\#_4, \_]$

$$pAcc(y_{true}, y_0) = 3/4$$

Pitch Accuracy

$$y_{true} = [C_4, \_, D_4, \_, E_4, \_, F_4, \_]$$

$$y_1 = [C_4, \_, D_4, \_, E_4, \_, \_, F_4]$$

Pitch Accuracy

$$y_{true} = [C_4, \_, D_4, \_, E_4, \_, F_4, \_]$$

$$y_1 = [C_4, \_, D_4, \_, E_4, \_, \_, F_4]$$

$$pAcc(y_{true}, y_1) = 3/4$$

Rhythm Accuracy

True data

[60, 128, 128, 128, 62, 128, 128, 128]

Pred data

[60, 128, 128, 128, 62, 128, 64, 128]

pAcc = 2 / (2 + 1) = 0.67

rAcc = 5 / (5 + 1) = 0.83

[60, 128, 128, 128, 128, 128, 62, 128, 128, 128, 128, 128]

[60, 128, 128, 128, 128, 128, 62, 128, 128, 64, 128, 128]

pAcc = 2 / (2 + 1) = 0.67

rAcc = 9 / (9 + 1) = 0.90

Position F1

- We propose a new metric to deal with these issues: Position Score
- Pros:
  - Disambiguates Pitch Accuracy
  - Standardize Rhythm Accuracy

Silence Density

$$x = [D_3, \_, \_, \_, E_3, \_, \_, \_, \times, \times, \times, \times, A_3, \_, \_, \_]$$

$$S(x) = 0.25$$

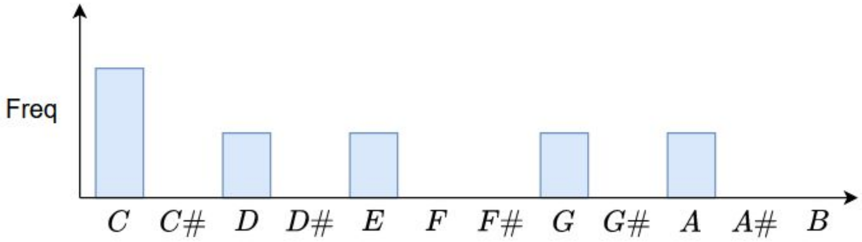
Pitch Class Entropy

$$x = [C_4, \_, D_4, \_, C_4, \_, E_4, \_, G_4, \_, \_, \_, A_4, \_, \_, \_]$$

Pitch Class Entropy

$$x = [C_4, \_, D_4, \_, C_4, \_, E_4, \_, G_4, \_, \_, \_, A_4, \_, \_, \_]$$

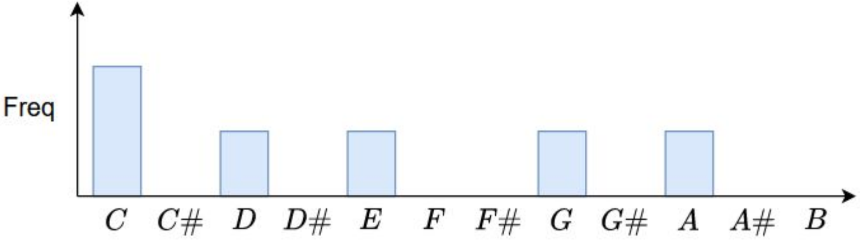
$$\overrightarrow{h}_{pitch(m_i)} =$$



Pitch Class Entropy

$$x = [C_4, \_, D_4, \_, C_4, \_, E_4, \_, G_4, \_, \_, \_, A_4, \_, \_, \_]$$

$$\vec{h}_{pitch(m_i)} =$$



$$\mathcal{H}_{m_i} = \mathcal{H}(\vec{h}_{pitch(m_i)}) = -\sum_{i=0}^{11} h_i \log_2(h_i)$$

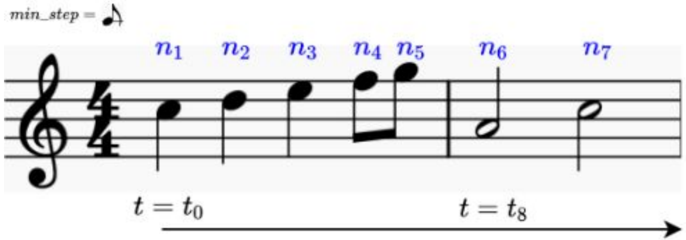
Groove Pattern Similarity

$$x = [1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0]$$

$$y = [1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0]$$

$$\mathcal{GS}(x, y) = 14/16$$

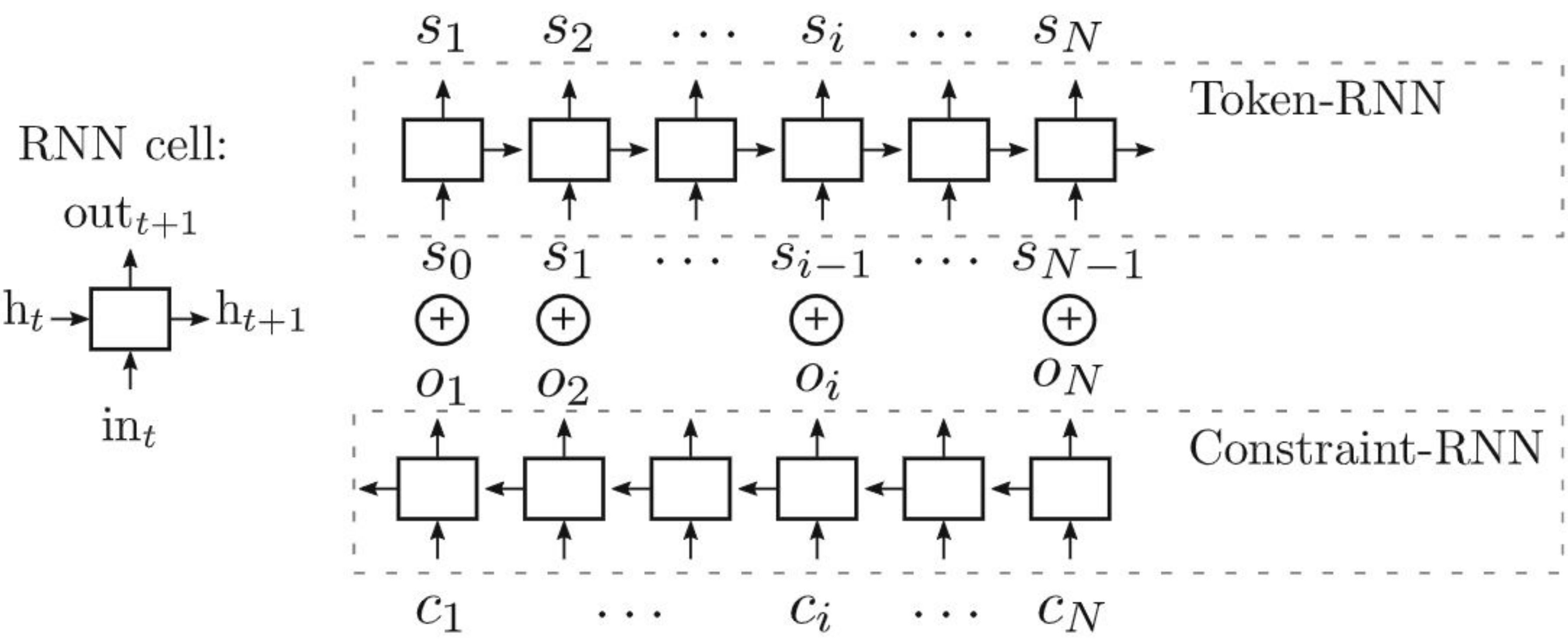
Representation - REMI



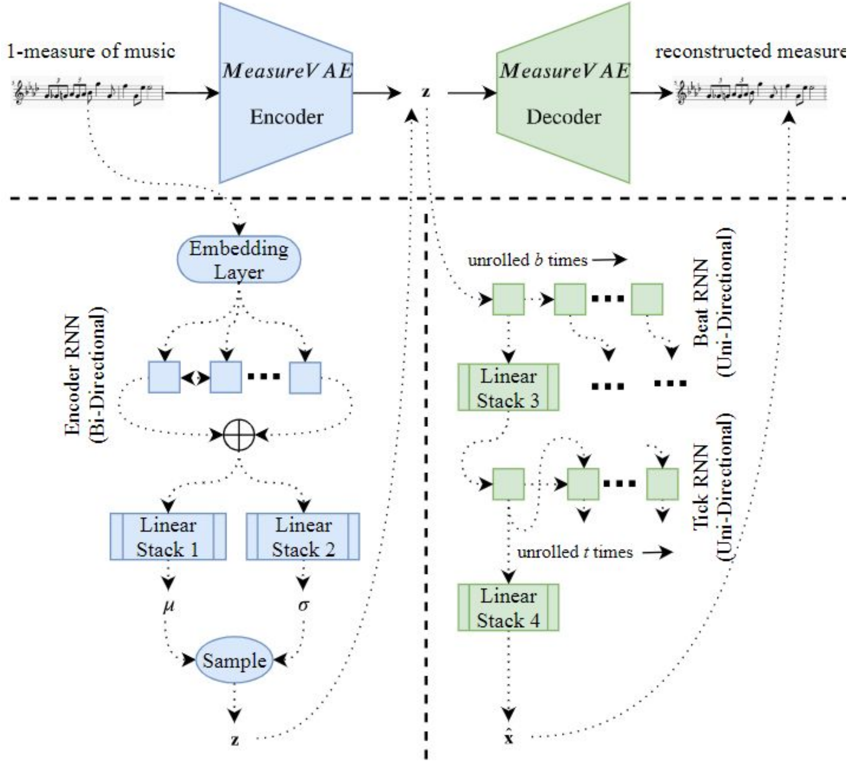
$$x = [\ [n_1, n_2, n_3, n_4, n_5], [n_6, n_7]\ ]$$

n	Tempo	Bar Start	Position	Pitch	Velocity	Duration
n_1	120	1	0 | 8	C_4	90	♩
n_2	120	0	2 | 8	D_4	90	♩
n_3	120	0	4 | 8	E_4	90	♩
n_4	120	0	6 | 8	F_4	90	♪
n_5	120	0	7 | 8	G_4	90	♪
n_6	120	1	0 | 8	A_3	90	𝅗𝅥
n_7	120	0	4 | 8	C_4	90	𝅗𝅥

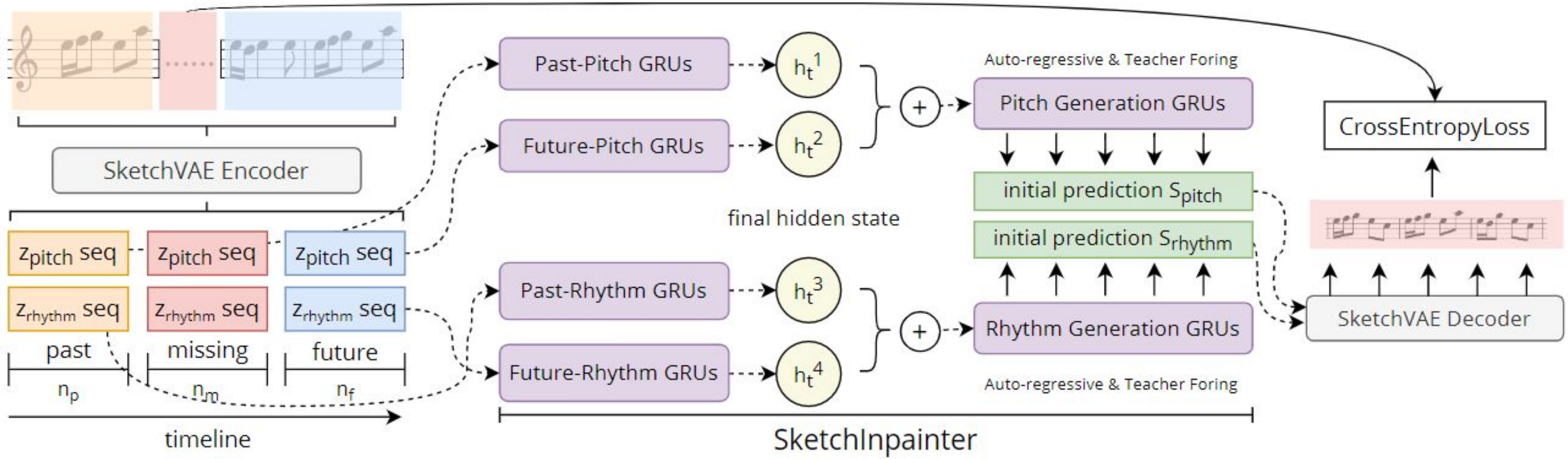
Models - Anticipation-RNN



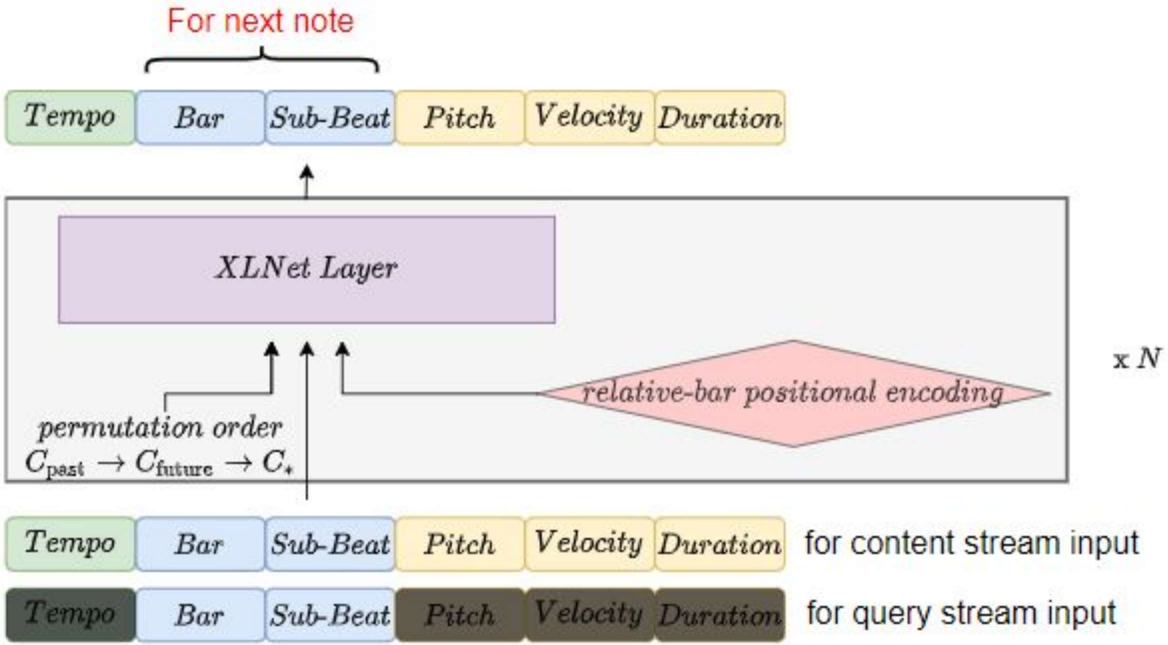
Models - Music InpaintNet



Models - Music SketchNet



Models - VLI



For next note
Tempo
Bar
Sub-Beat
Pitch
Velocity
Duration
XLNet Layer
permutation order
$C_{\text{past}} \rightarrow C_{\text{future}} \rightarrow C_*$
relative-bar positional encoding
x N
Tempo
Bar
Sub-Beat
Pitch
Velocity
Duration
for content stream input
Tempo
Bar
Sub-Beat
Pitch
Velocity
Duration
for query stream input