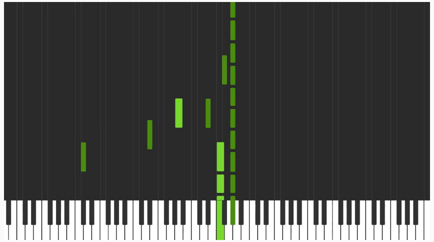
Midi2Hands

Hand placement for automatic piano transcription

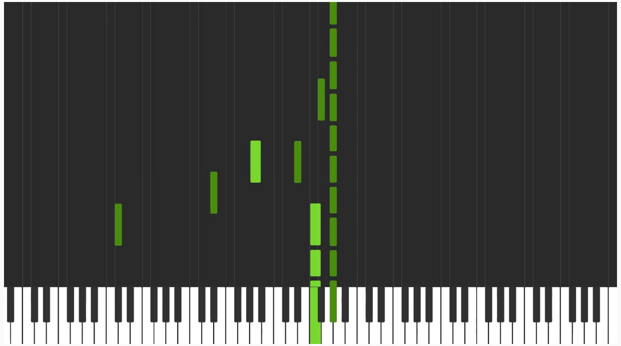
Automatic Music Transcription

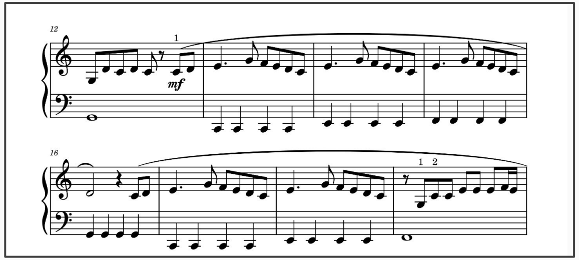
Audio (mp3)

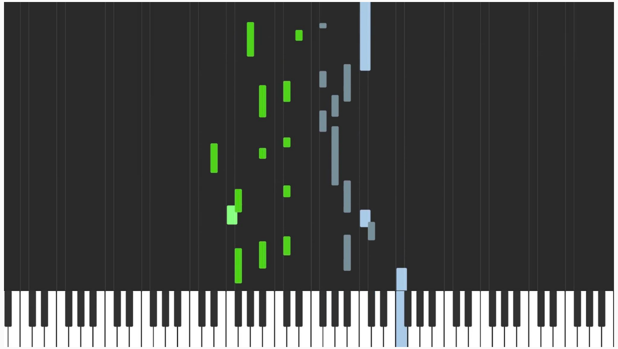


MIDI

Hand Assignment







Dataset and research questions

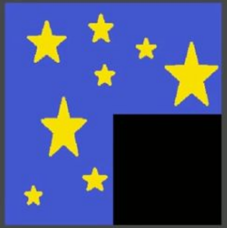
Dataset

- 122 piano performances in midi format (350k) events
- One keyboard for each hand
- 10-fold cross validation

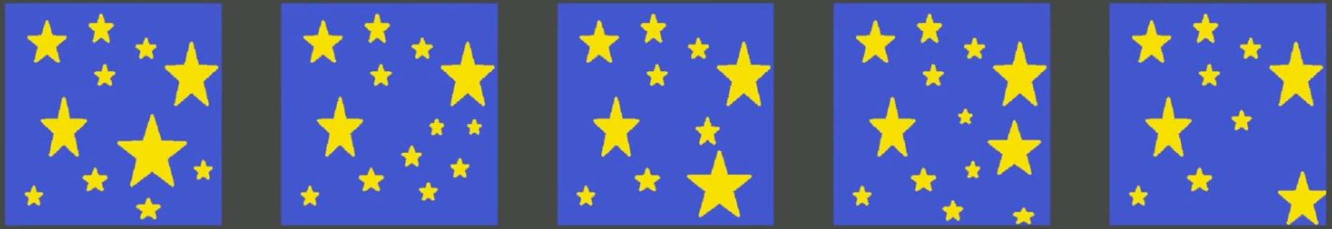
Research questions

1. Can the performance be improved by using a generative model instead of a discriminatory model?
2. How does the window size affect performance?

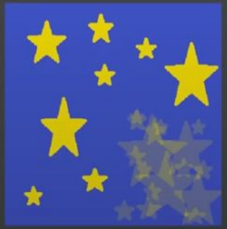
Input



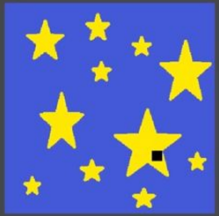
Plausible Completions



Model Output

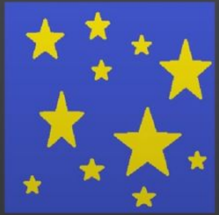


Input



Plausible Completions

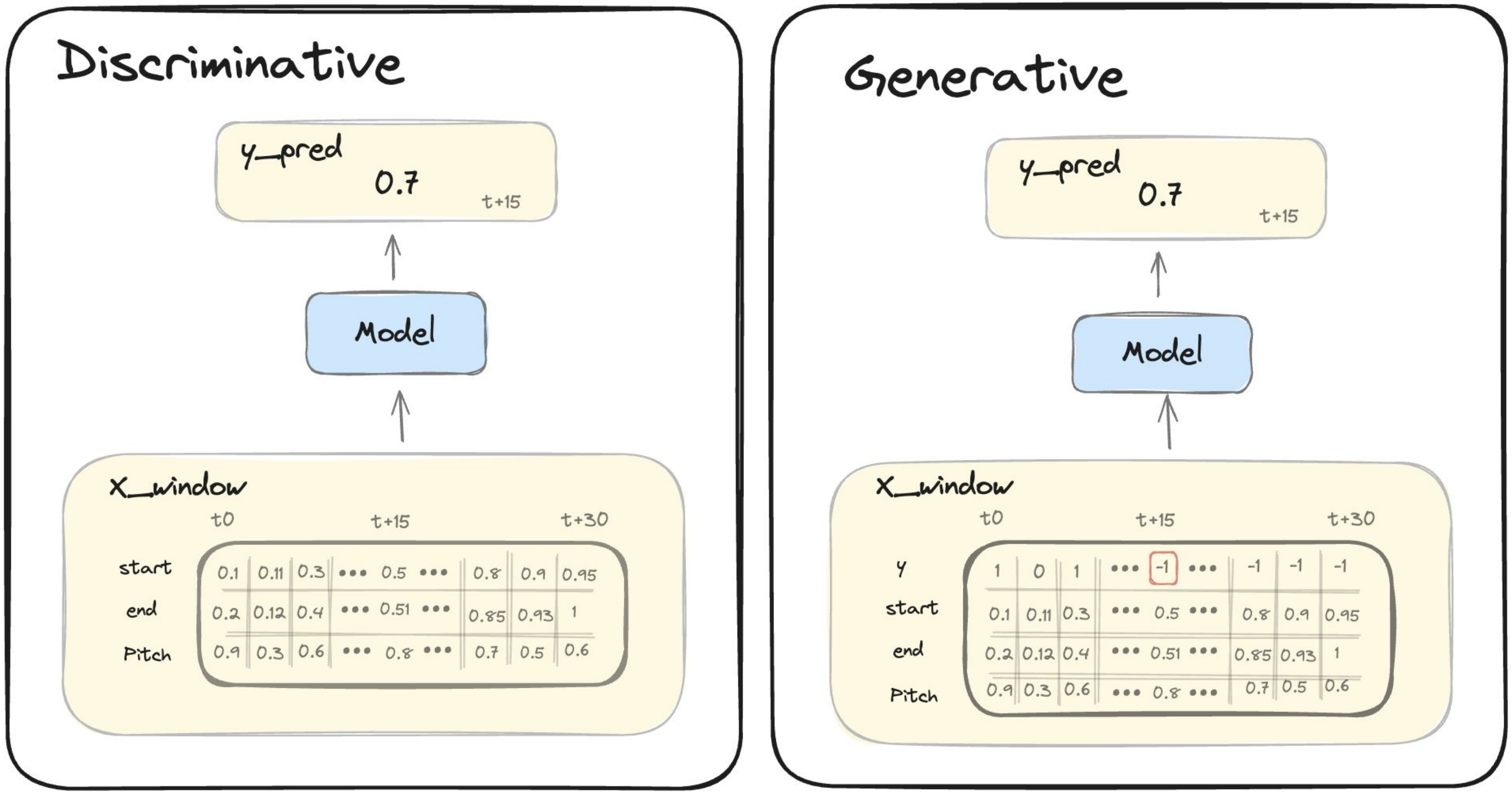
Model Output



$$P(X) = y$$

Discriminatory or generative modeling

$$P(X,Y) = y$$



Discriminative

X_window

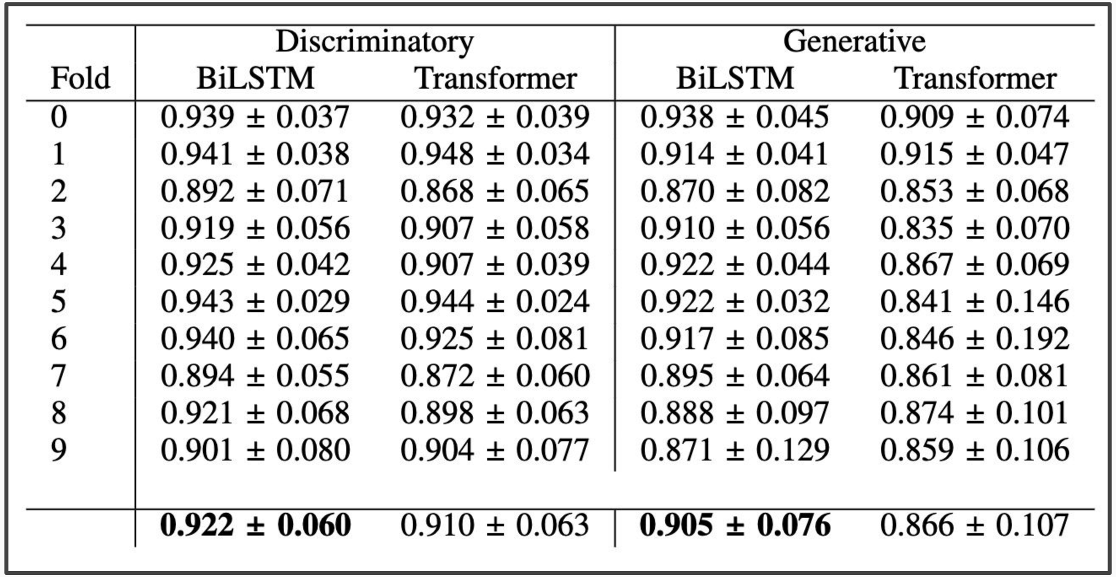
	t0			t+15			t+30
start	0.1	0.11	0.3	••• 0.5 •••	0.8	0.9	0.95
end	0.2	0.12	0.4	••• 0.51 •••	0.85	0.93	1
Pitch	0.9	0.3	0.6	••• 0.8 •••	0.7	0.5	0.6

Generative

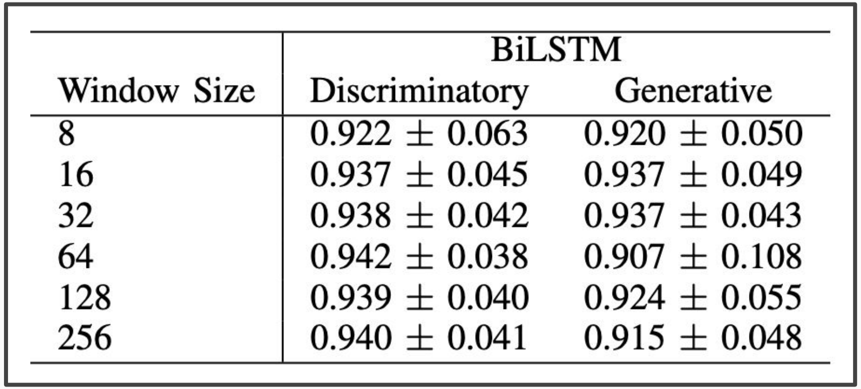
X_window

	t0			t+15			t+30
y	1	0	1	••• -1 •••	-1	-1	-1
start	0.1	0.11	0.3	••• 0.5 •••	0.8	0.9	0.95
end	0.2	0.12	0.4	••• 0.51 •••	0.85	0.93	1
Pitch	0.9	0.3	0.6	••• 0.8 •••	0.7	0.5	0.6

Results



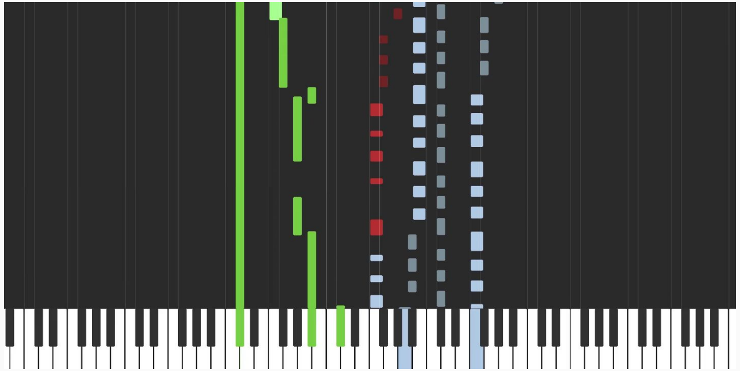
Fold	Discriminatory		Generative	
	BiLSTM	Transformer	BiLSTM	Transformer
0	0.939 ± 0.037	0.932 ± 0.039	0.938 ± 0.045	0.909 ± 0.074
1	0.941 ± 0.038	0.948 ± 0.034	0.914 ± 0.041	0.915 ± 0.047
2	0.892 ± 0.071	0.868 ± 0.065	0.870 ± 0.082	0.853 ± 0.068
3	0.919 ± 0.056	0.907 ± 0.058	0.910 ± 0.056	0.835 ± 0.070
4	0.925 ± 0.042	0.907 ± 0.039	0.922 ± 0.044	0.867 ± 0.069
5	0.943 ± 0.029	0.944 ± 0.024	0.922 ± 0.032	0.841 ± 0.146
6	0.940 ± 0.065	0.925 ± 0.081	0.917 ± 0.085	0.846 ± 0.192
7	0.894 ± 0.055	0.872 ± 0.060	0.895 ± 0.064	0.861 ± 0.081
8	0.921 ± 0.068	0.898 ± 0.063	0.888 ± 0.097	0.874 ± 0.101
9	0.901 ± 0.080	0.904 ± 0.077	0.871 ± 0.129	0.859 ± 0.106
	0.922 ± 0.060	0.910 ± 0.063	**0.905 ± 0.076**	0.866 ± 0.107

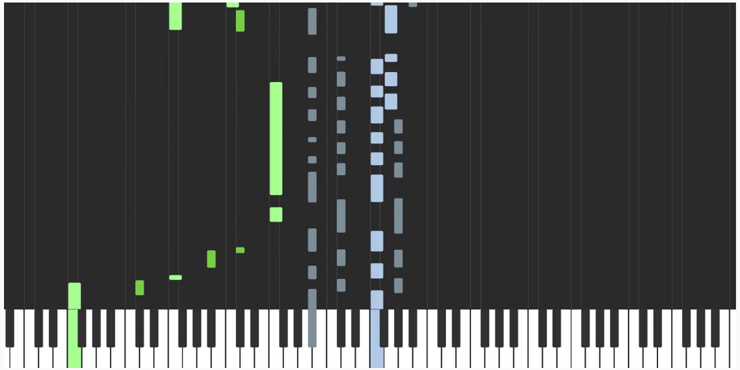


Window Size	BiLSTM	
	Discriminatory	Generative
8	0.922 ± 0.063	0.920 ± 0.050
16	0.937 ± 0.045	0.937 ± 0.049
32	0.938 ± 0.042	0.937 ± 0.043
64	0.942 ± 0.038	0.907 ± 0.108
128	0.939 ± 0.040	0.924 ± 0.055
256	0.940 ± 0.041	0.915 ± 0.048

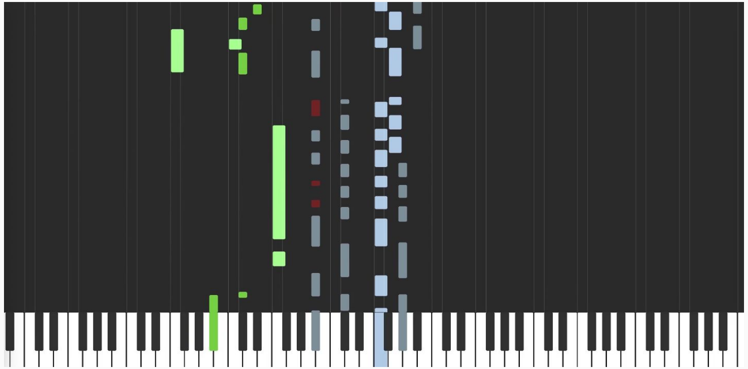
Typical errors

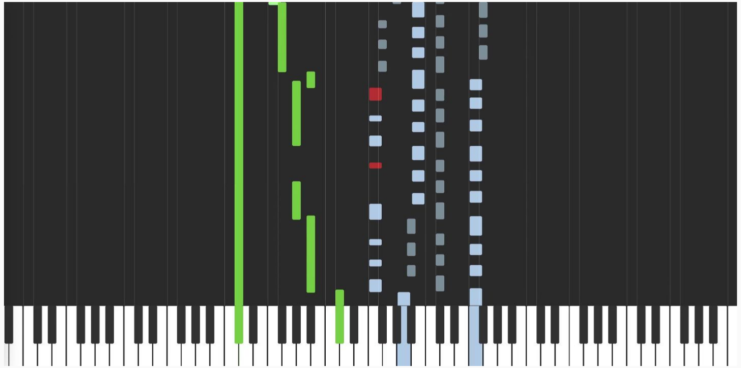
Generative





Discriminatory





Conclusions

We need **larger and more complete datasets** to use accuracy as a fair metric when comparing discriminatory and generative model although generative models show promising results.

Additionally a **more comprehensive study** of the interplay of window size and model architecture need to be conducted

References

1. **Automatic Piano Transcription with Hierarchical Frequency-Time Transformer**
    - Keisuke Toyama, Taketo Akama, Yukara Ikemiya, Yuhta Takida, Wei-Hsiang Liao, Yuki Mitsufuji

2. **High-resolution Piano Transcription with Pedals by Regressing Onset and Offset Times**
    - Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, Yuxuan Wang

3. **Detecting Hands in Piano MIDI Data**
    - Aristotelis Hadjakos, Simon Waloschek, Alexander Leemhuis

4. **Sequence-to-Sequence Piano Transcription with Transformers**
    - Curtis Hawthorne, Ian Simon, Rigel Swavely, Ethan Manilow, Jesse Engel