TOWARDS MULTI-INSTRUMENT DRUM TRANSCRIPTION

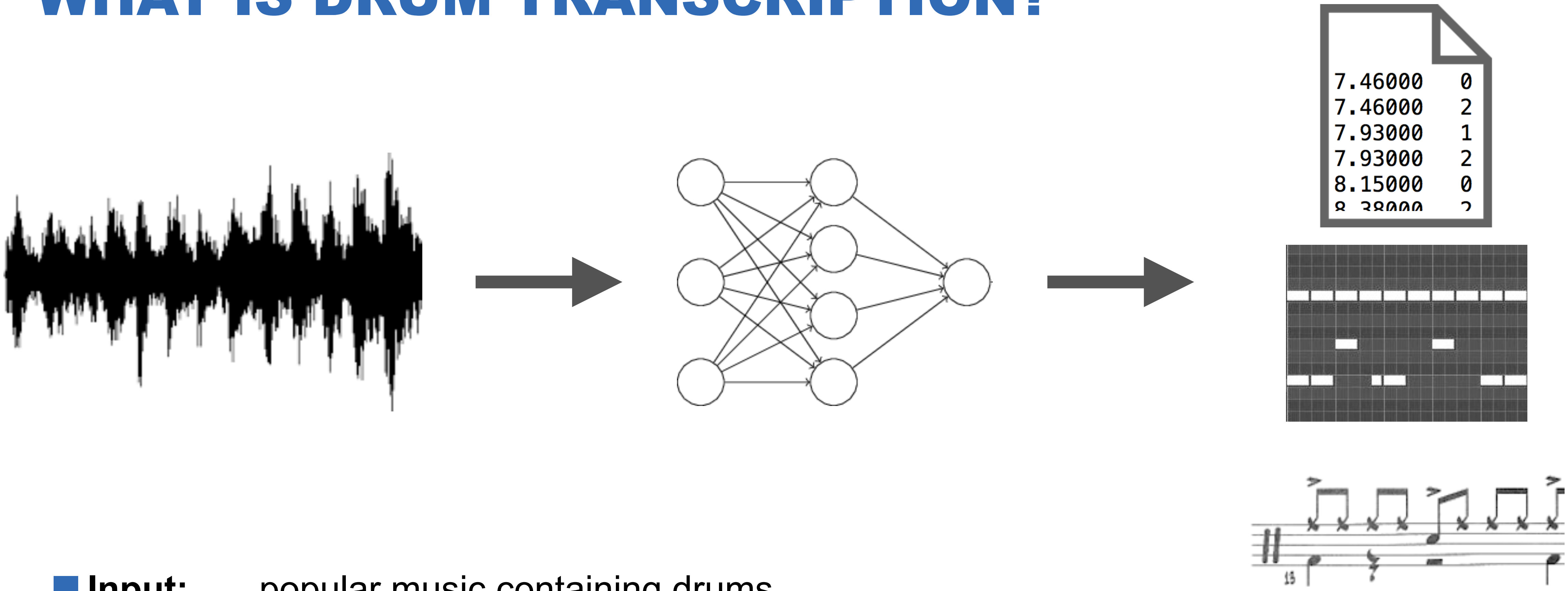
Richard Vogl[1,2], Gerhard Widmer[2], Peter Knees[1]

richard.vogl@tuwien.ac.at, gerhard.widmer@jku.at, peter.knees@tuwien.ac.at

[1] TU WIEN TECHNISCHE UNIVERSITÄT WIEN ifs mir group

[2] JKU JOHANNES KEPLER UNIVERSITÄT LINZ Department of Computational Perception

WHAT IS DRUM TRANSCRIPTION?

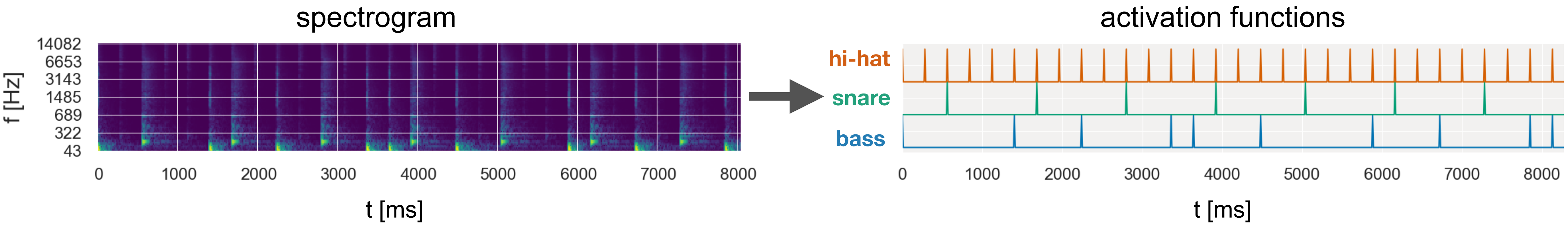


```
7.46000   0
7.46000   2
7.93000   1
7.93000   2
8.15000   0
8.38000   2
```

- **Input:** popular music containing drums
- **Output:** symbolic representation of notes played by drum instruments

STATE OF THE ART

- Current state-of-the-art systems:
  - End-to-end / **activation-function-based** approaches
  - **NN** based approaches and **NMF** approaches


spectrogram

activation functions

- Overview Article

  Wu, C.-W., Dittmar, C., Southall, C.,Vogl, R., Widmer, G., Hockman, J., Müller, M., Lerch, A.:
  “**An Overview of Automatic Drum Transcription**,” IEEE TASLP, vol. 26, no. 9, Sept. 2018.

FOCUS OF THIS WORK

FOCUS OF THIS WORK

- SotA works focus bass drum (**BD**) snare (**SD**) and hi-hat (**HH**)

FOCUS OF THIS WORK

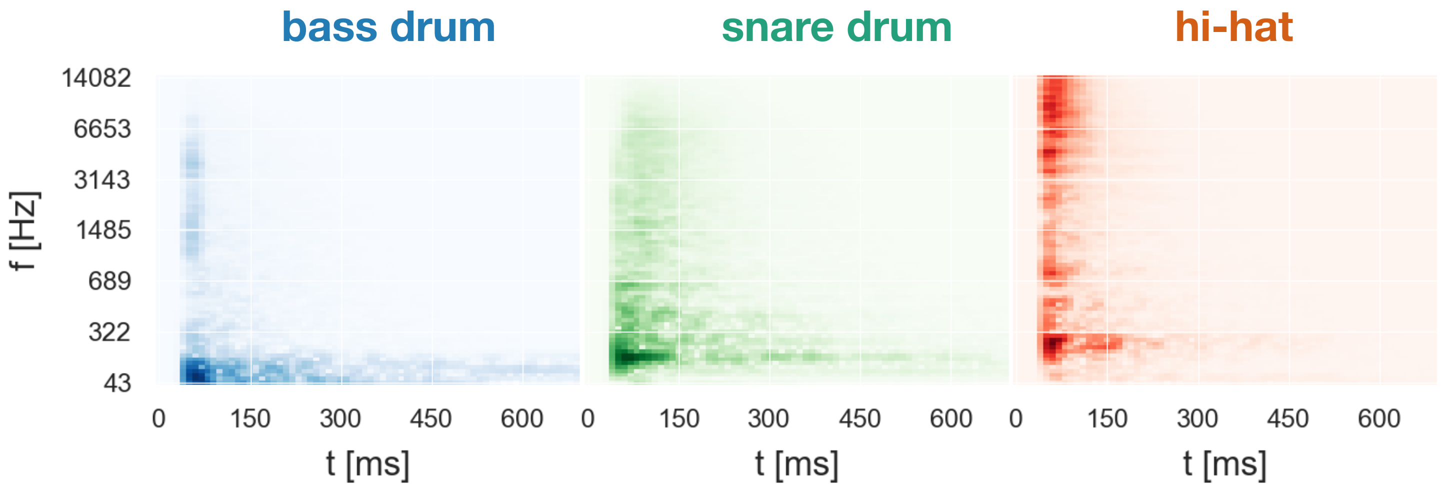
- SotA works focus bass drum (**BD**) snare (**SD**) and hi-hat (**HH**)
  - Make up **majority of notes** in datasets

FOCUS OF THIS WORK

- SotA works focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**

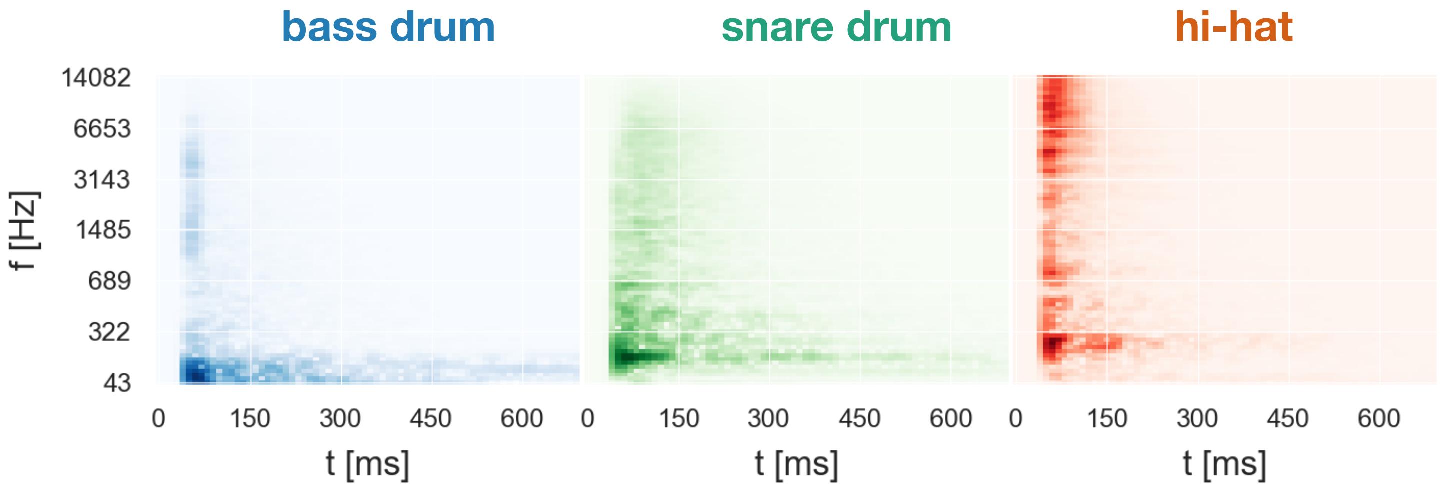
FOCUS OF THIS WORK

- SotA works focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy** distribution



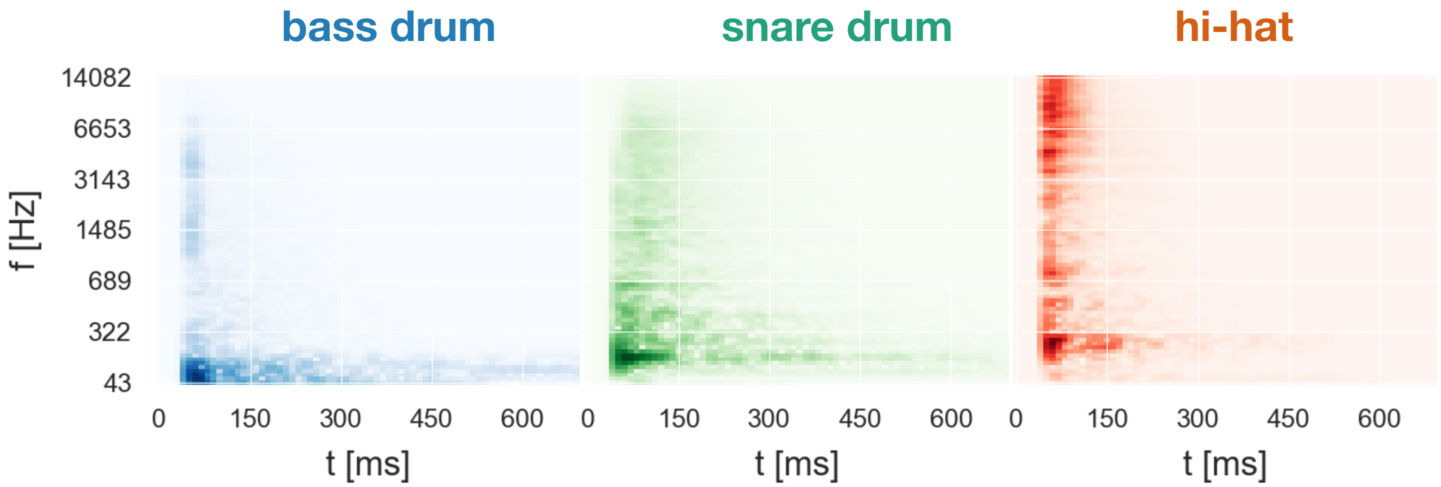
FOCUS OF THIS WORK

- SotA works focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy** distribution



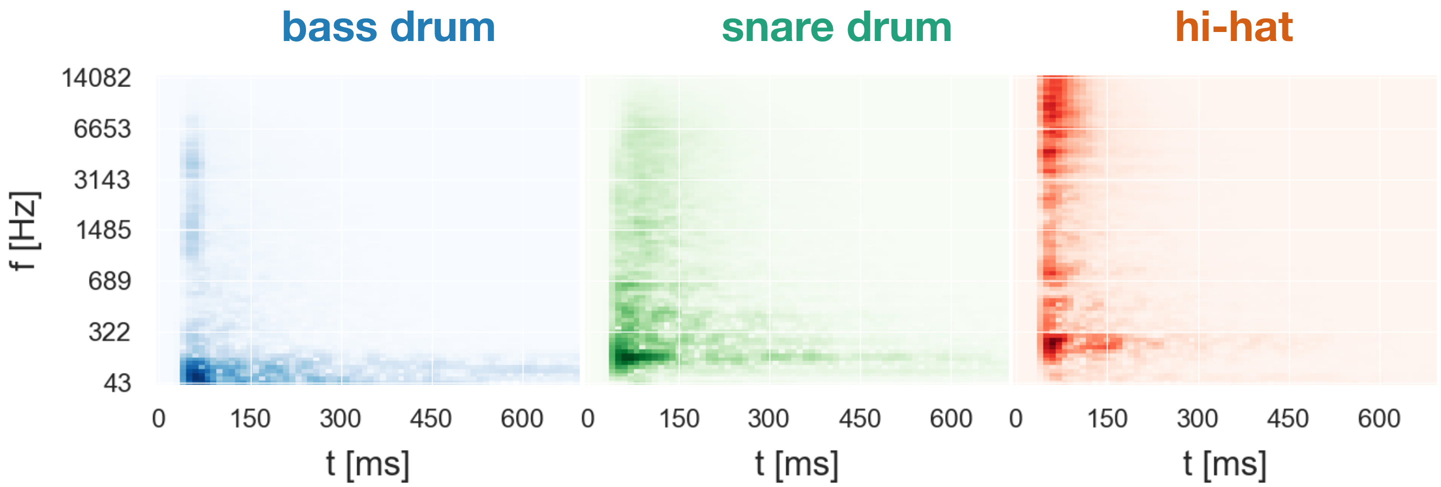
FOCUS OF THIS WORK

- SotA works focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy distribution**



FOCUS OF THIS WORK

- SotA works focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy distribution**

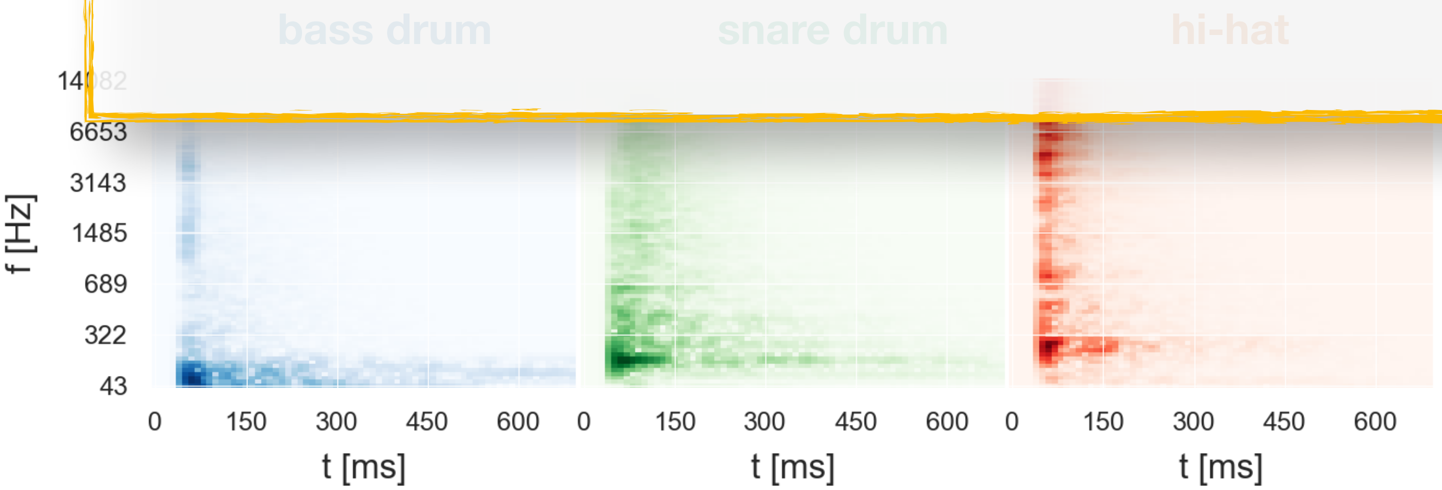


FOCUS OF THIS WORK

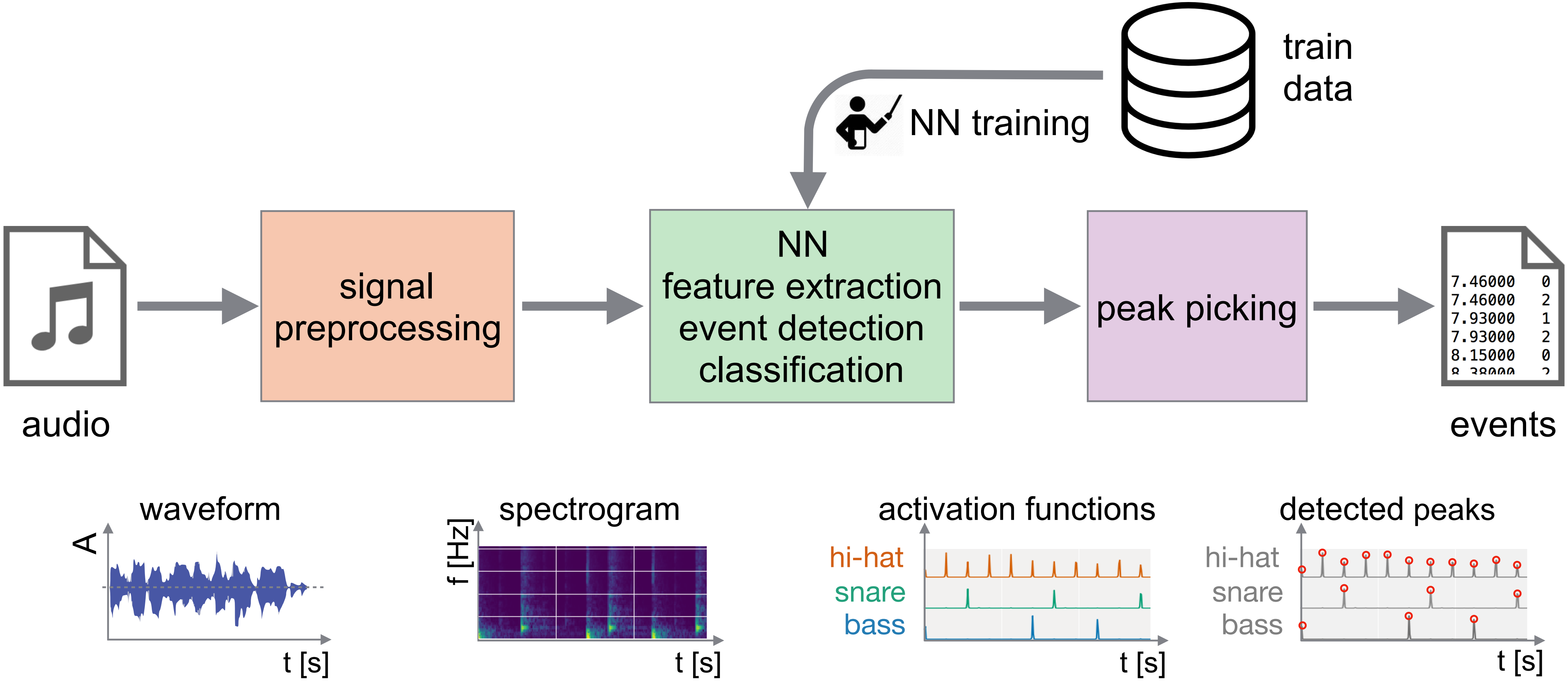
- SotA works focus bass drum (**BD**) snare (**SD**) and hi-hat (**HH**)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy** distribution

Other instruments are important!

→ Increase number of instruments for drum transcription



SYSTEM OVERVIEW

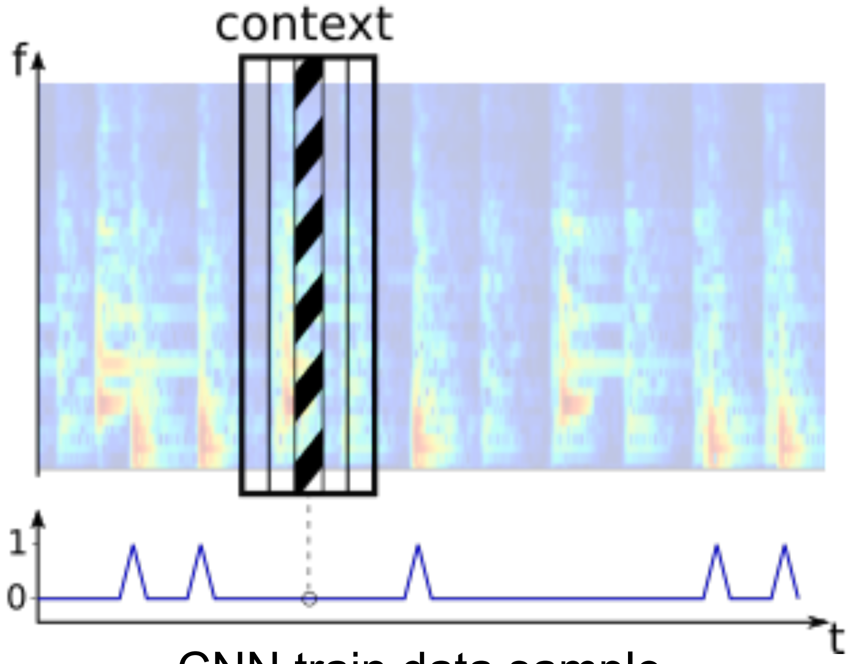


```
7.46000   0
7.46000   2
7.93000   1
7.93000   2
8.15000   0
8.38000   2
```

NETWORK ARCHITECTURES

NETWORK ARCHITECTURES

- Convolutional NN (**CNN**)
  - Convolutions capture **local correlations**
  - **Acoustic modeling** of drum sounds

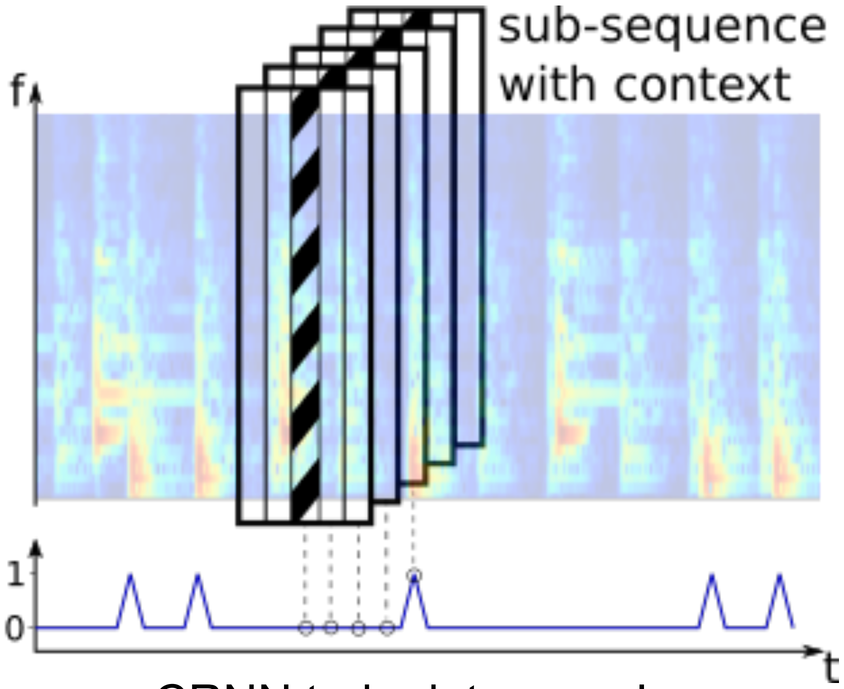


CNN train data sample

NETWORK ARCHITECTURES

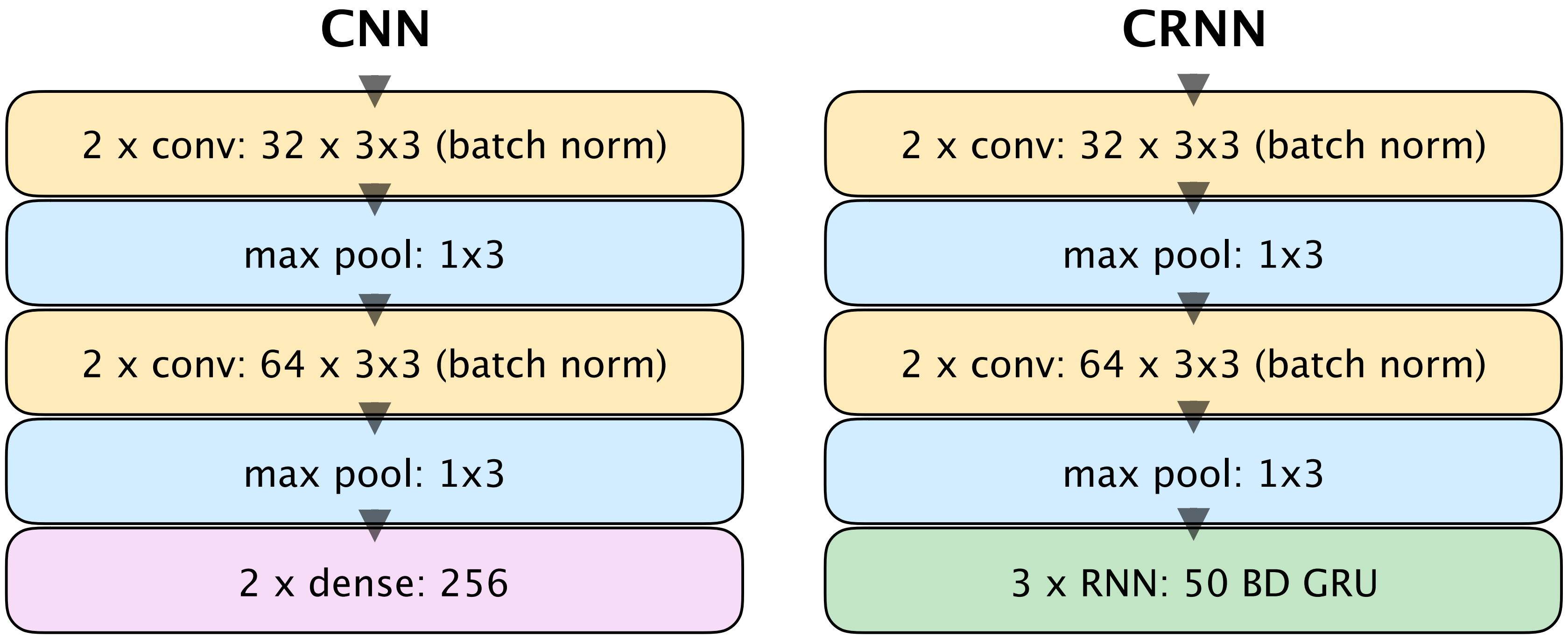
- Convolutional NN (**CNN**)
  - Convolutions capture **local correlations**
  - **Acoustic modeling** of drum sounds

- Convolutional RNN (**CRNN**)
  - **”best of both worlds”**
  - Low-level CNN for **acoustic modeling**
  - Higher-level RNN for **repetitive pattern modeling**



CRNN train data sample

NETWORK ARCHITECTURES



CNN

2 x conv: 32 x 3x3 (batch norm)

max pool: 1x3

2 x conv: 64 x 3x3 (batch norm)

max pool: 1x3

2 x dense: 256

CRNN

2 x conv: 32 x 3x3 (batch norm)

max pool: 1x3

2 x conv: 64 x 3x3 (batch norm)

max pool: 1x3

3 x RNN: 50 BD GRU

- Early stopping
- Batch normalization
- L2 norm
- Dropout (30%)
- ADAM optimizer

	frames	context	conv. layers	rec. layers	dense layers
CNN	—	25	see figure	—	2x256
CRNN	400	13		3 x 50 BD GRU	—

DATASETS

DATASETS

- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers / drum kits
  - 64 tracks, total duration: **1h**

♫

DATASETS

- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers / drum kits
  - 64 tracks, total duration: **1h**

♫

DATASETS

- **ENST-Drums** [Gillet and Richard 2006] ♫
  - Recordings, three drummers / drum kits
  - 64 tracks, total duration: **1h**

- **MDB Drums** [Southall et al. 2017] ♫
  - Drum annotations for Medley DB subset
  - 23 tracks, total duration: **20m**

DATASETS

- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers / drum kits
  - 64 tracks, total duration: **1h**

- **MDB Drums** [Southall et al. 2017]
  - Drum annotations for Medley DB subset
  - 23 tracks, total duration: **20m**

DATASETS

- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers / drum kits
  - 64 tracks, total duration: **1h**

♫

- **MDB Drums** [Southall et al. 2017]
  - Drum annotations for Medley DB subset
  - 23 tracks, total duration: **20m**

♫

- **RBMA13-Drums** [Vogl et al. 2017]
  - Music from 2013 Red Bull Music Academy, different styles
  - 27 tracks, total duration: **1h 43m**

♫

DATASETS

- **ENST-Drums** [Gillet and Richard 2006] ♫
  - Recordings, three drummers / drum kits
  - 64 tracks, total duration: **1h**

- **MDB Drums** [Southall et al. 2017] ♫
  - Drum annotations for Medley DB subset
  - 23 tracks, total duration: **20m**

- **RBMA13-Drums** [Vogl et al. 2017] ♫
  - Music from 2013 Red Bull Music Academy, different styles
  - 27 tracks, total duration: **1h 43m**

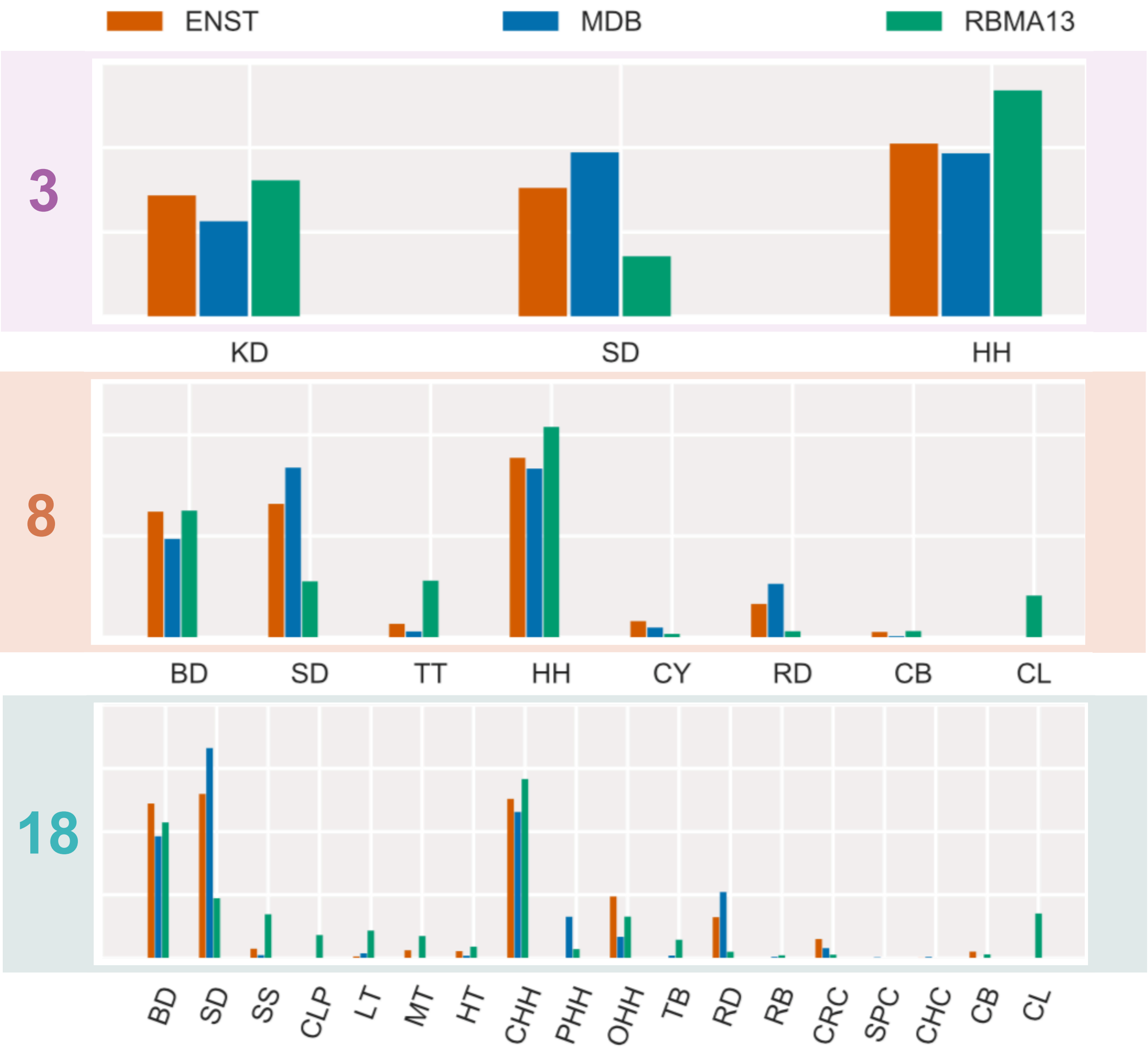
DATASETS

number of classes			
3	8	18	instrument name
BD	BD	BD	bass drum
SD	SD	SD	snare drum
		SS	side stick
		CLP	hand clap
	TT	HT	hight tom
		MT	mid tom
		LT	low tom
HH	HH	CHH	closed hi-hat
		PHH	pedal hi-hat
		OHH	open hi-hat
		TB	tambourine
	RD	RD	ride cymbal
	BE	RB	ride bell
		CB	cowbell
	CY	CRC	crash cymbal
		SPC	splash cymbal
		CHC	Chinese cymbal
	CL	CL	clave/sticks

DATASETS

number of classes			
3	8	18	instrument name
BD	BD	BD	bass drum
SD	SD	SD	snare drum
		SS	side stick
		CLP	hand clap
	TT	HT	hight tom
		MT	mid tom
		LT	low tom
HH	HH	CHH	closed hi-hat
		PHH	pedal hi-hat
		OHH	open hi-hat
		TB	tambourine
	RD	RD	ride cymbal
	BE	RB	ride bell
		CB	cowbell
	CY	CRC	crash cymbal
		SPC	splash cymbal
		CHC	Chinese cymbal
	CL	CL	clave/sticks

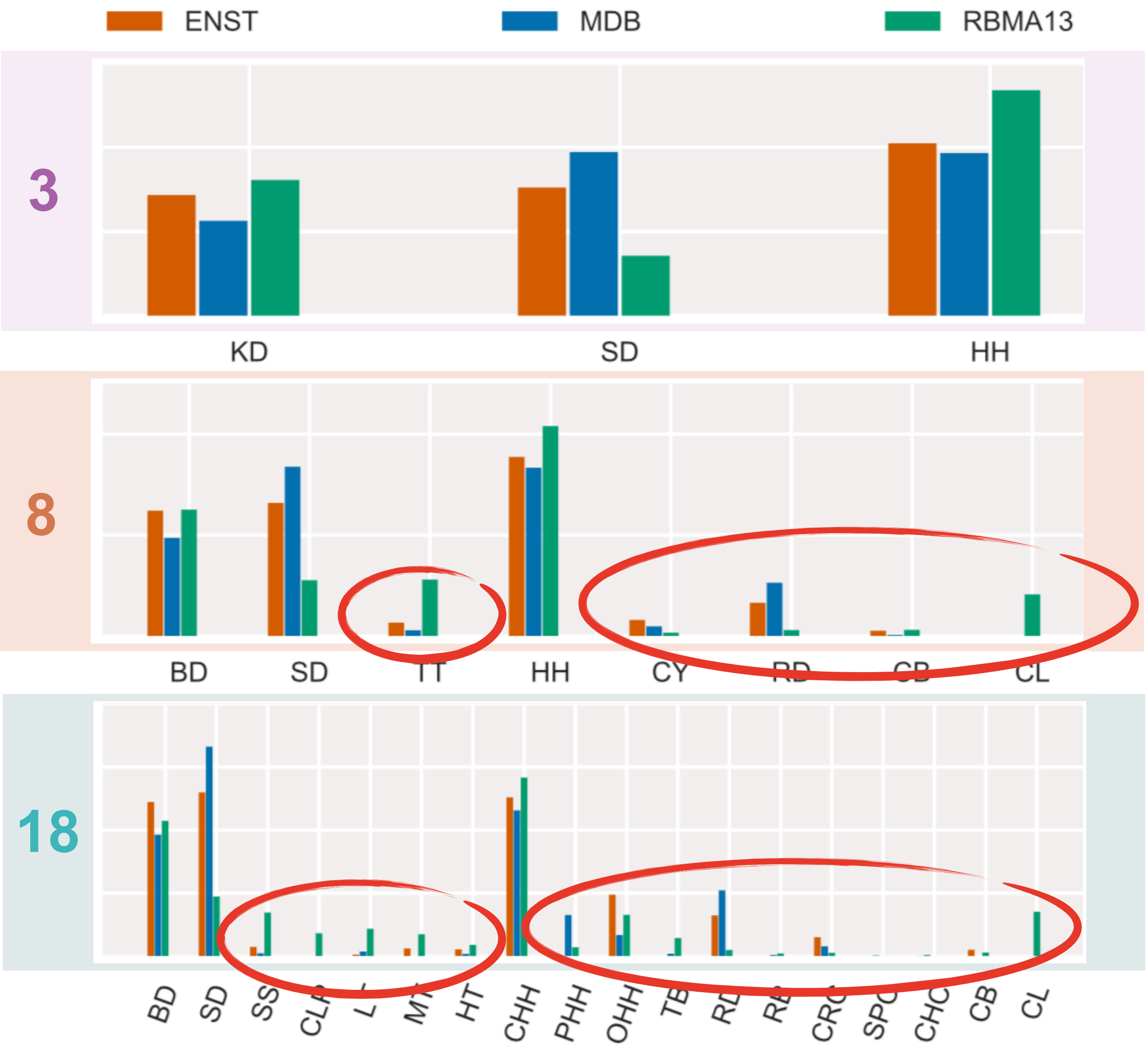
relative frequency of instrument onsets



DATASETS

number of classes			
3	8	18	instrument name
BD	BD	BD	bass drum
SD	SD	SD	snare drum
		SS	side stick
		CLP	hand clap
	TT	HT	hight tom
		MT	mid tom
		LT	low tom
HH	HH	CHH	closed hi-hat
		PHH	pedal hi-hat
		OHH	open hi-hat
		TB	tambourine
	RD	RD	ride cymbal
	BE	RB	ride bell
		CB	cowbell
	CY	CRC	crash cymbal
		SPC	splash cymbal
		CHC	Chinese cymbal
	CL	CL	clave/sticks

relative frequency of instrument onsets



SYNTHETIC DATASET

NEW!

SYNTHETIC DATASET

NEW!

Synthetic dataset from **MIDI** songs

- Mix of different genres, **full songs**

SYNTHETIC DATASET

NEW!

Synthetic dataset from **MIDI** songs

- Mix of different genres, **full songs**
- Optional **accompaniment**

SYNTHETIC DATASET

NEW!

Synthetic dataset from **MIDI** songs

- Mix of different genres, **full songs**
- Optional **accompaniment**
- Diverse drum sounds (**57 different drum kits**, acoustic and electronic)

SYNTHETIC DATASET

NEW!

Synthetic dataset from **MIDI** songs

- Mix of different genres, **full songs**
- Optional **accompaniment**
- Diverse drum sounds (**57 different drum kits**, acoustic and electronic)
- Varying quality, **no vocals**!

SYNTHETIC DATASET

NEW!

Synthetic dataset from **MIDI** songs

- Mix of different genres, **full songs**
- Optional **accompaniment**
- Diverse drum sounds (**57 different drum kits**, acoustic and electronic)
- Varying quality, **no vocals**!
- 4197 tracks, total duration: **259h**

♫

SYNTHETIC DATASET

NEW!

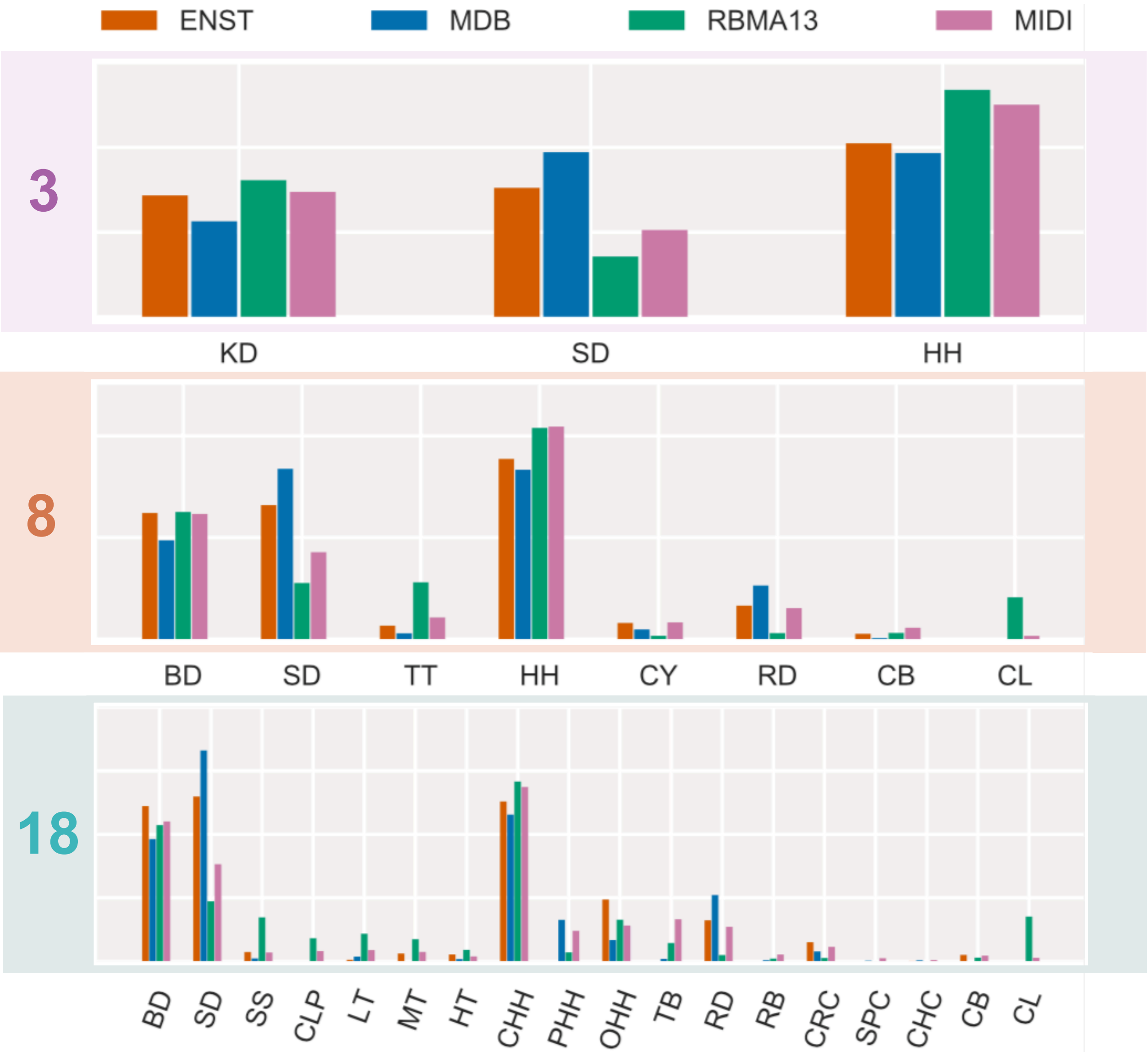
Synthetic dataset from **MIDI** songs

- Mix of different genres, **full songs**
- Optional **accompaniment**
- Diverse drum sounds (**57 different drum kits**, acoustic and electronic)
- Varying quality, **no vocals**!
- 4197 tracks, total duration: **259h**

♫

SYNTHETIC DATASET

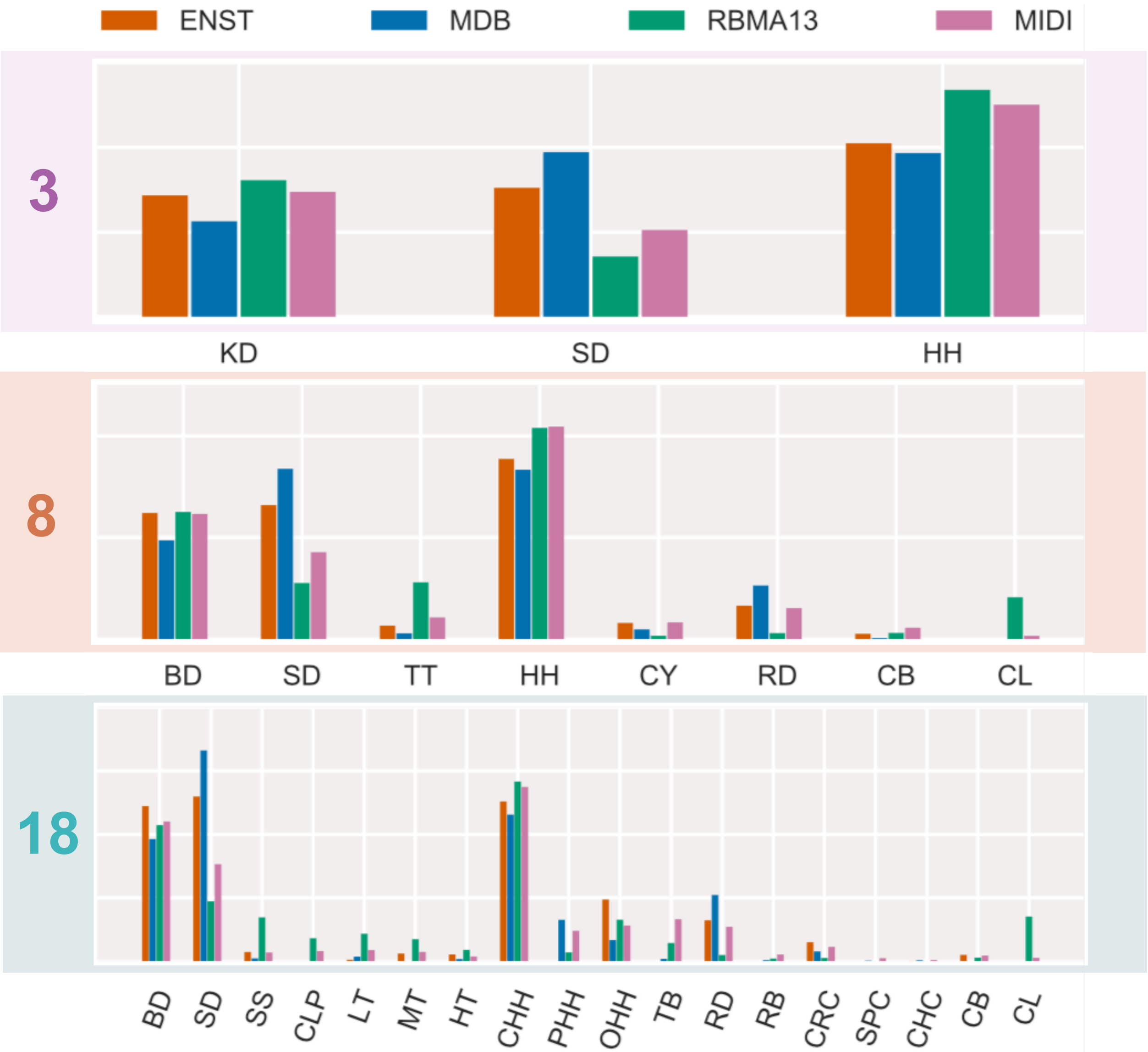
relative frequency of instrument onsets



SYNTHETIC DATASET

- Follows the **same relative instrument distribution**

relative frequency of instrument onsets

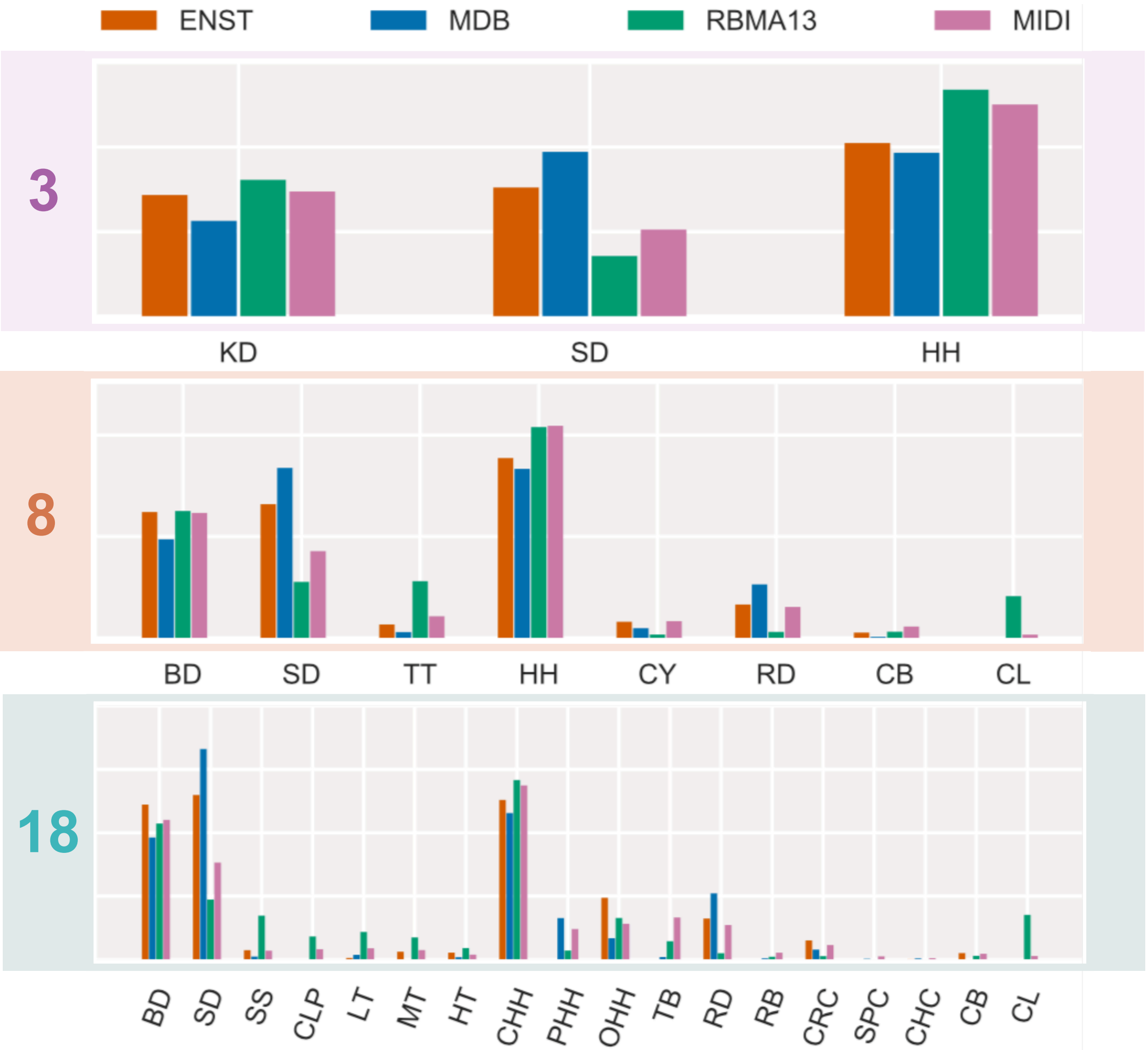


SYNTHETIC DATASET

- Follows the **same relative instrument distribution**
  - same bias for instruments

    same problems during training

relative frequency of instrument onsets

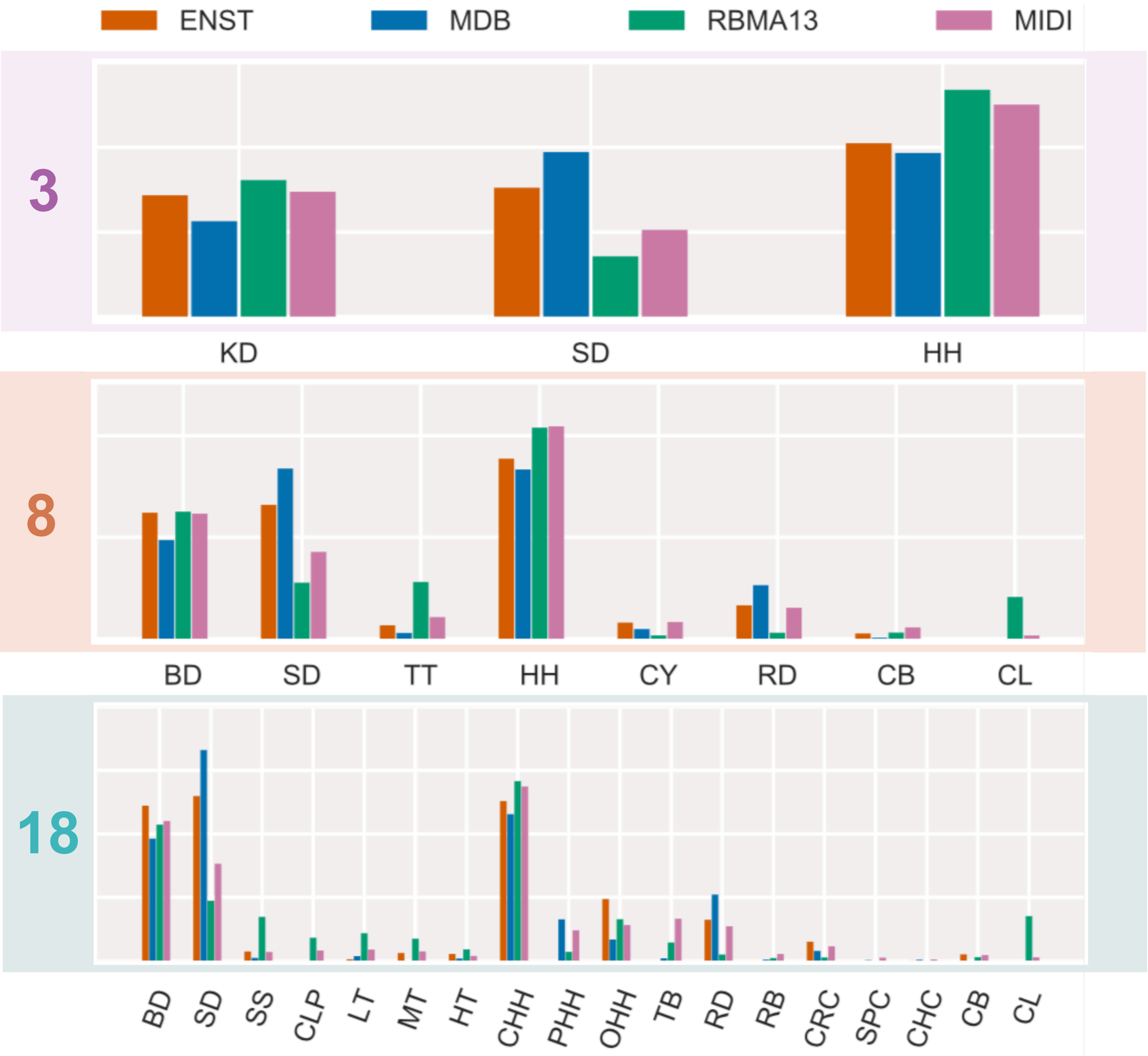


SYNTHETIC DATASET

- Follows the **same relative instrument distribution**
  - – same bias for instruments

    same problems during training
  - + datasets are **representative samples**

relative frequency of instrument onsets



BALANCING OF SYNTHETIC DATASET

BALANCING OF SYNTHETIC DATASET

- **Swap instruments** for individual tracks

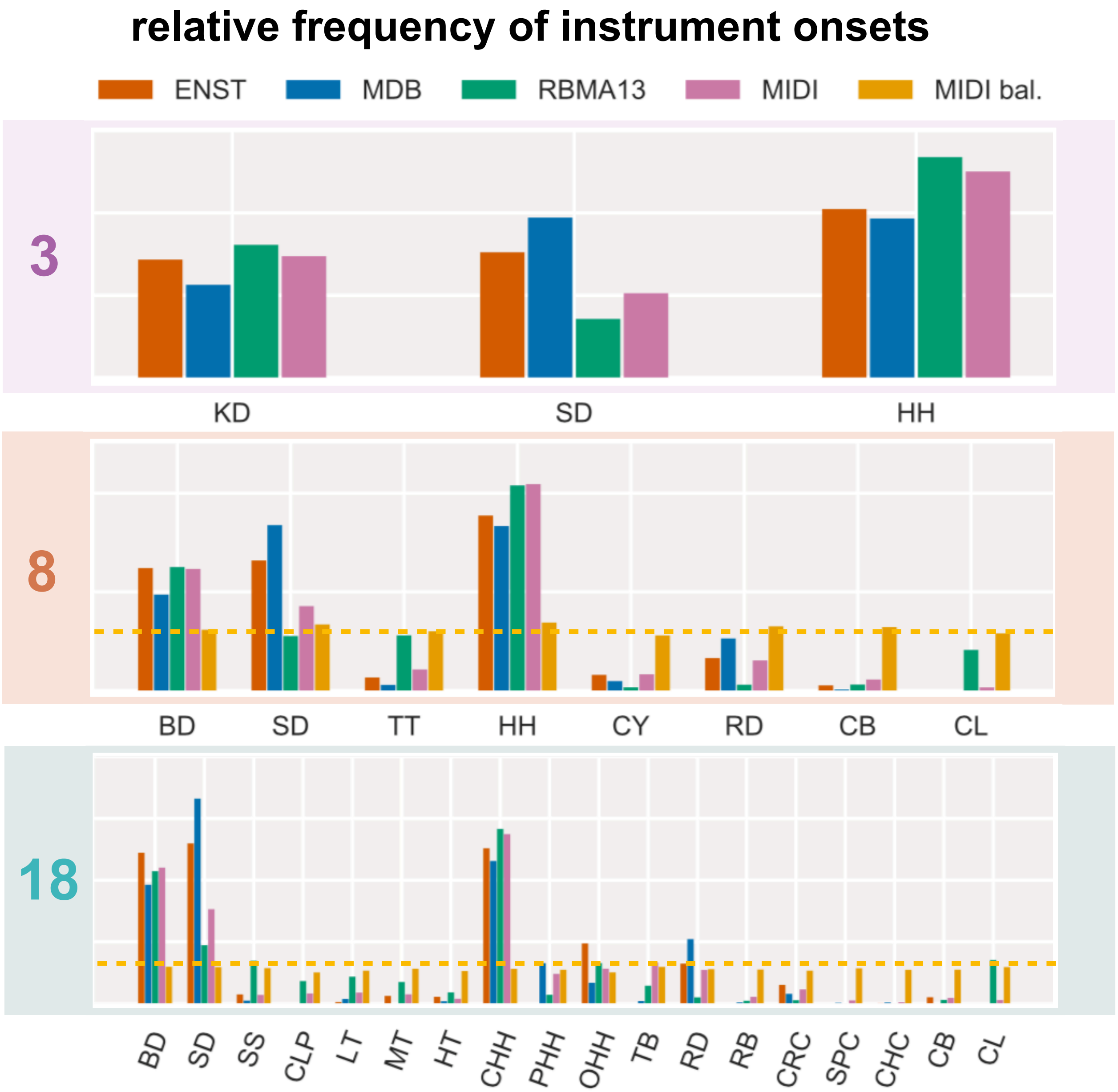
BALANCING OF SYNTHETIC DATASET

- **Swap instruments** for individual tracks
- **Artificial balancing** of instrument distribution

BALANCING OF SYNTHETIC DATASET

- **Swap instruments** for individual tracks
- **Artificial balancing** of instrument distribution

♫



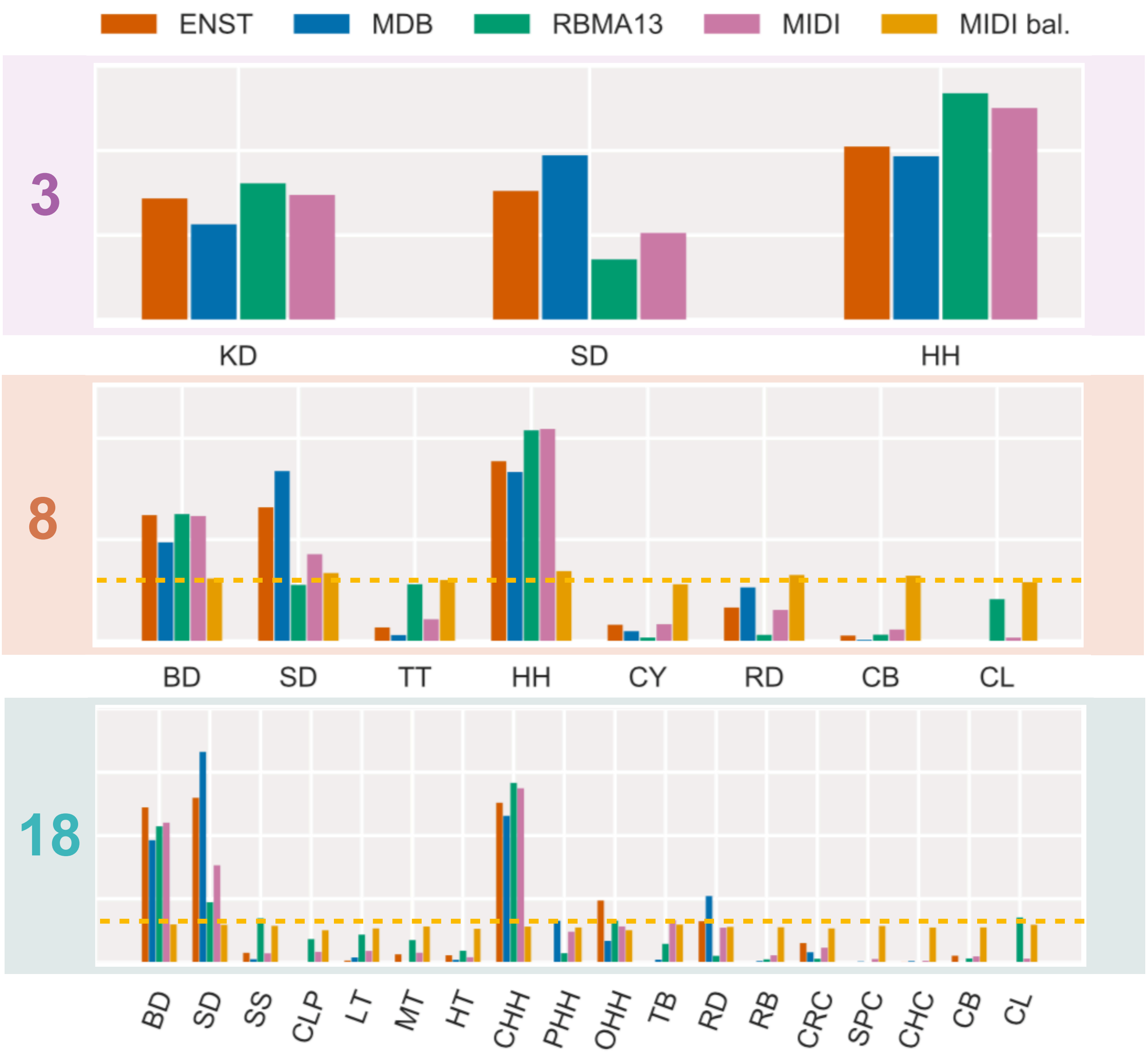
relative frequency of instrument onsets

BALANCING OF SYNTHETIC DATASET

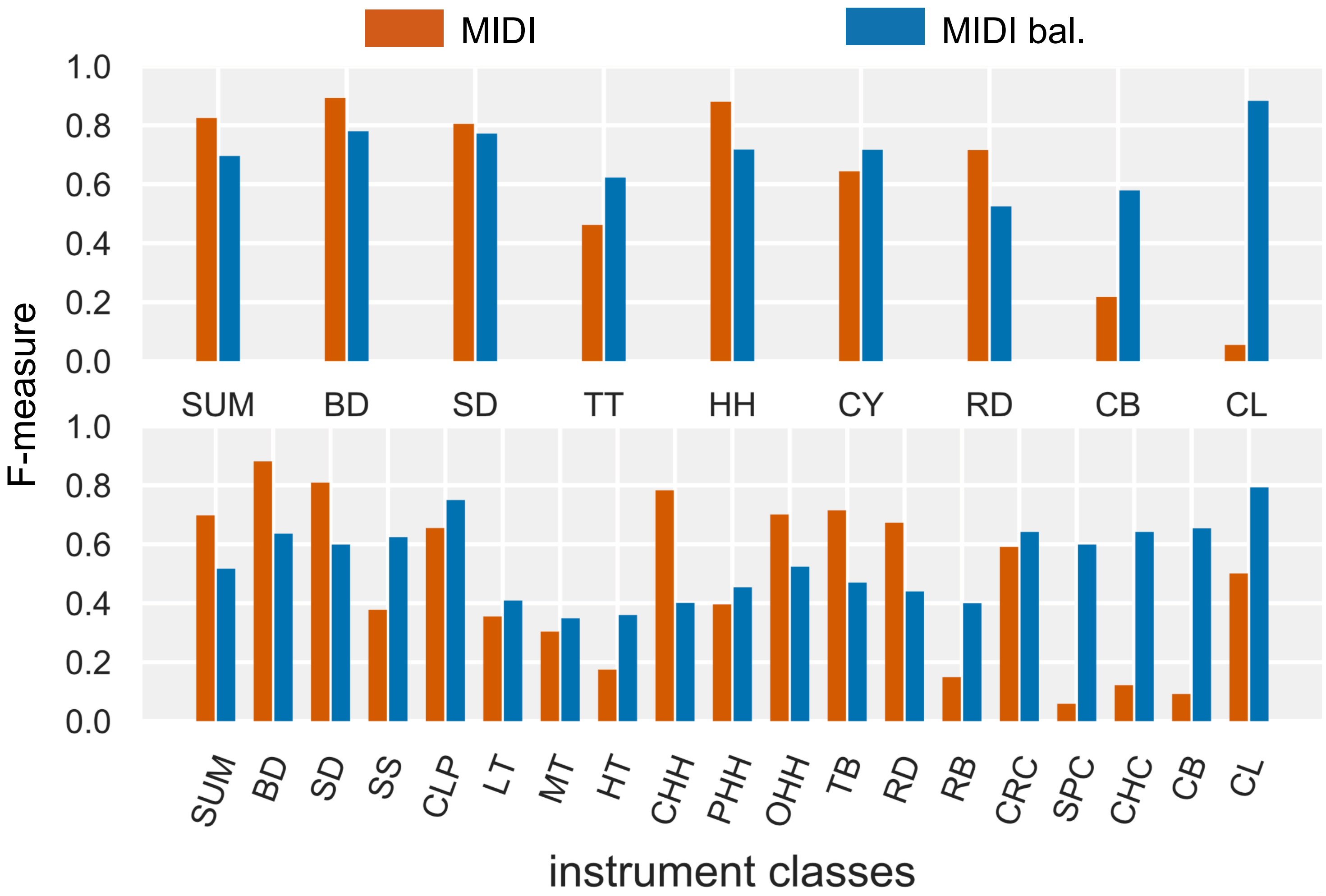
- **Swap instruments** for individual tracks
- **Artificial balancing** of instrument distribution

♫

relative frequency of instrument onsets

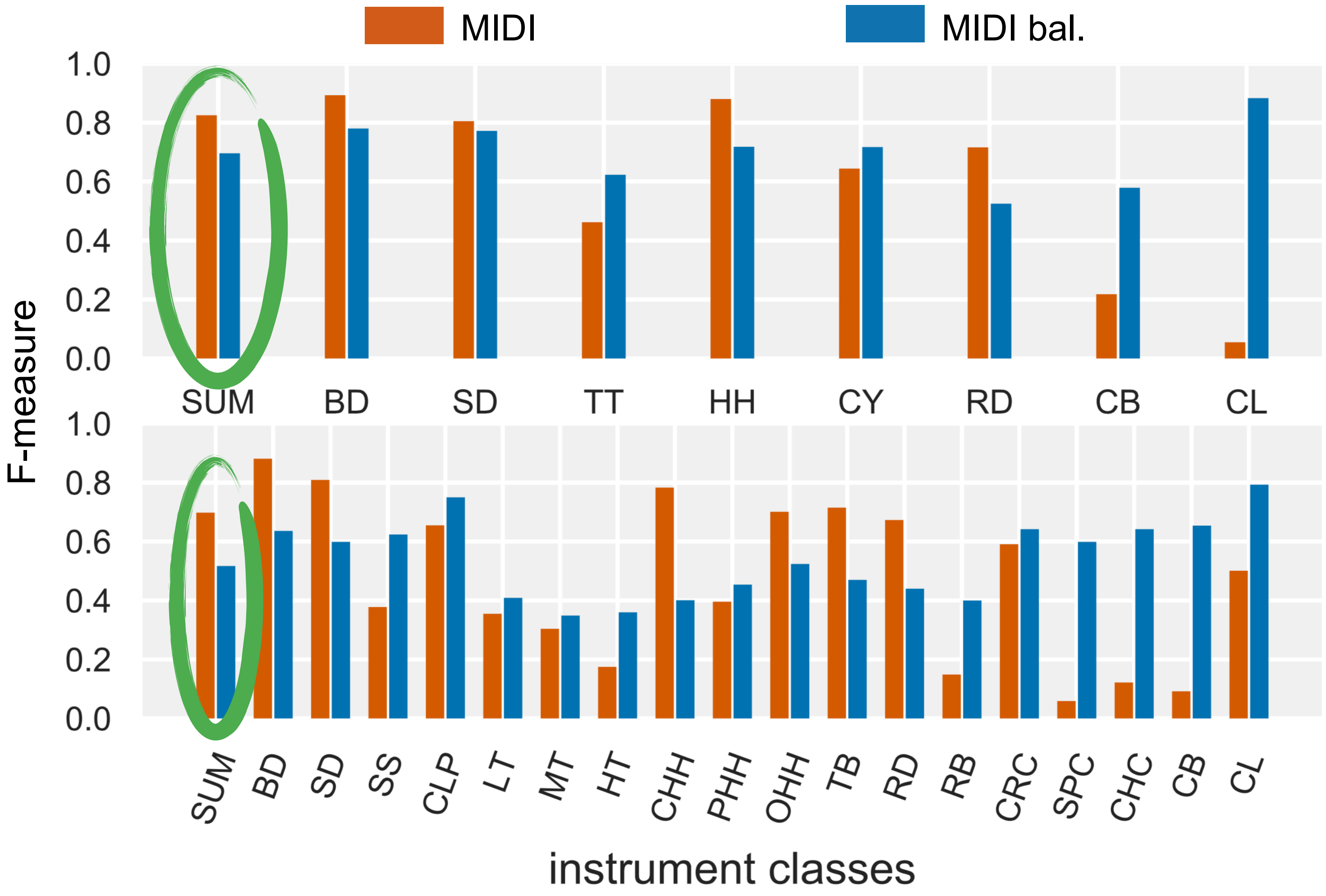


RESULTS ON SYNTHETIC DATA



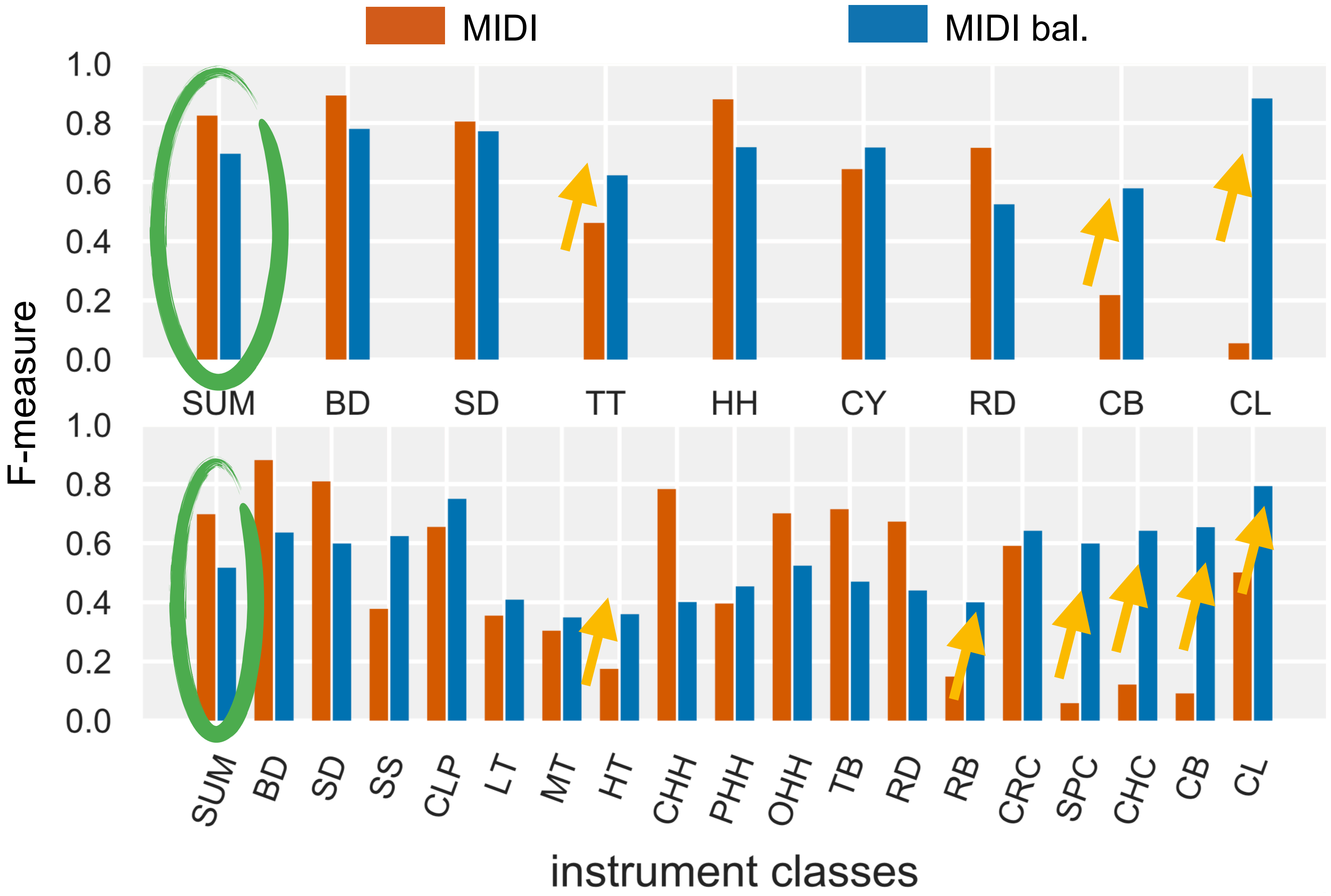
RESULTS ON SYNTHETIC DATA

- **Overall performance** for MIDI bal. **is worse**
  - It is a harder task



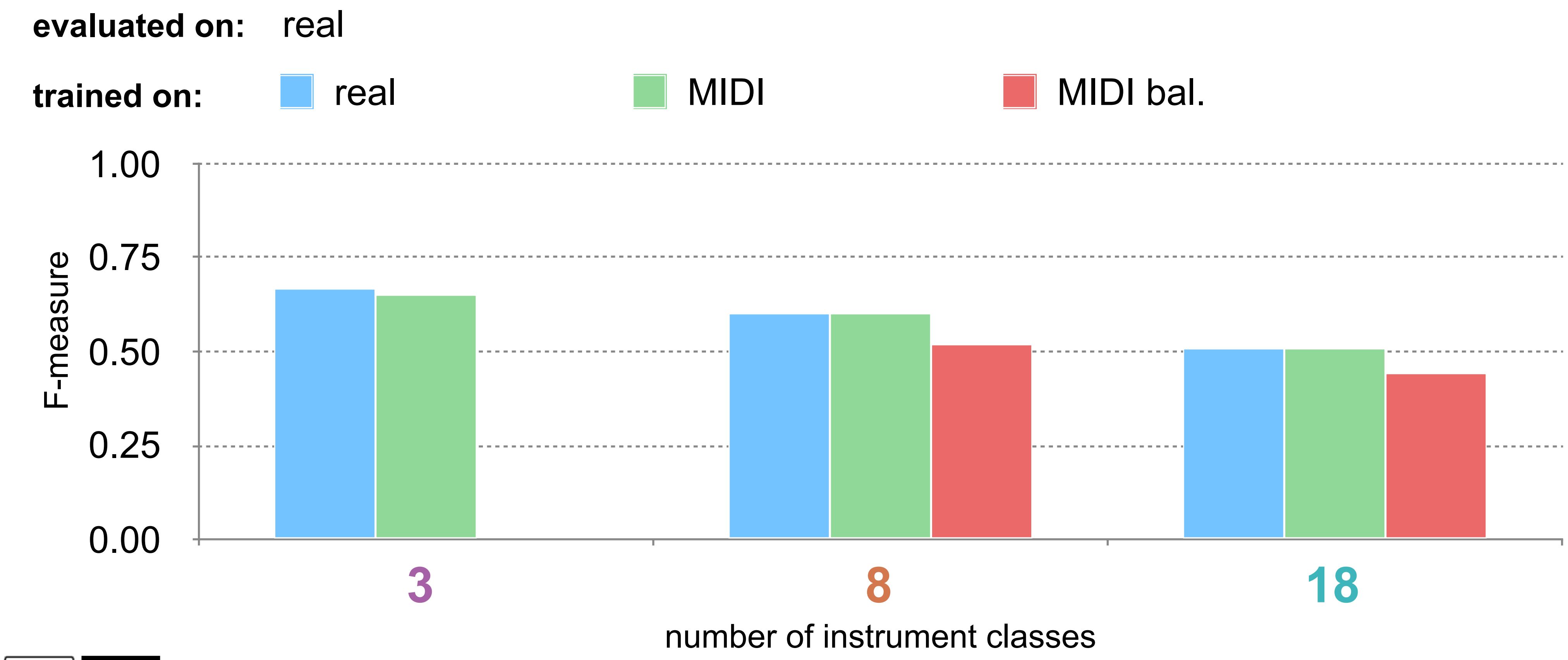
RESULTS ON SYNTHETIC DATA

- **Overall performance** for MIDI bal. **is worse**
  - It is a harder task
- Performance of **underrepresented instruments improves**
  - Providing more samples forces the network to learn formerly sparsely used instruments



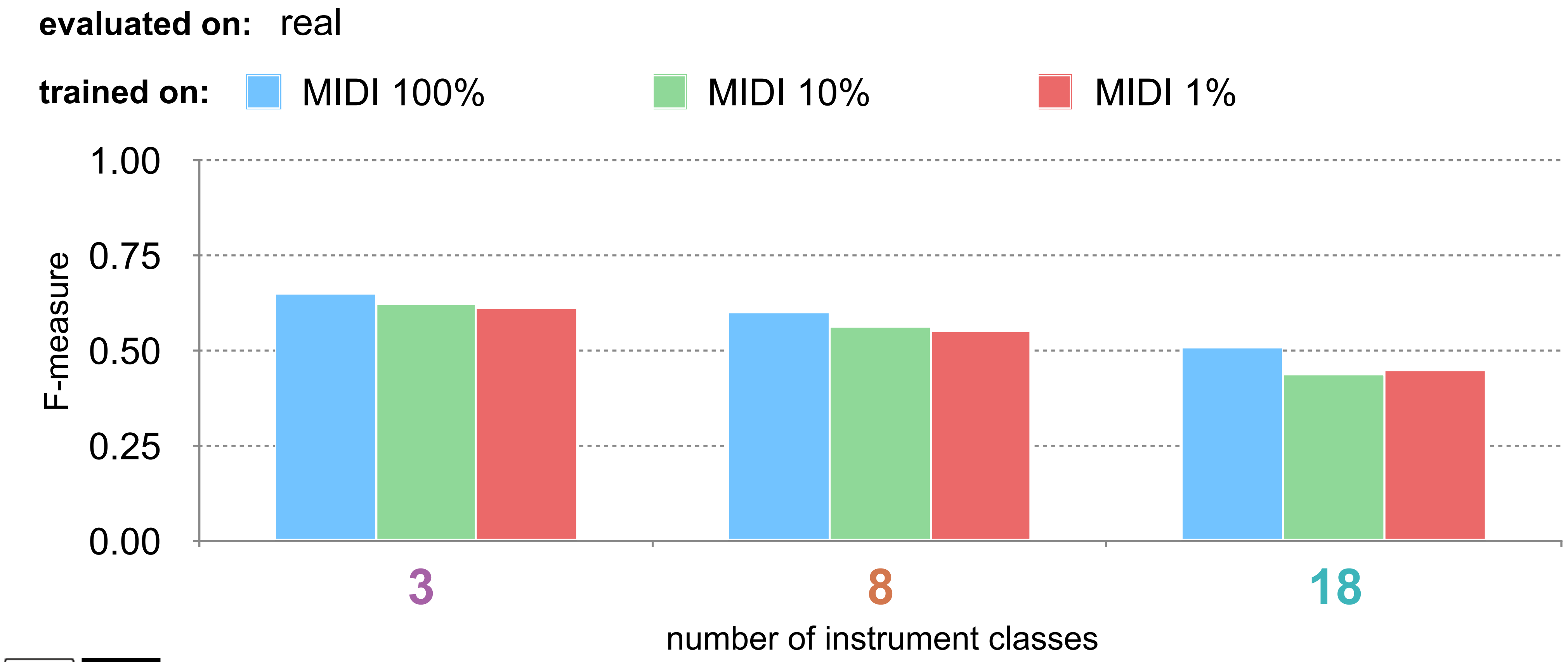
OVERALL PERFORMANCE ON REAL DATA

- Model trained on synthetic data **performs well** on real-world data (ENST + MDB + RBMA)

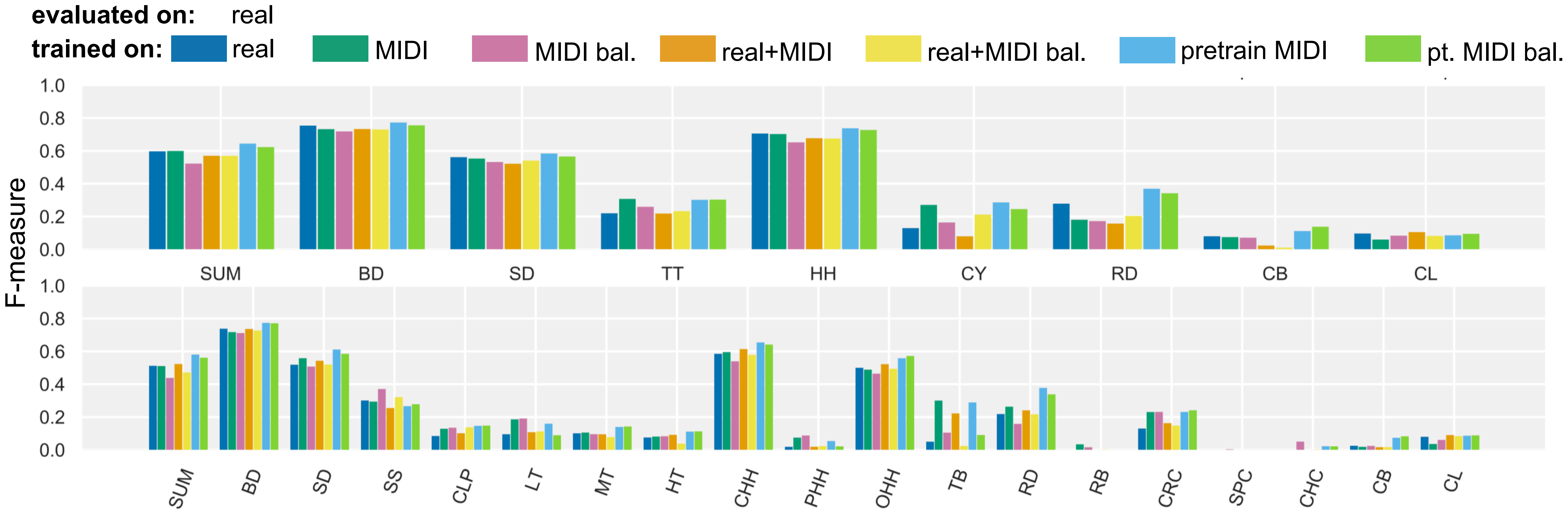


RESULTS FOR DIFFERENT SIZES

- Performance **decreases**, but **not drastically**

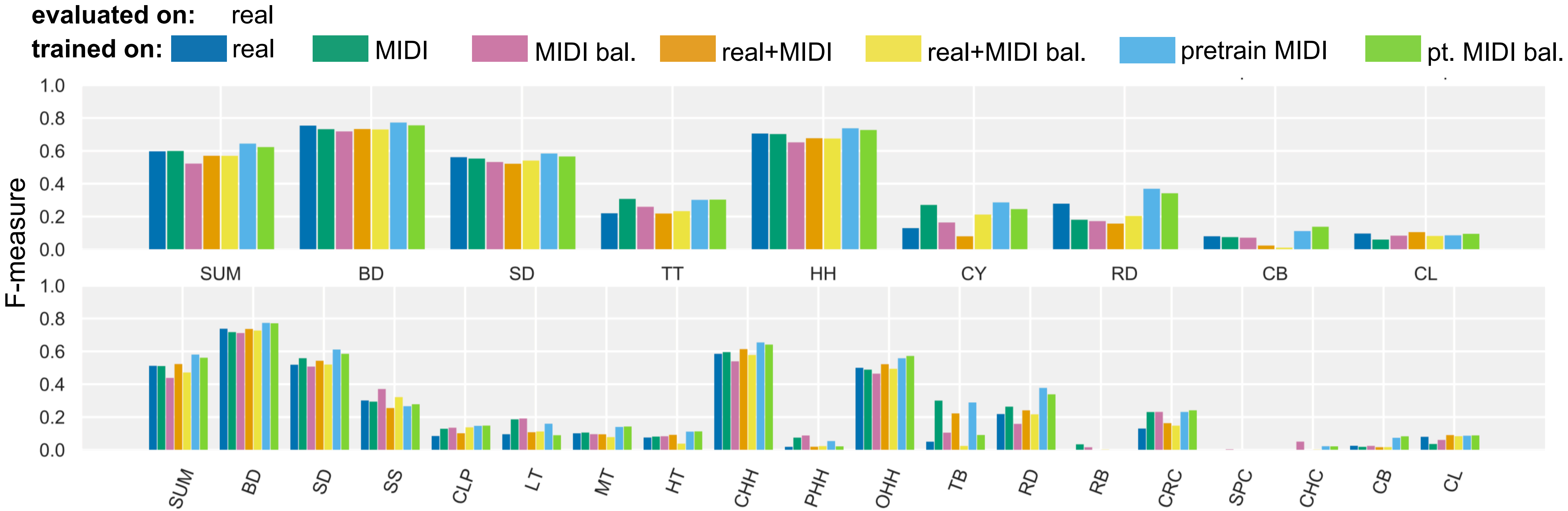


PERFORMANCE FOR INSTRUMENTS



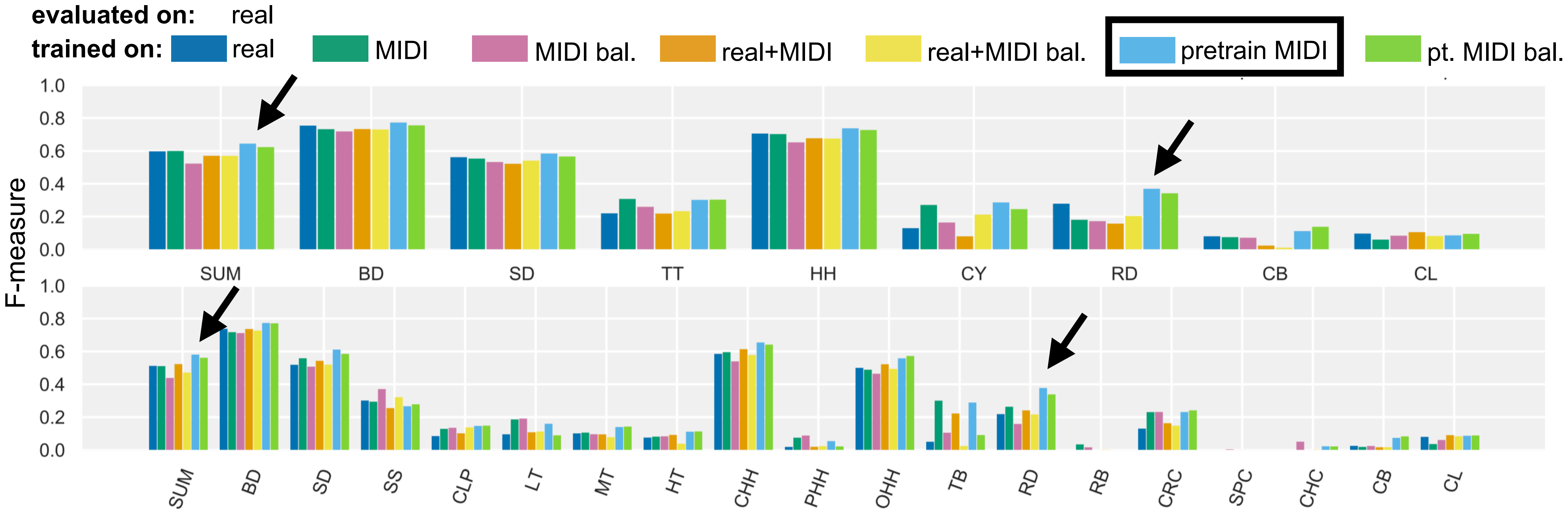
PERFORMANCE FOR INSTRUMENTS

- Improvements observed on balanced synthetic data **do not translate to real-world data**



PERFORMANCE FOR INSTRUMENTS

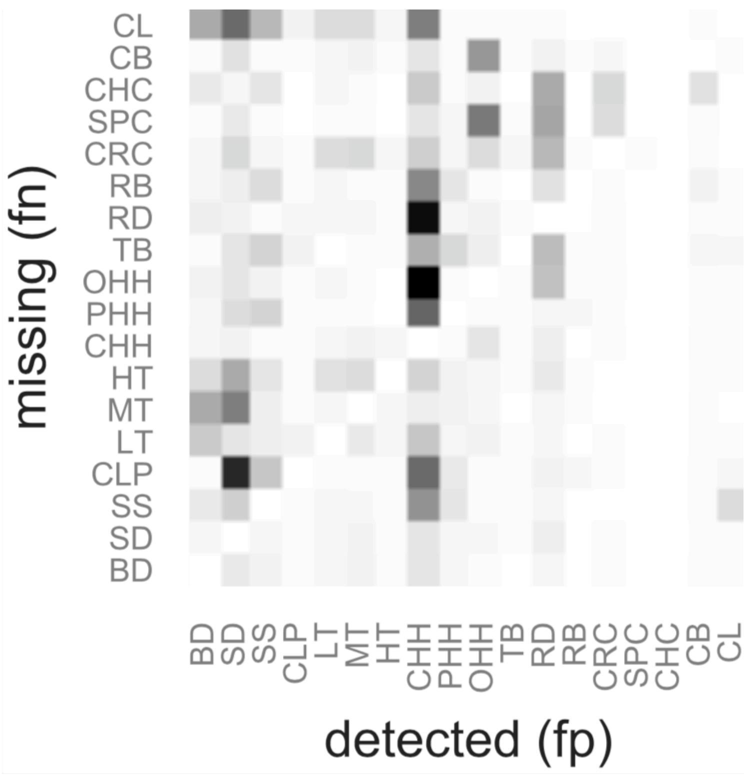
- Improvements observed on balanced synthetic data **do not translate to real-world data**
- Small improvements **using pre-training**



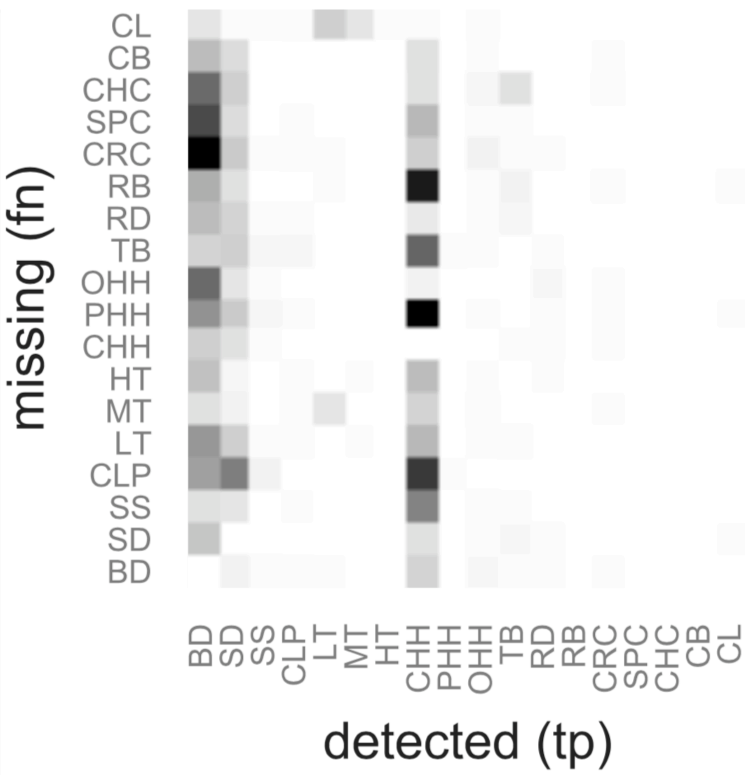
INSTRUMENT CONFUSIONS

trained on: real+MIDI
evaluated on: real+MIDI

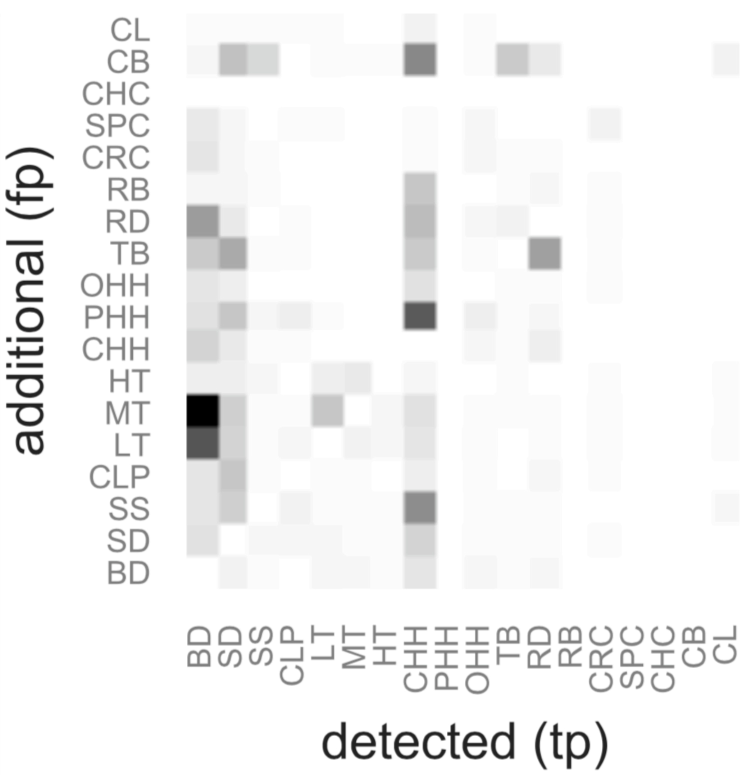
confusions



hidden onsets



additional onsets



CAN YOU HEAR THE DIFFERENCE?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

1: low tom
2: bass drum
3: bass drum

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

1: low tom
2: bass drum
3: bass drum

- Which **cymbal** is it?

hi-hat
splash cymbal
crash cymbal
Chinese cymbal
ride cymbal

?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

1: low tom
2: bass drum
3: bass drum

- Which **cymbal** is it?

hi-hat
splash cymbal
crash cymbal
Chinese cymbal
ride cymbal

?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass drum** or **low tom**?

♫ ♫

1: low tom
2: bass drum
3: bass drum

- Which **cymbal** is it?

hi-hat
splash cymbal
crash cymbal
Chinese cymbal
ride cymbal

?

♫ ♫

CAN YOU HEAR THE DIFFERENCE?

- **Bass** drum or **low tom**?

♫ ♫

1: low tom
2: bass drum
3: bass drum

- Which **cymbal** is it?

hi-hat
splash cymbal
crash cymbal
Chinese cymbal
ride cymbal

?

♫ ♫

1: crash
2: ride
3: China

CONCLUSIONS

- Publicly available large scale **synthetic dataset** `http://ifs.tuwien.ac.at/~vogl/dafx2018/`
  - Optional with **balanced instruments**
  - **Generalizes** well to real data
- Dataset size important but not that critical
- Balancing **did not improve** performance on real-world data
  - Recurrent layers learn **untypical patters**
- **Pre-training** with synthetic data provides small improvement
- **Mistakes** are *understandable*
  - Focus more on context