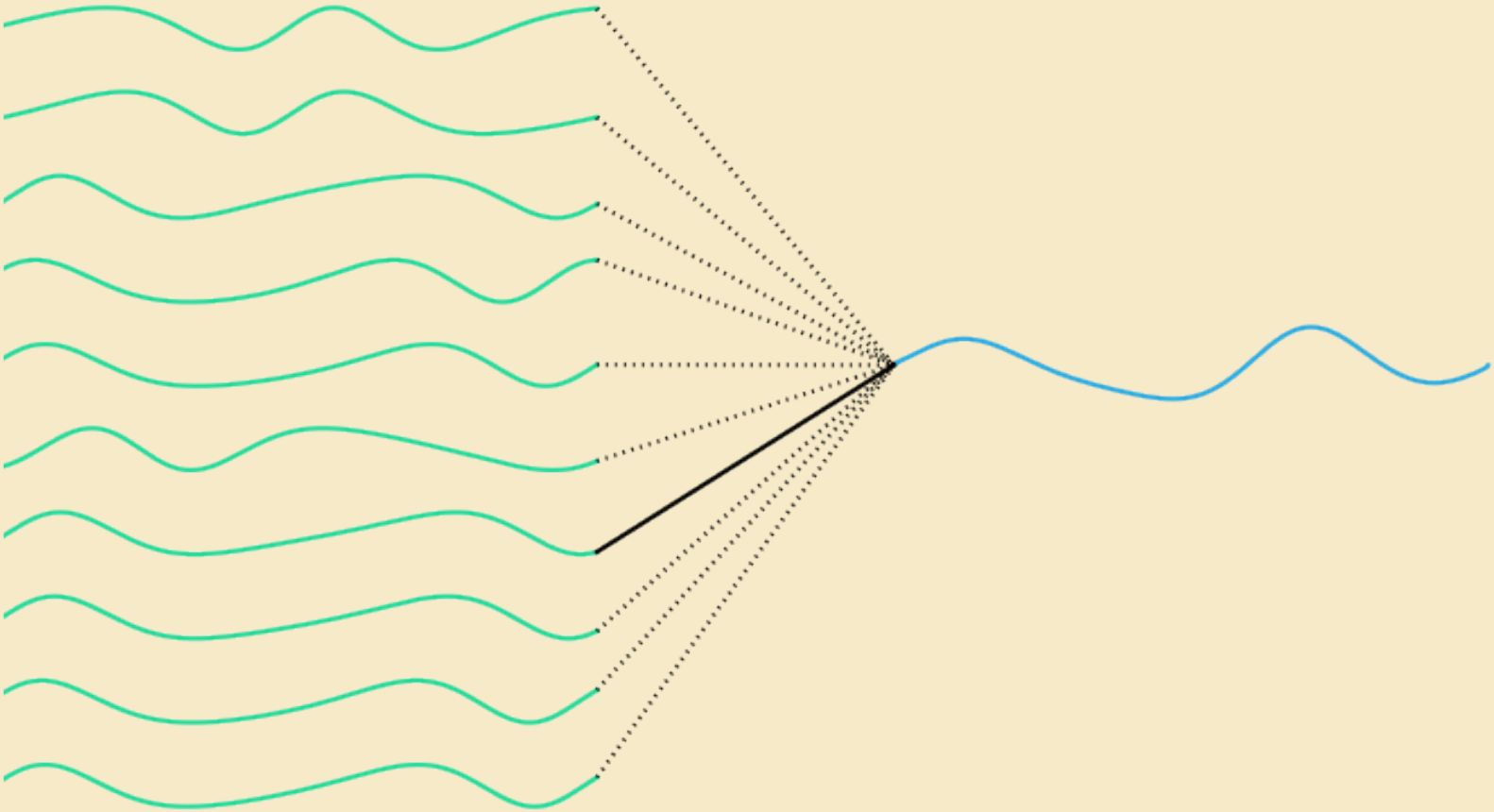
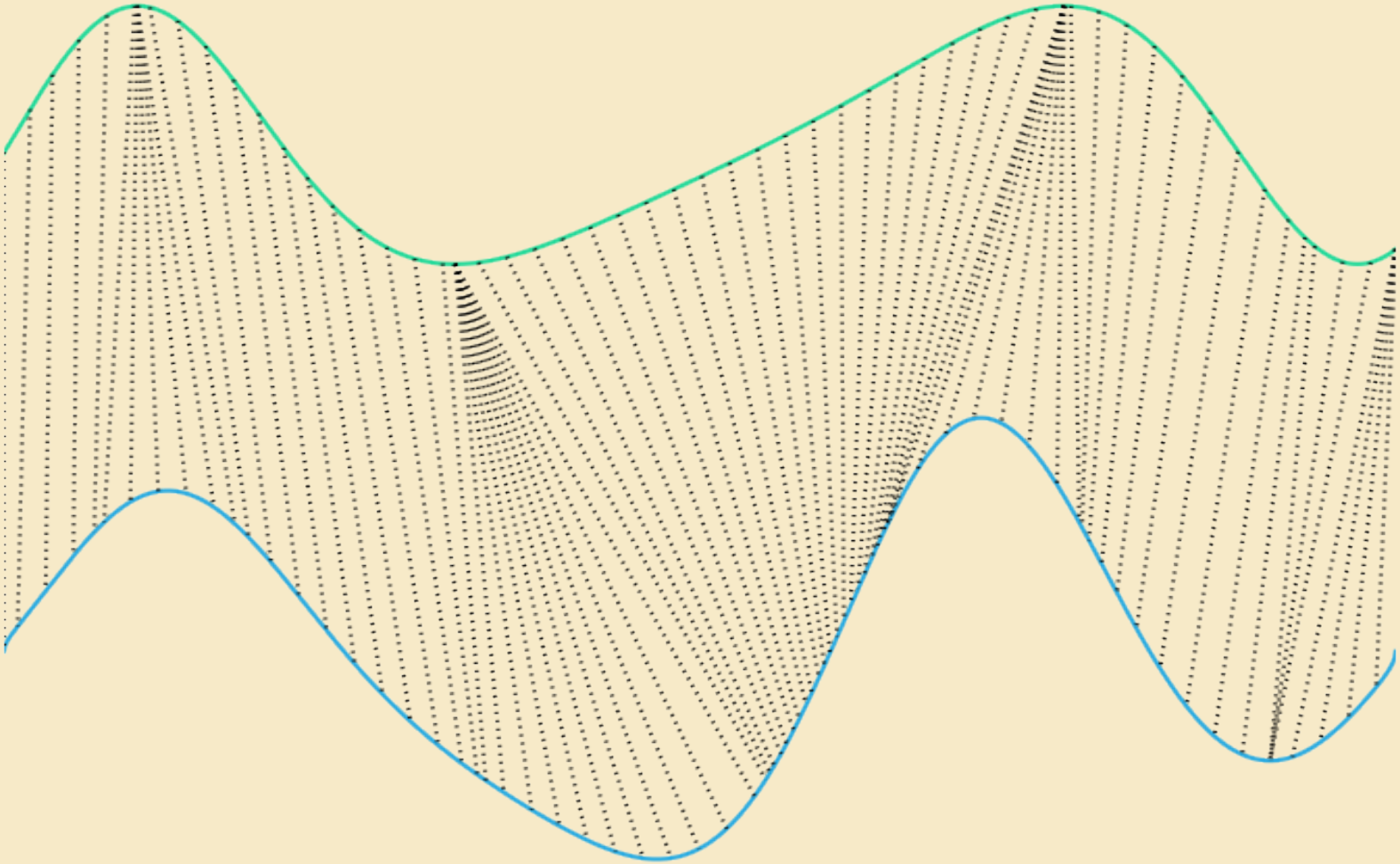
Accelerating Large-Scale Sequence Retrieval with Convolutional Networks

Colin Raffel
IIT Bombay
December 23rd, 2015

Sequence Retrieval

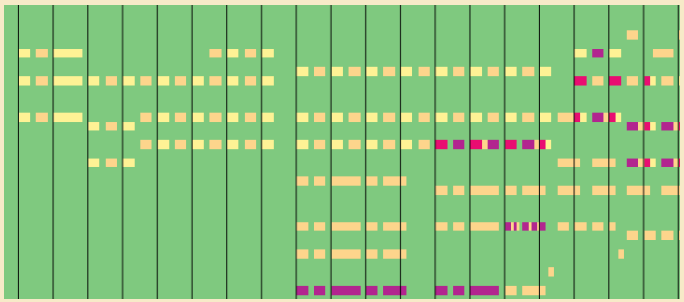


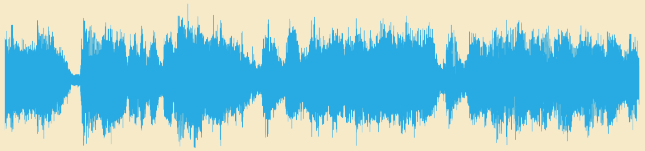
Dynamic Time Warping



Goal

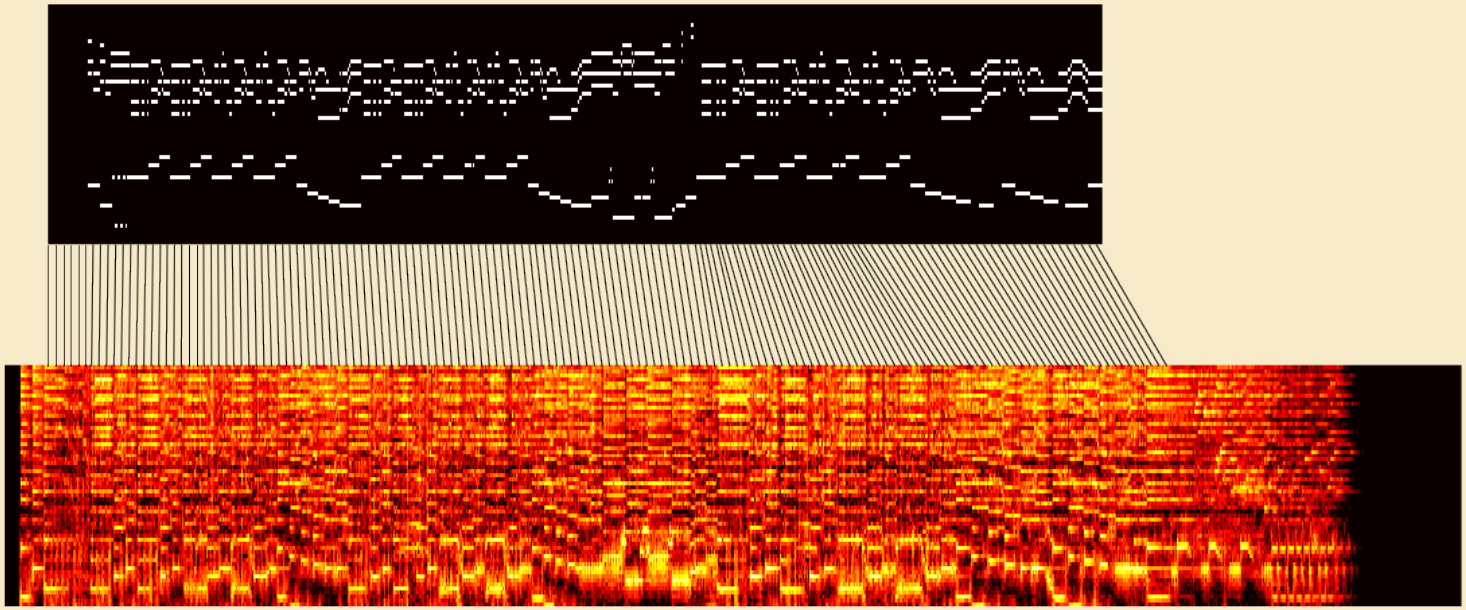
```
      artist: 'Tori Amos'
     release: 'LIVE AT MONTREUX'
       title: 'Smells Like Teen Spirit'
          id: 'TRKUYPW128F92E1FC0'
    duration: 216.4502
 sample_rate: 22050
   audio_md5: '8'
 7digitalid: 5764727
        year: 1992
```





Matching and Aligning

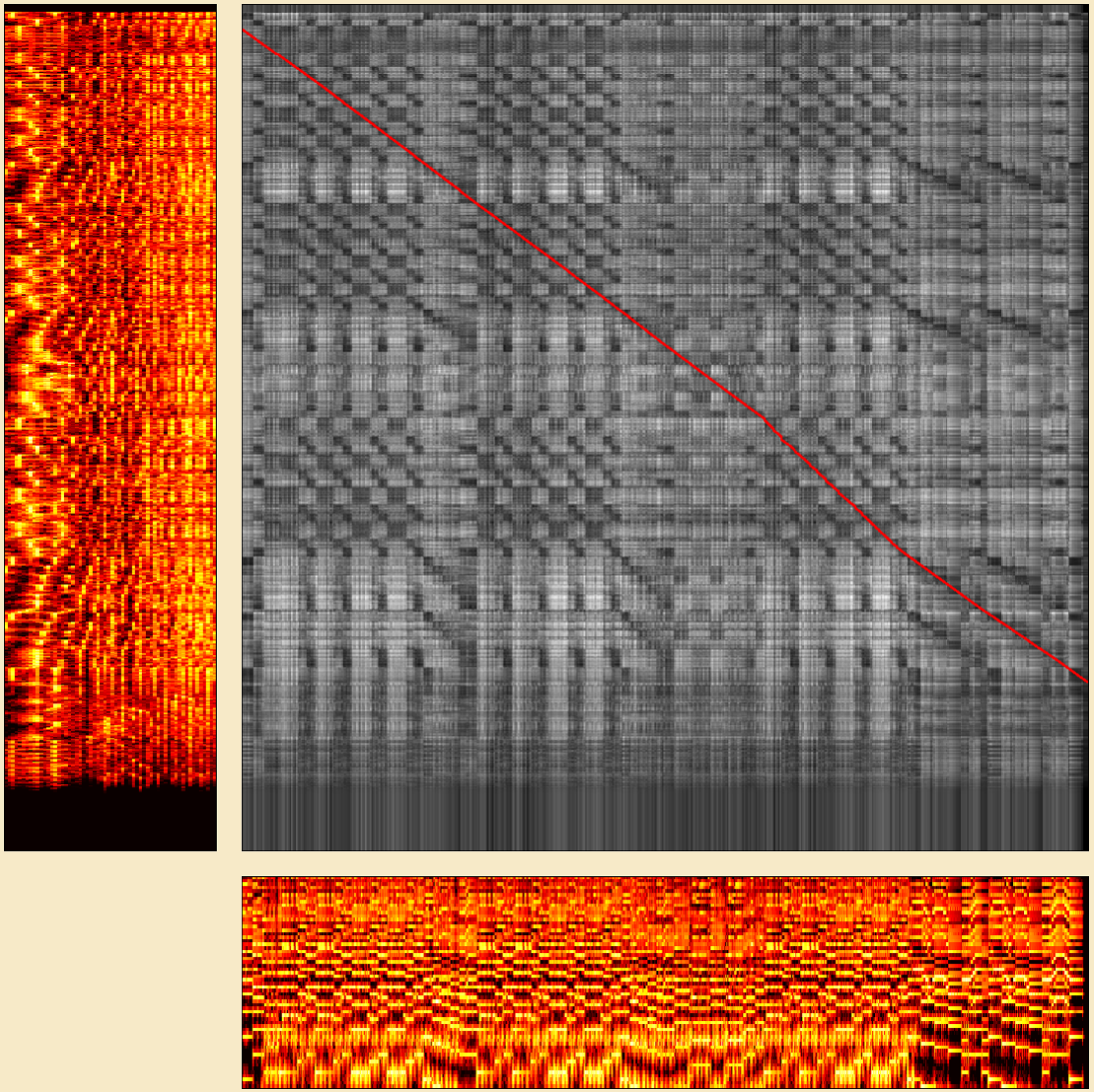
`Beatles/hello, goodbye.mid` ↔ `The Beatles/Hello.mp3`



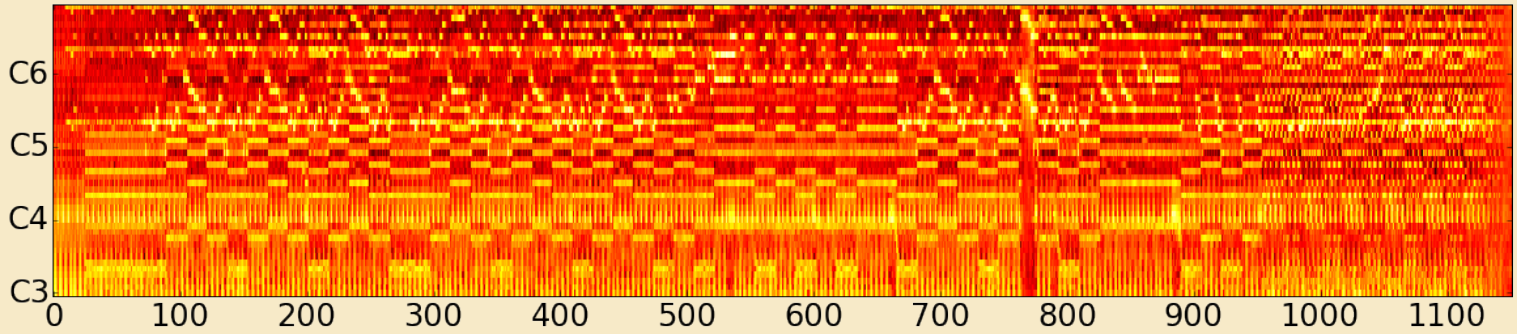
Matching by Metadata Won't Work

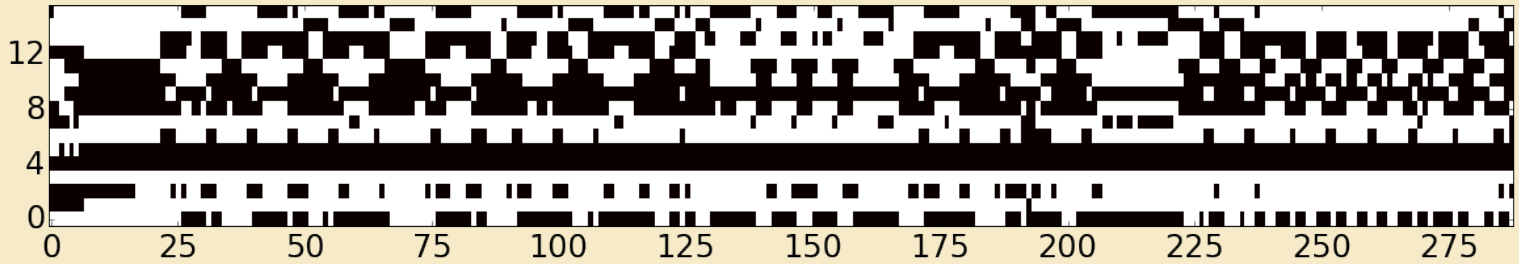
J/Jerseygi.mid
V/VARIA18O.MID
Carpenters/WeveOnly.mid
2009 MIDI/handy_man1-D105.mid
G/Garotos Modernos - Bailanta De Fronteira.mid
Various Artists/REWINDNAS.MID
GoldenEarring/Twilight_Zone.mid
Sure.Polyphone.Midi/Poly 2268.mid
d/danza3.mid
100%sure.polyphone.midi/Fresh.mid
rogers_kenny/medley.mid
2009 MIDI/looking_out_my_backdoor3-Bb192.mid

DTW: Natural, and Too Slow



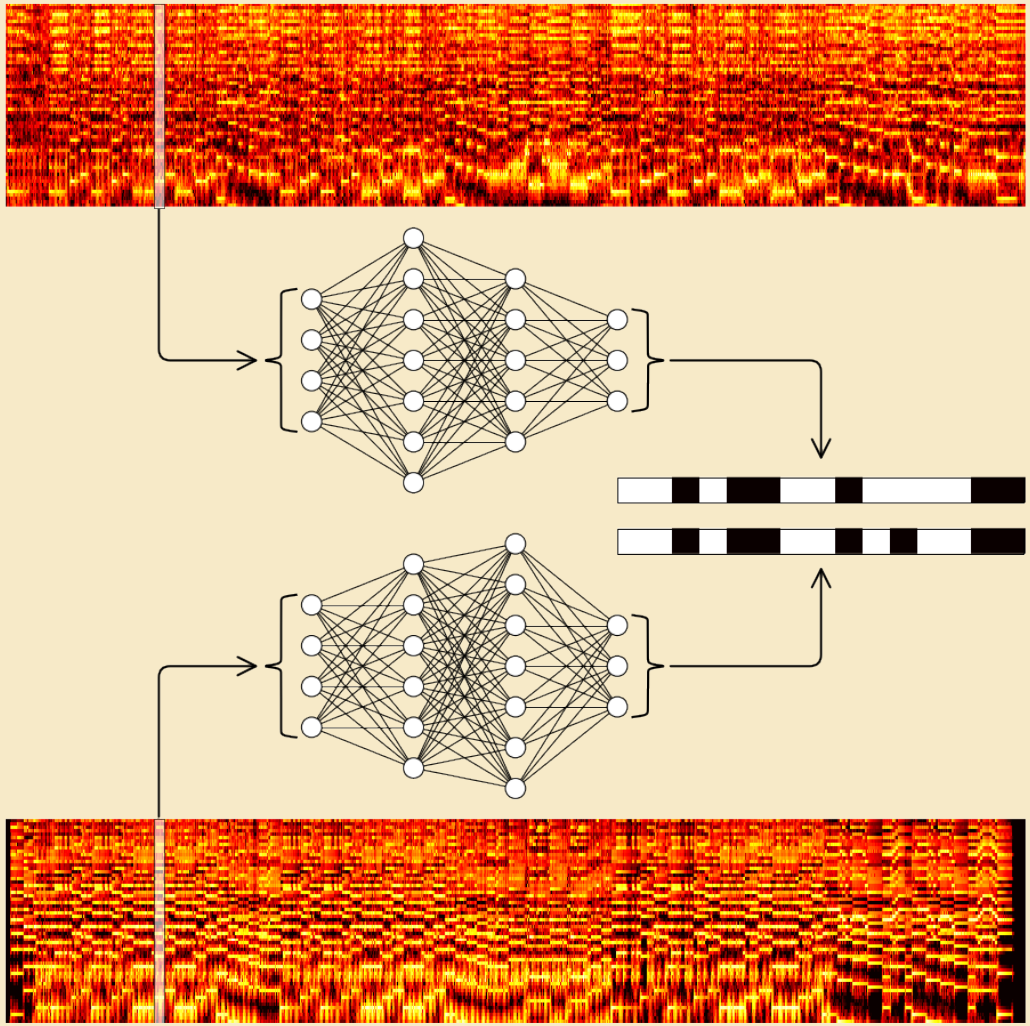
Hash Sequences



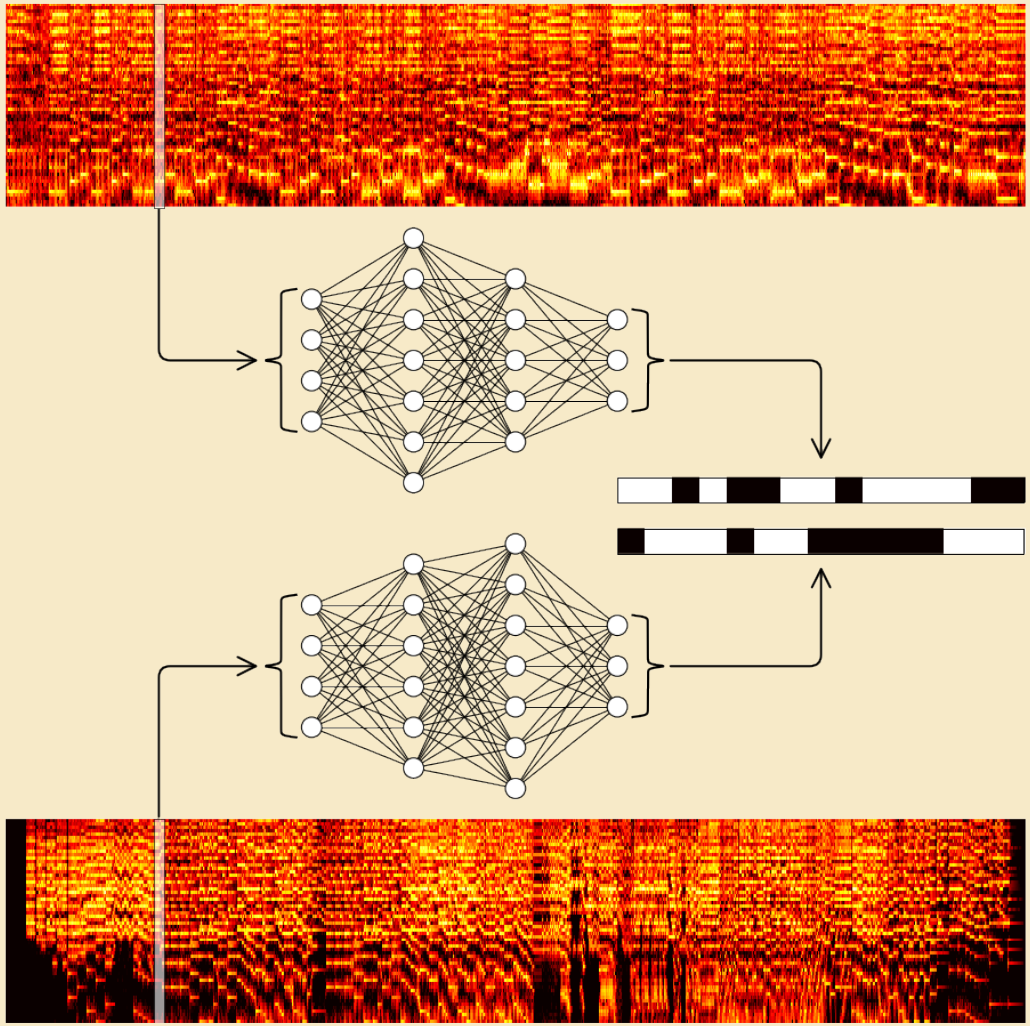


$$\texttt{distance}[m, n] = \texttt{bits\_set}[x[m] \oplus y[n]]$$

Similarity-Preserving Hashing



Similarity-Preserving Hashing



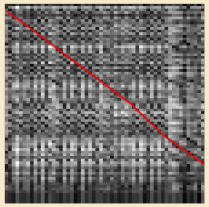
Collecting Data

140,910

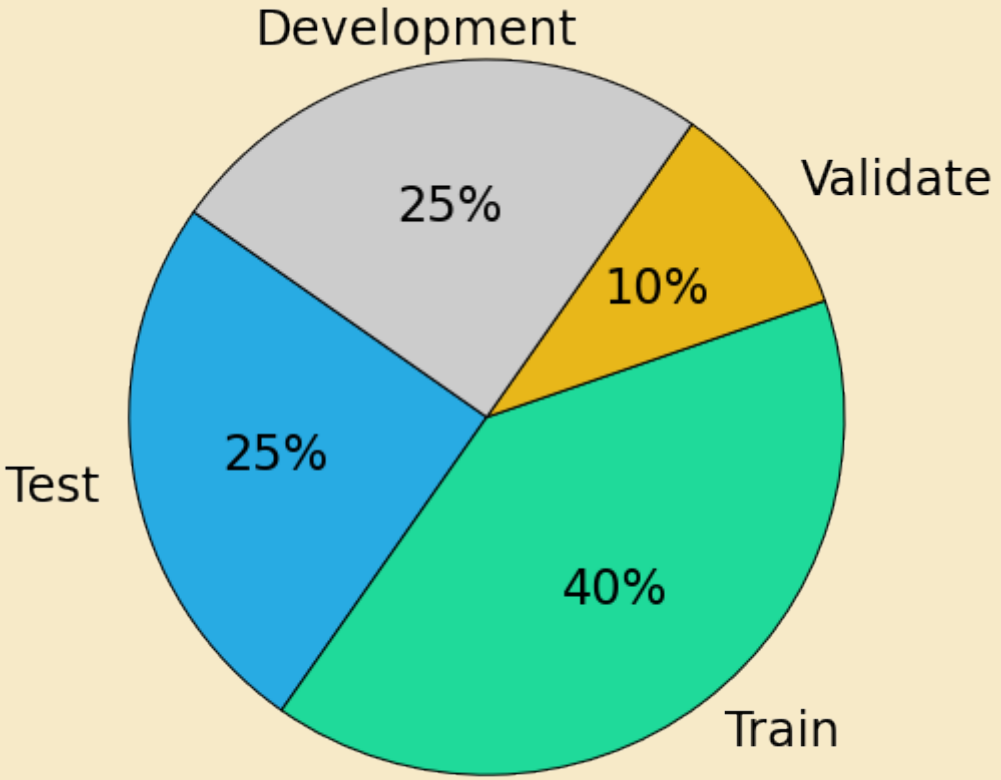
24,850

17,243

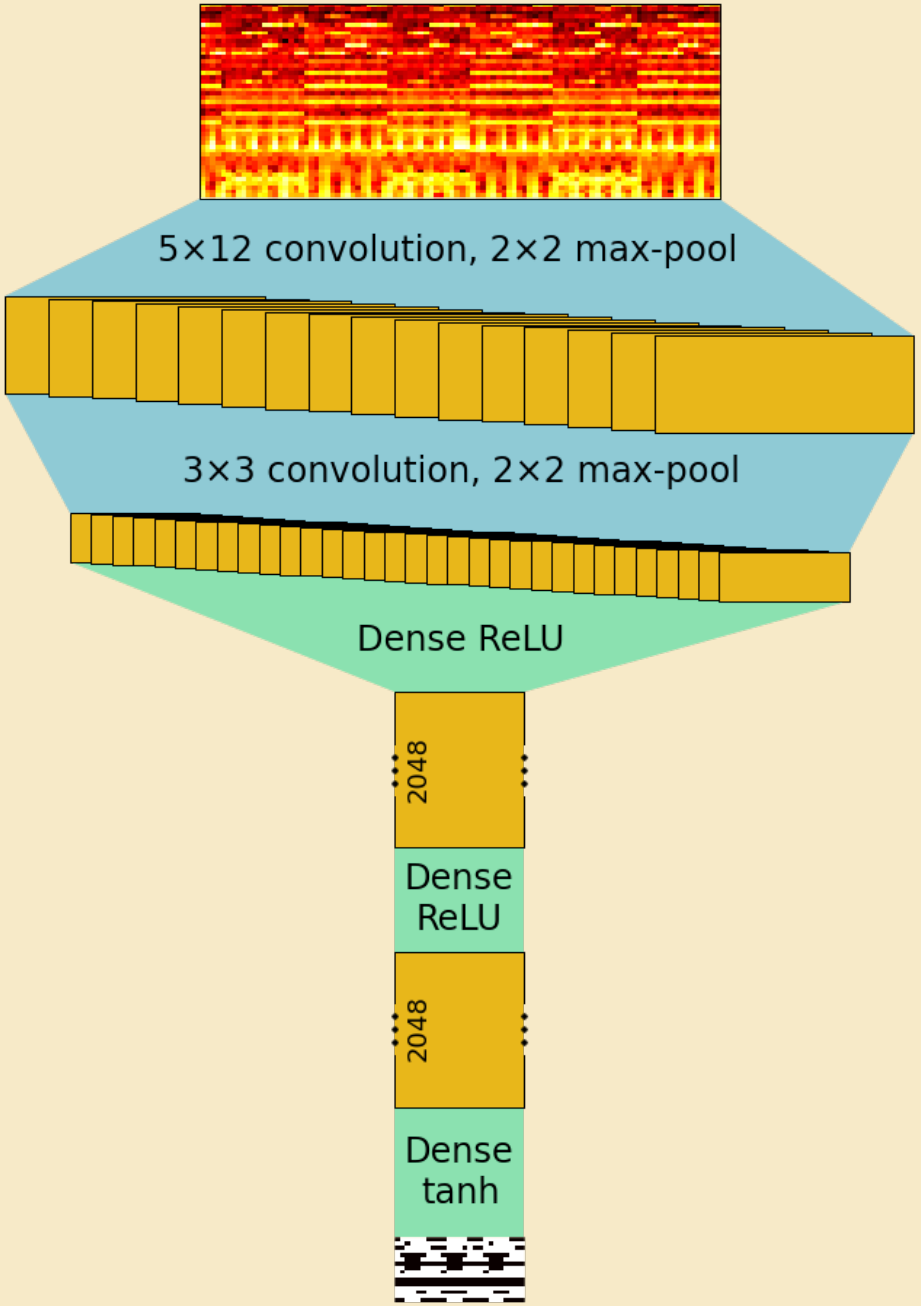
26,311



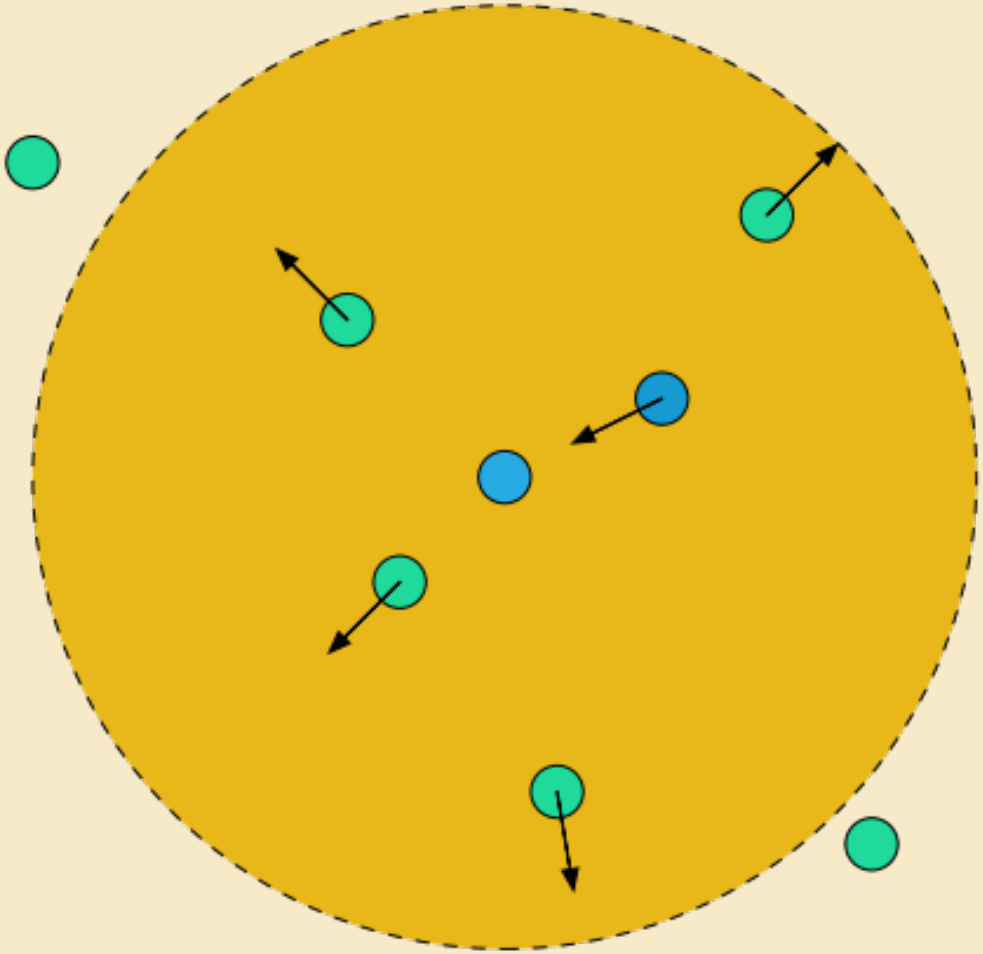
10,035



Network Structure

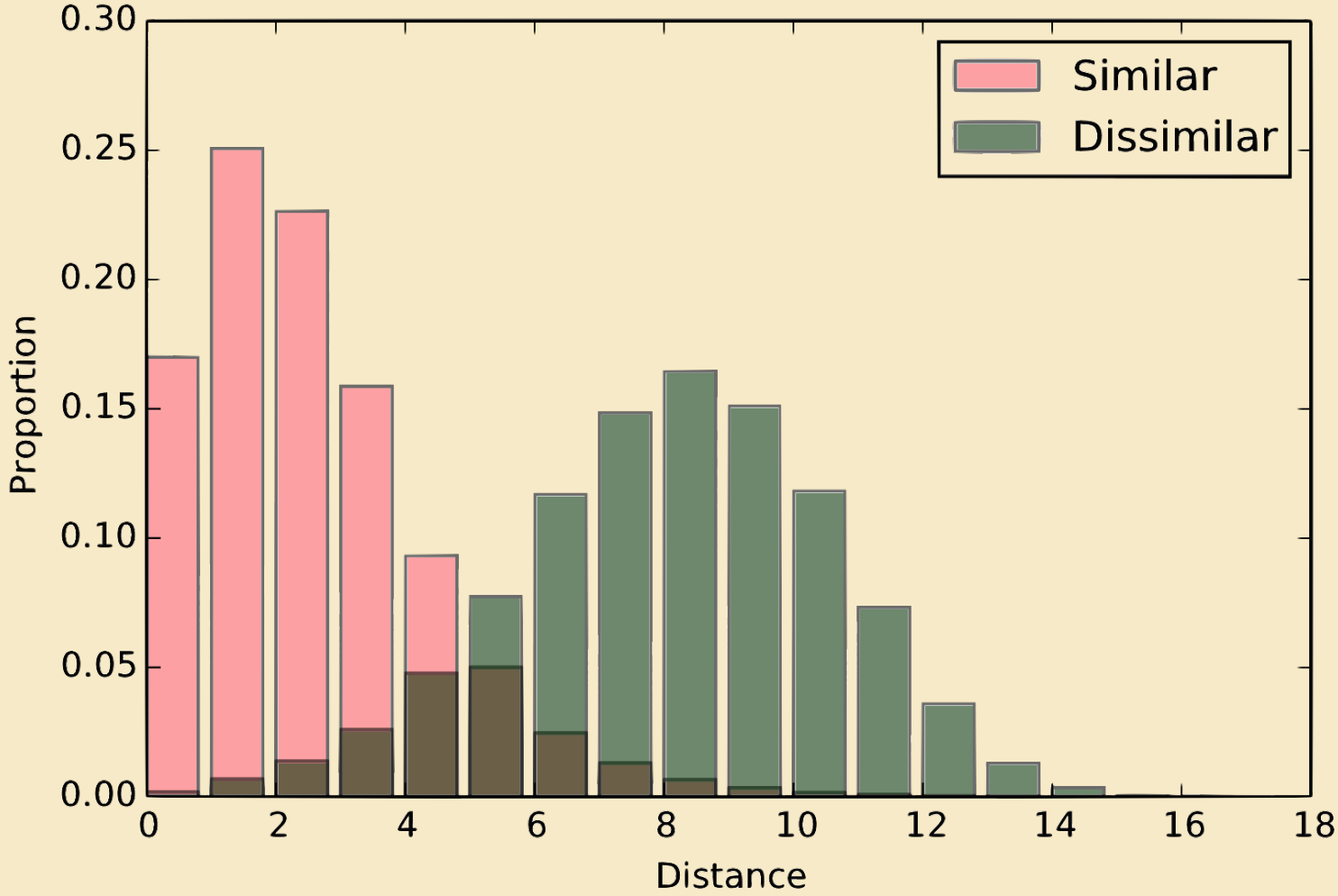


Loss Function

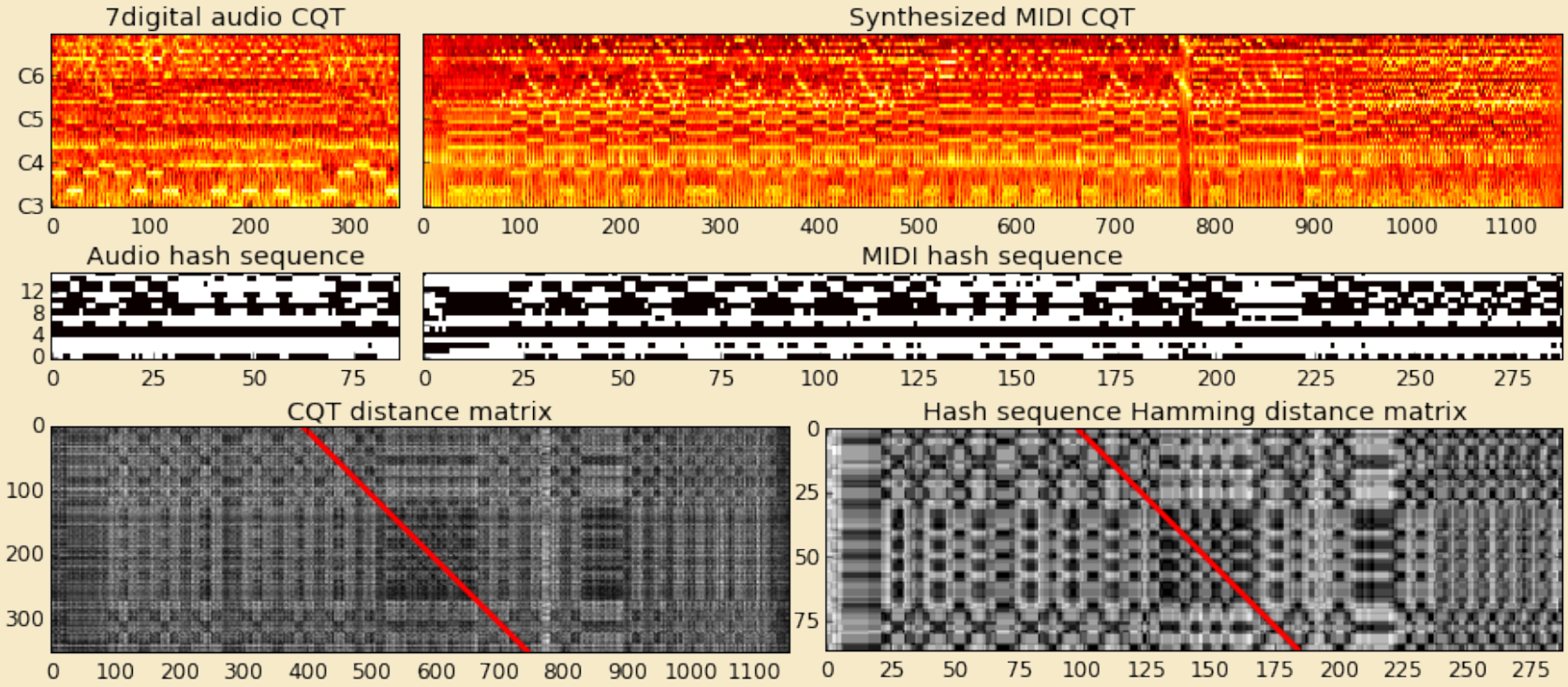


$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \|f(x) - g(y)\|_2^2 - \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y)\in\mathcal{N}} \max(0, m - \|f(x) - g(y)\|_2)^2$$

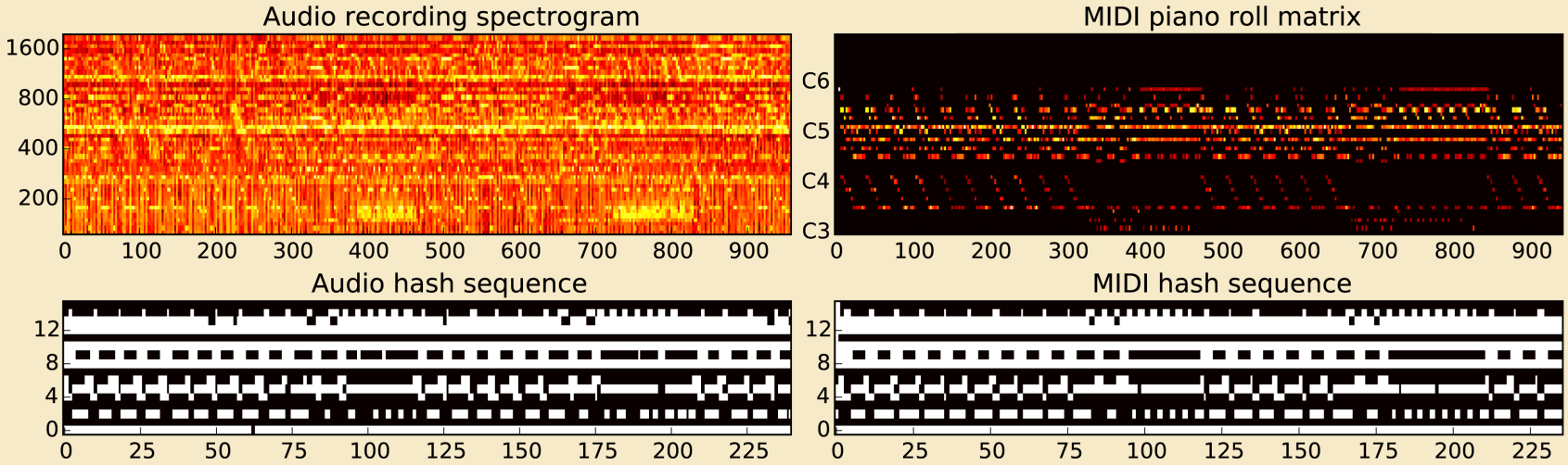
Validation Distance Distribution



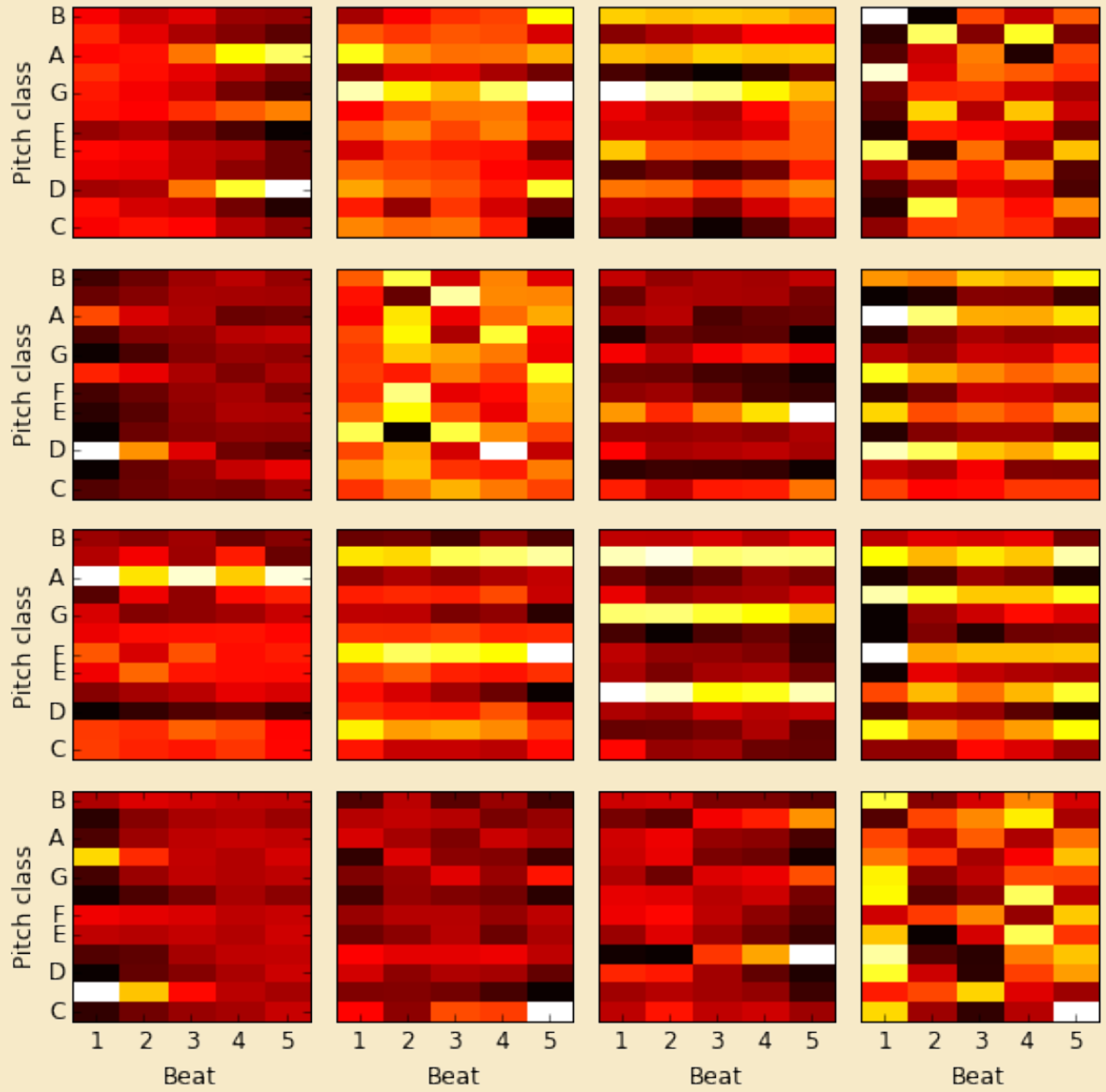
Example Sequence



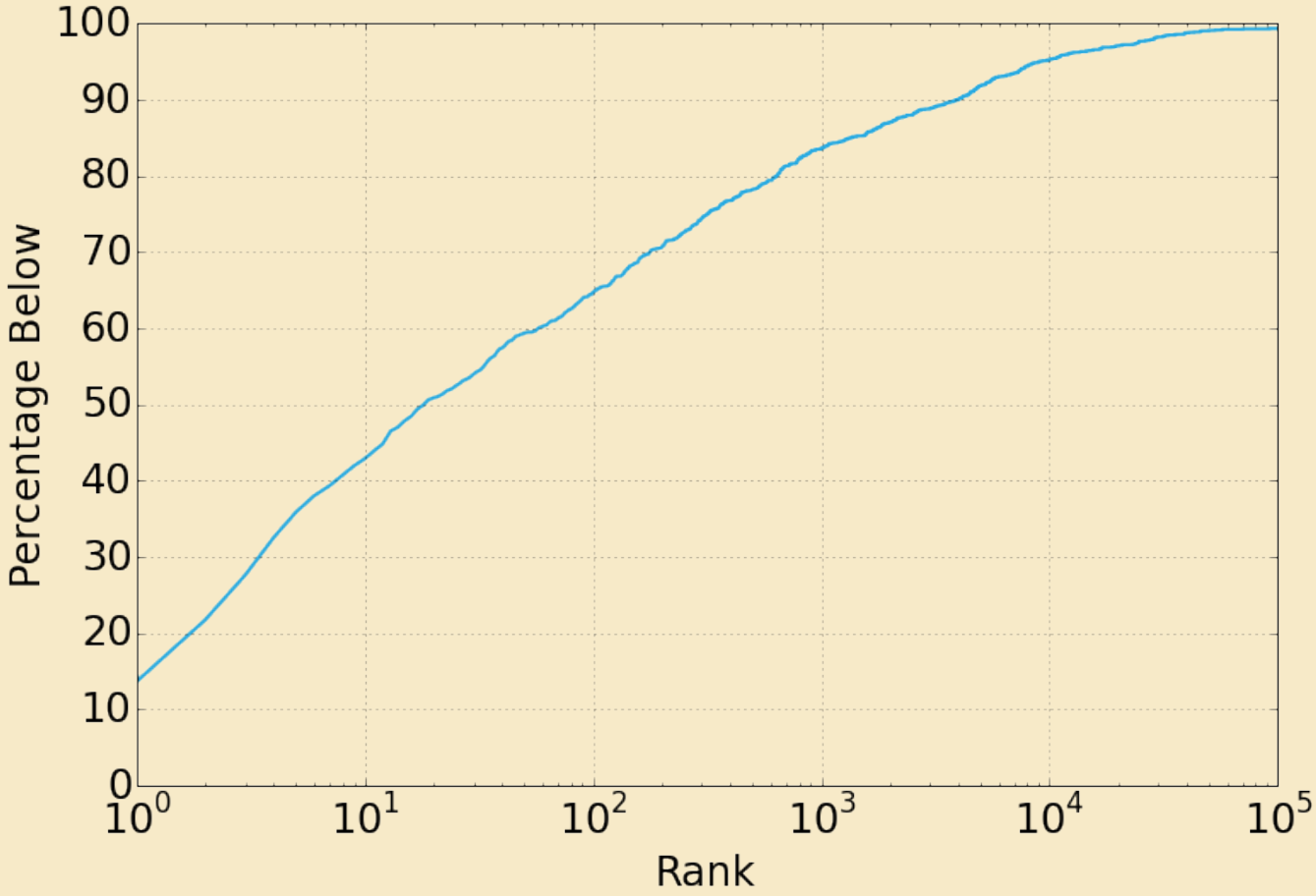
Side Note: Cross-Modal



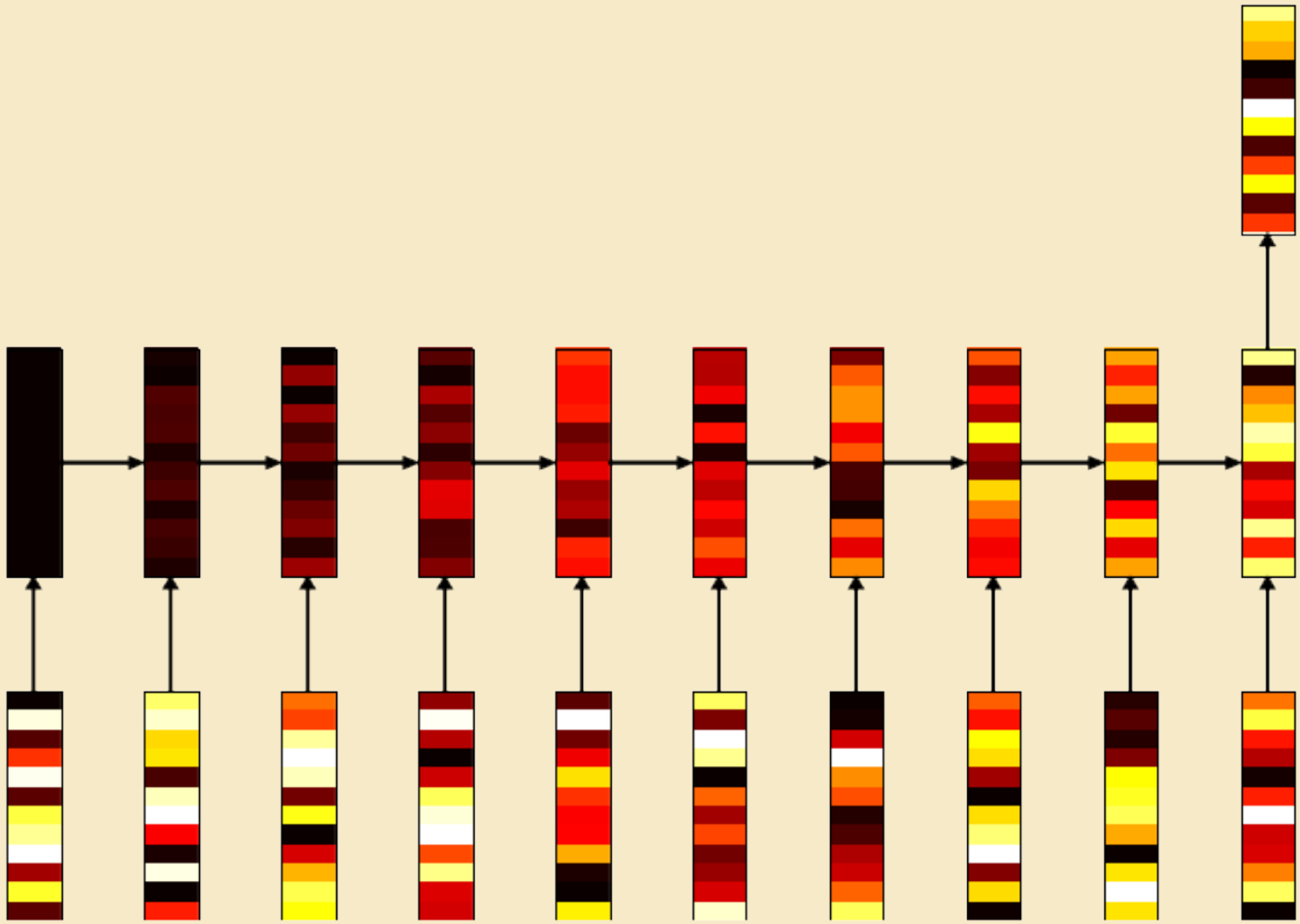
First Layer Filters



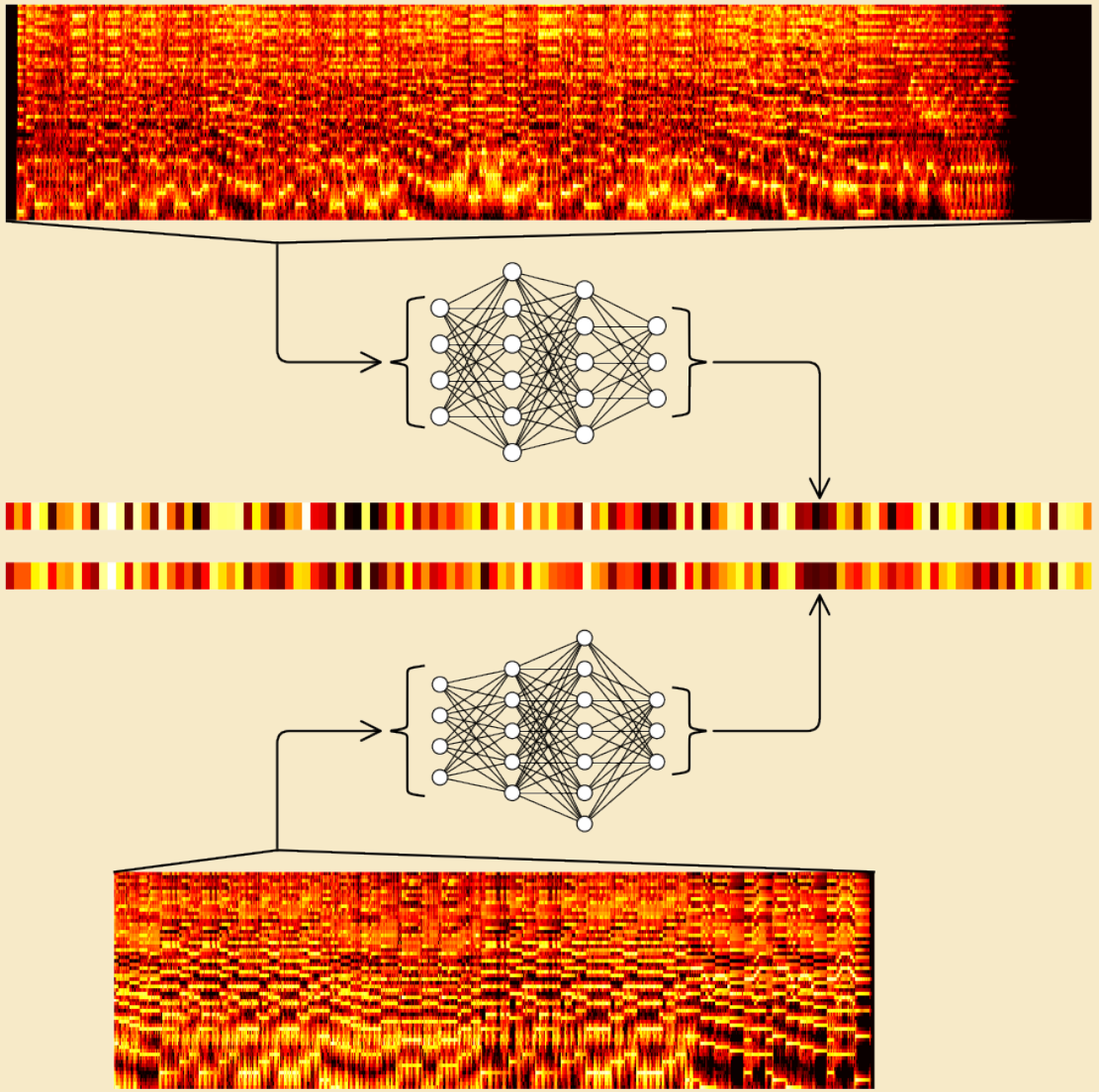
Test Set Matching Results



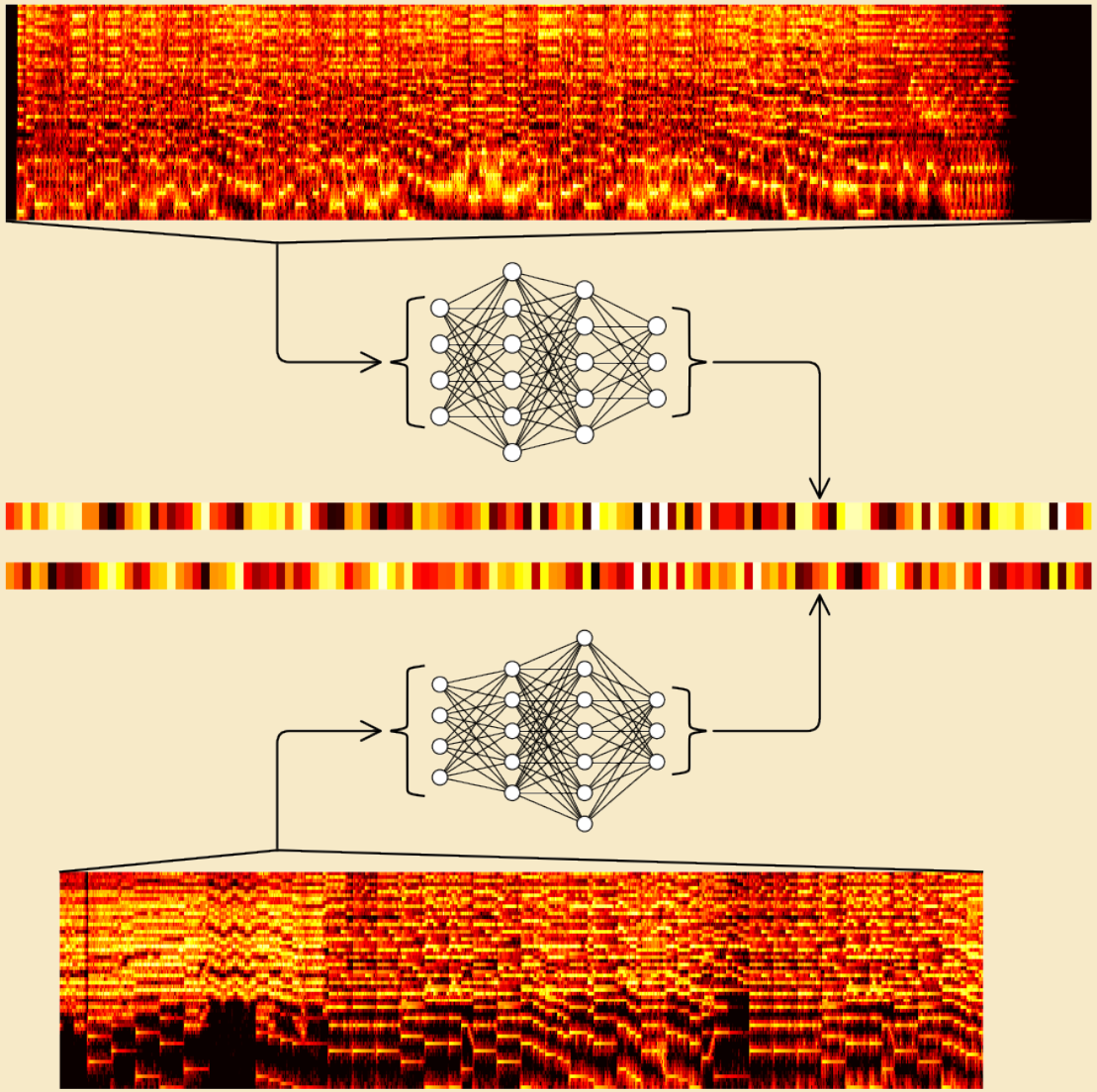
Sequence Embedding



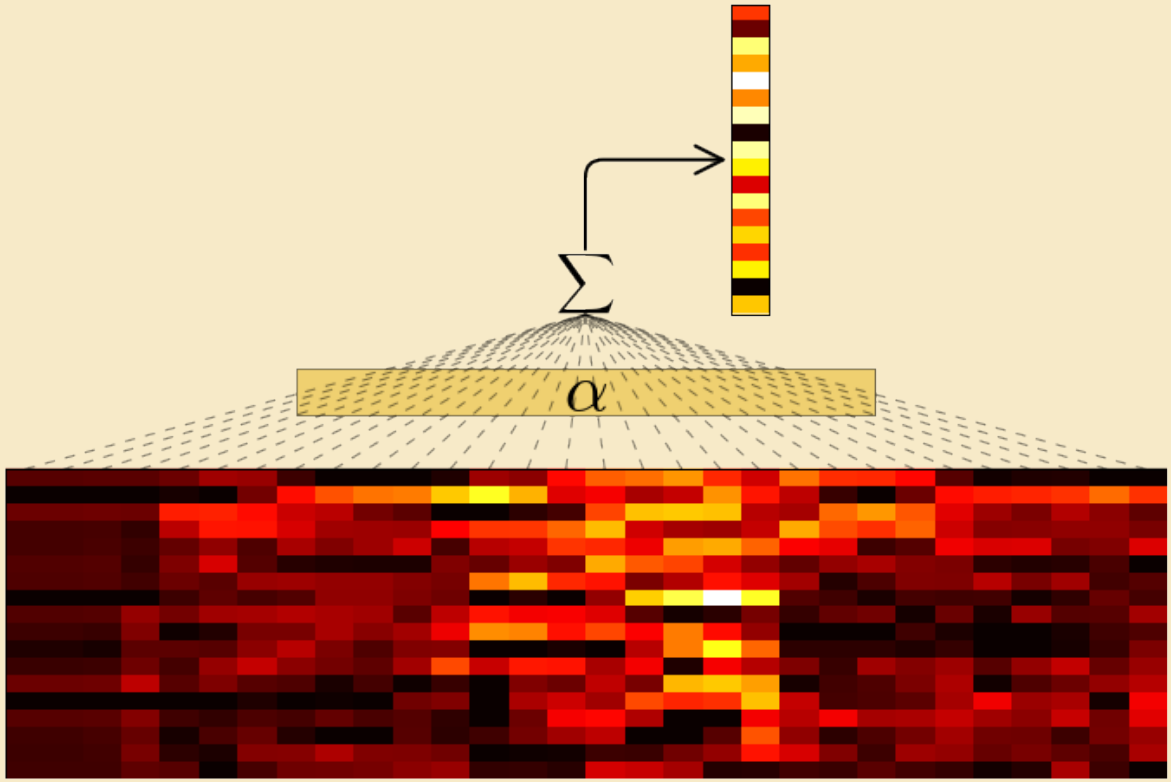
Sequence Embedding



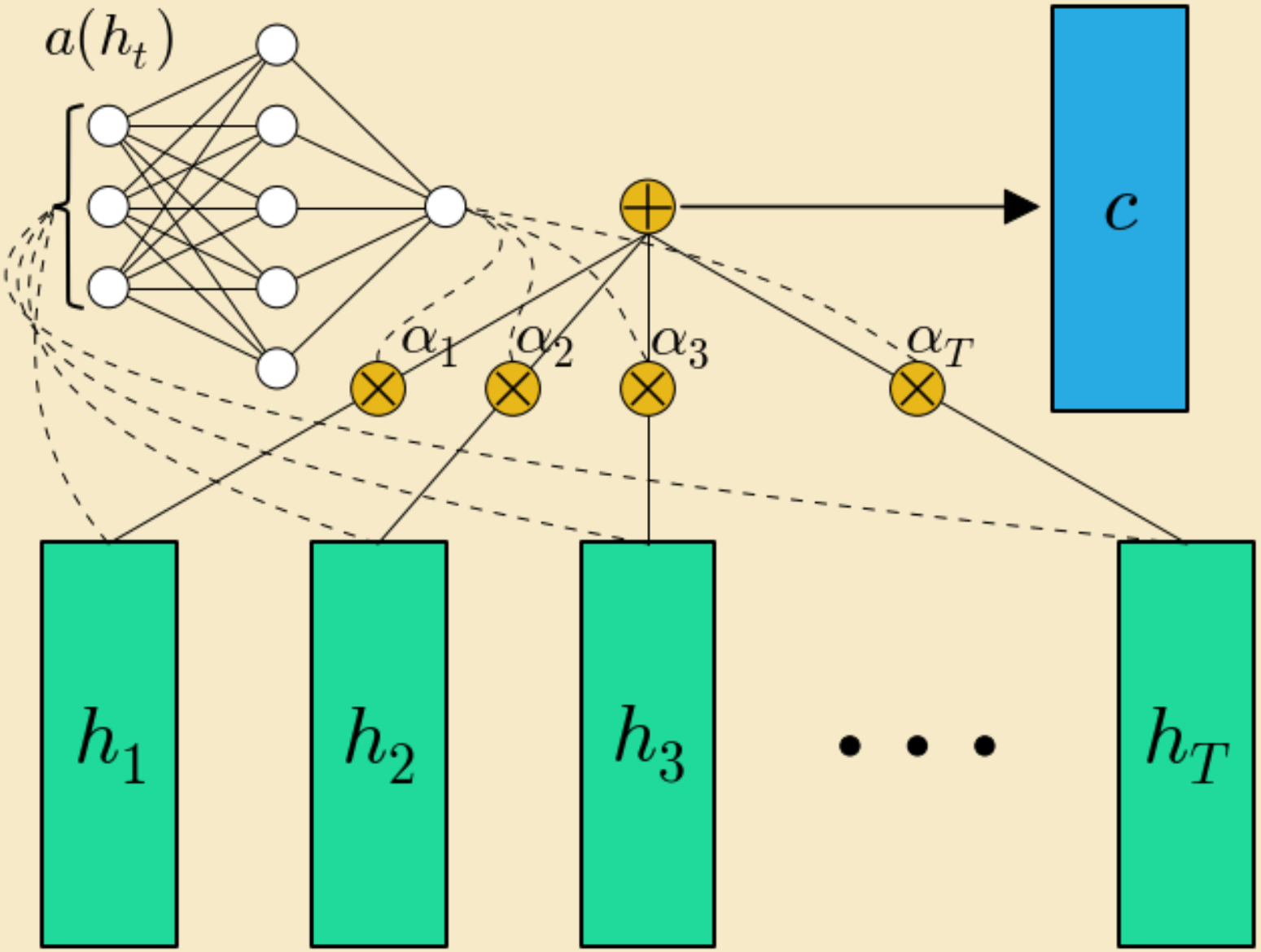
Sequence Embedding



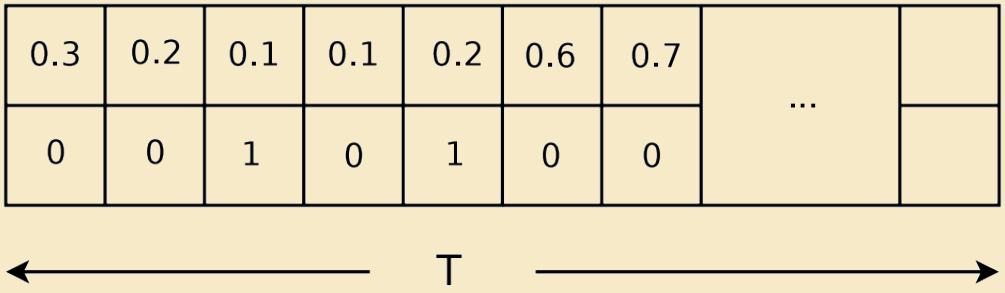
Attention



Feed-Forward Attention



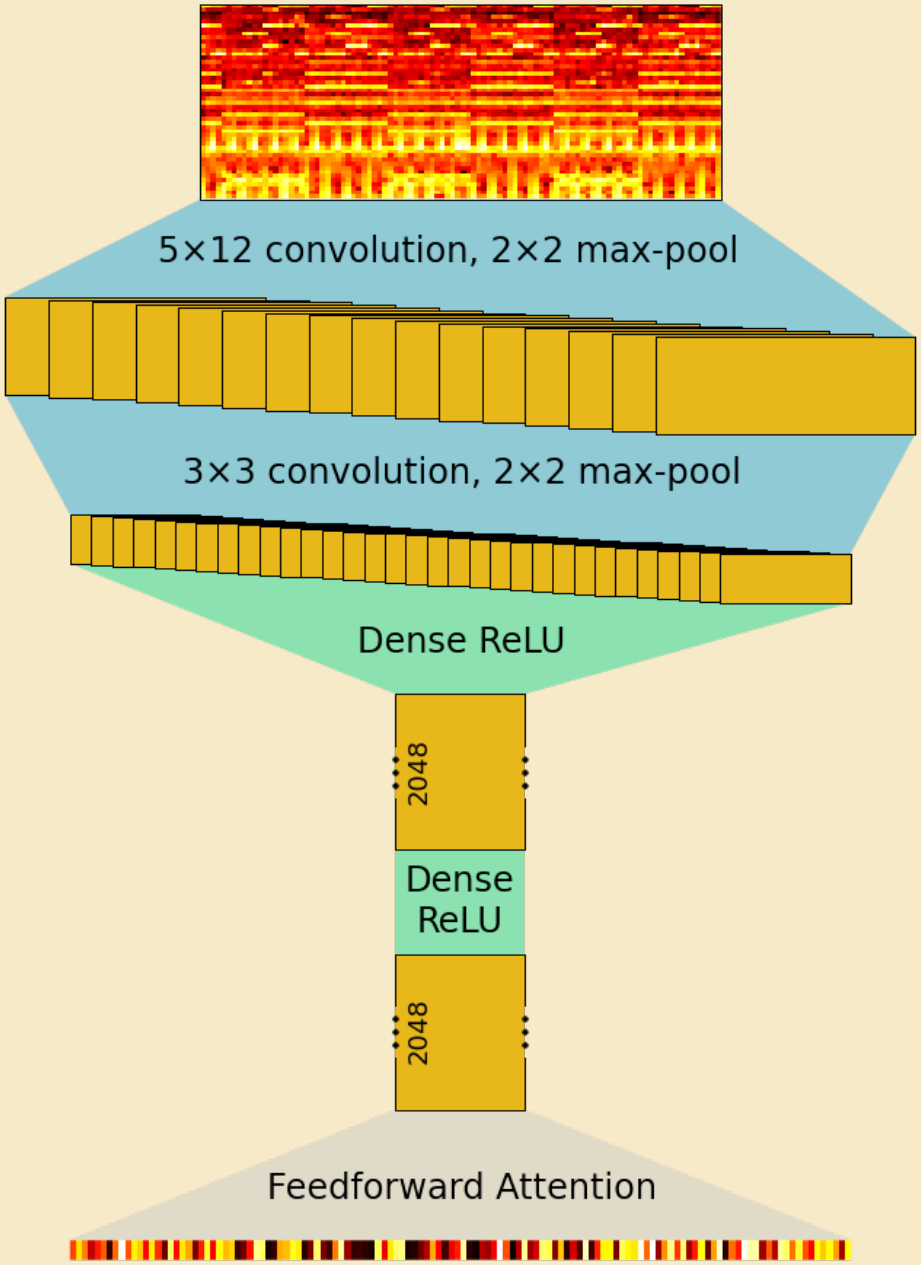
Side Note: Toy Problems



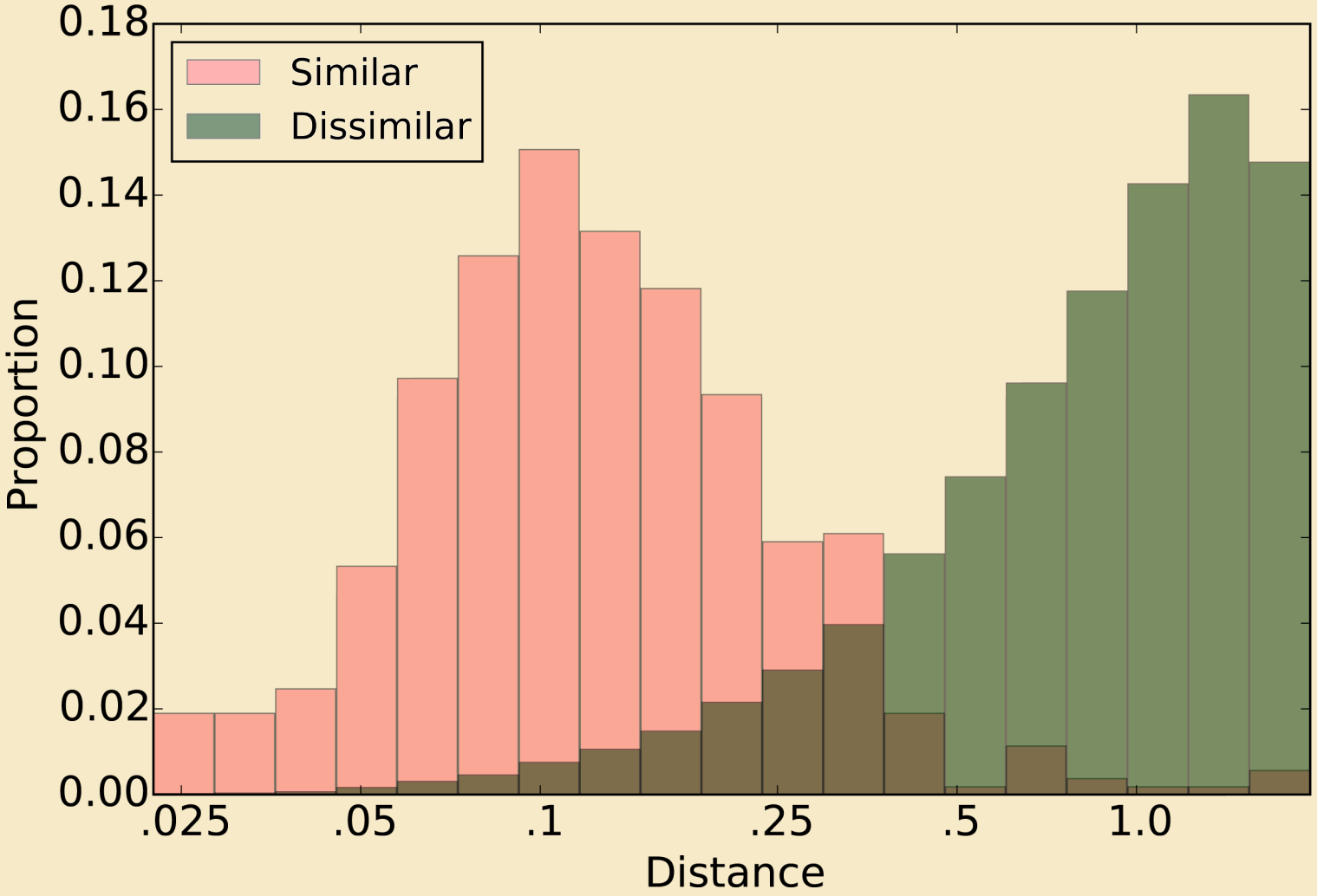
from Ilya Sutskever et al., *"On the Importance of Initialization and Momentum in Deep Learning"*

	Addition					
Sequence length (T_0)	50	100	500	1000	5000	10000
Attention	1	1	1	1	2	3
Unweighted	1	1	1	2	8	17
	Multiplication					
Sequence length (T_0)	50	100	500	1000	5000	10000
Attention	1	2	4	2	15	6
Unweighted	2	2	8	33	89.8%	80.8%

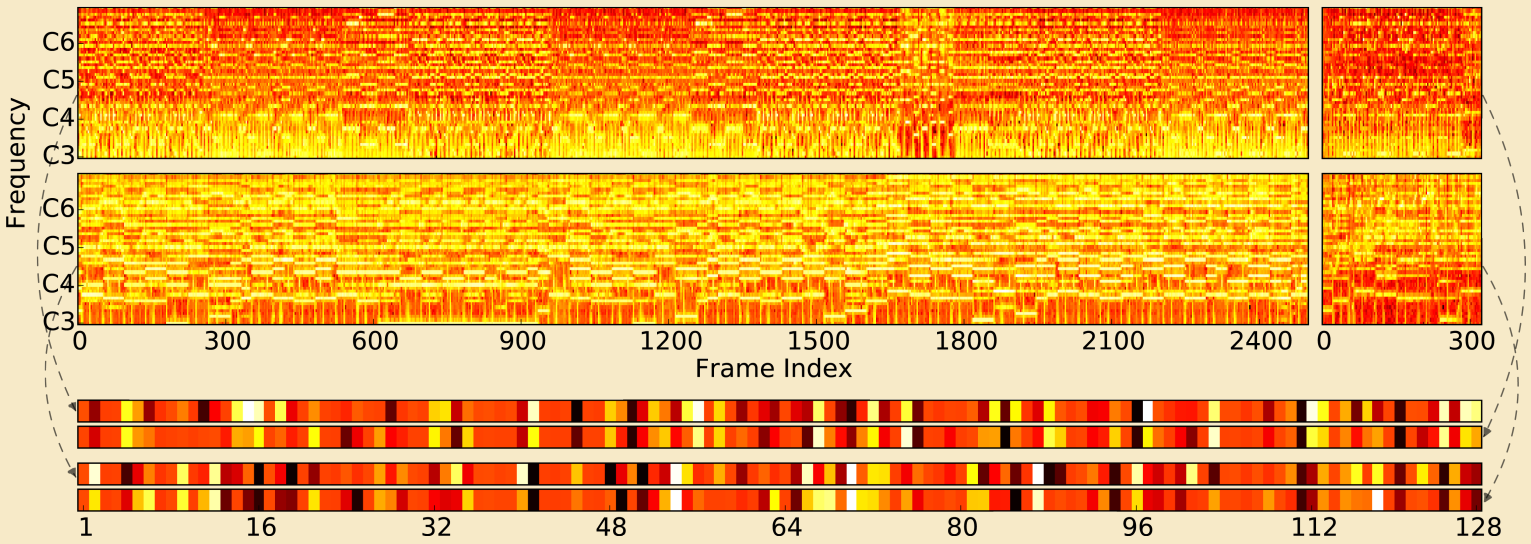
Network Structure



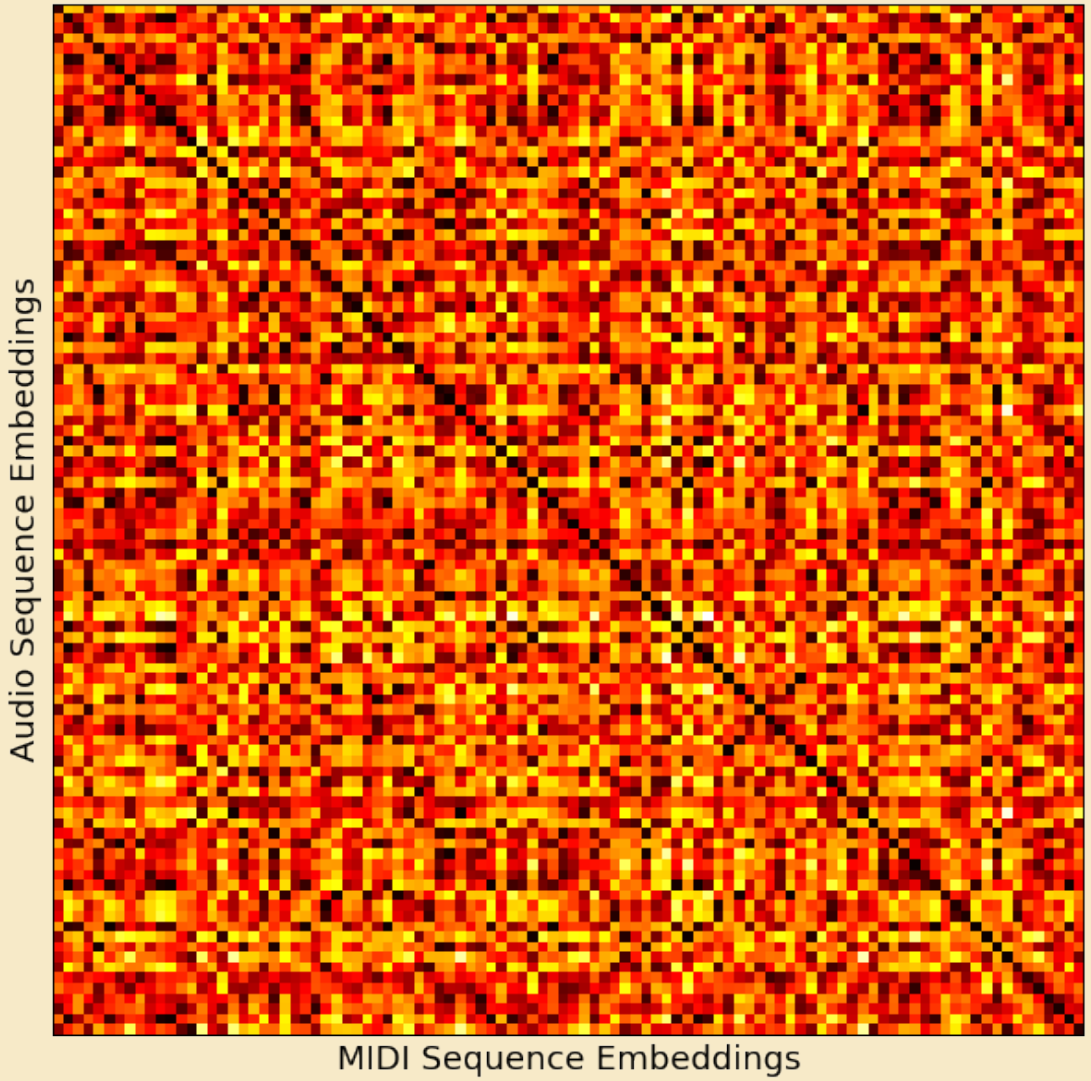
Validation Distance Distribution



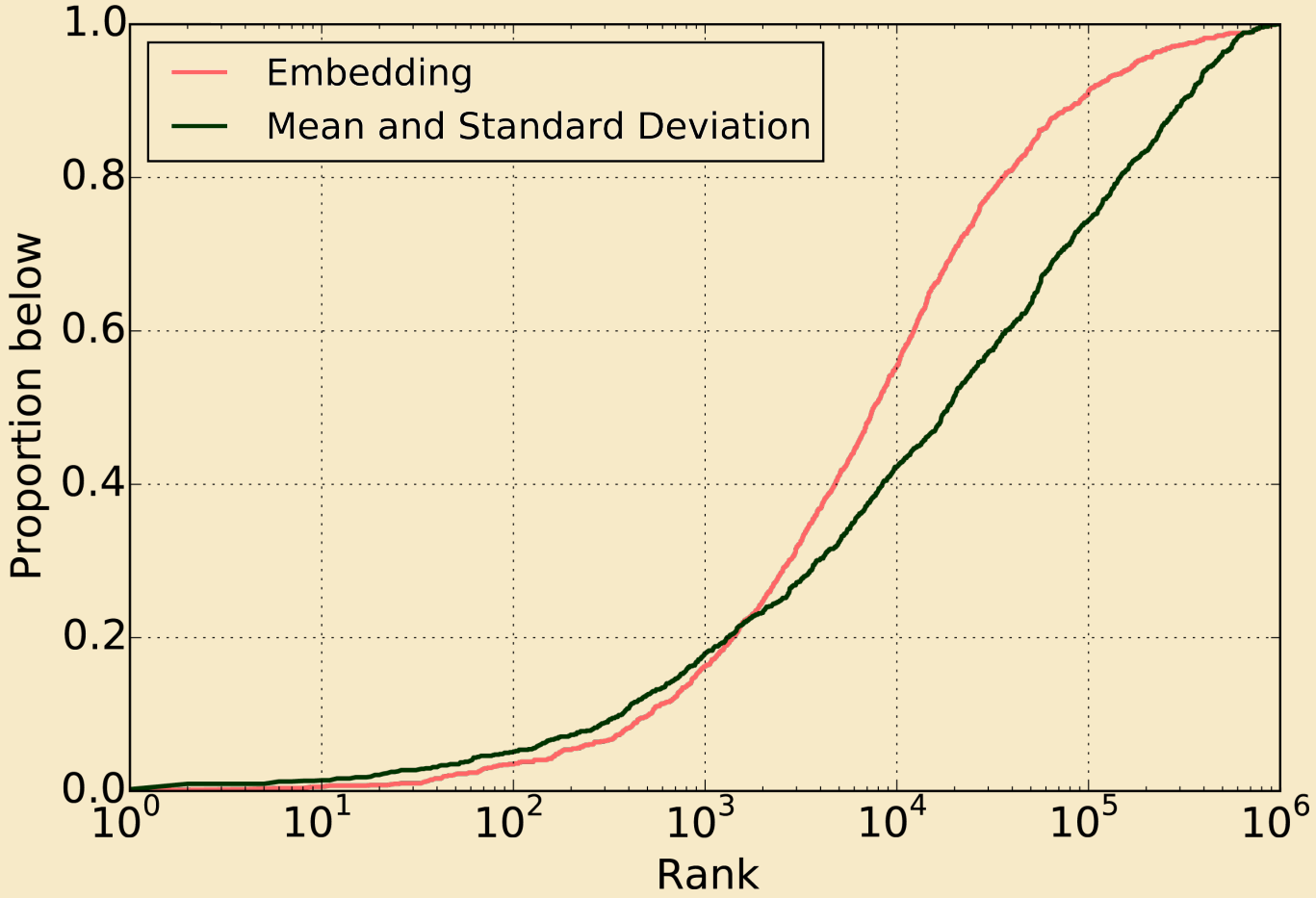
Example Embeddings



Embedding Distance Matrix



MIDI-to-MSD Matching



References

1. *"Large-Scale Content-Based Matching of MIDI and Audio Files"*, 16th International Society for Music Information Retrieval Conference, 2015
2. *"Pruning Subsequence Search with Attention-Based Embedding"*, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing
3. *"Accelerating Multimodal Sequence Retrieval with Convolutional Networks"*, NIPS Multimodal Machine Learning Workshop, 2015
4. *"Feedforward Networks with Attention Can Solve Some Long-Term Memory Problems"*, in preparation

`http://github.com/craffel/midi-dataset`
`http://github.com/craffel/sequence-embedding`

`craffel@gmail.com`