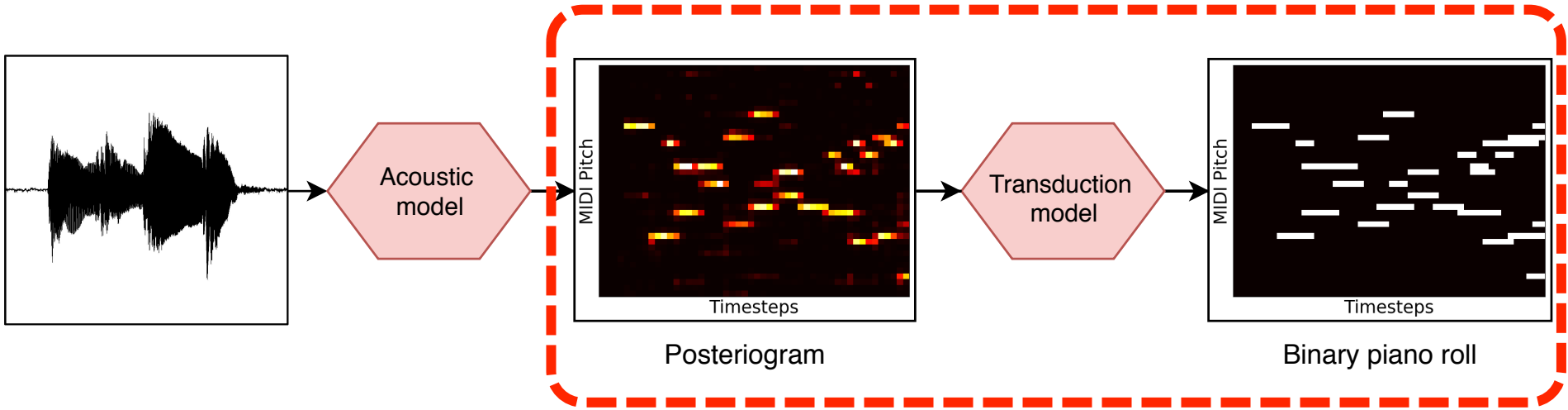
Queen Mary University of London

centre for digital music

A Comparative Study of Neural Models for Polyphonic Music Sequence Transduction

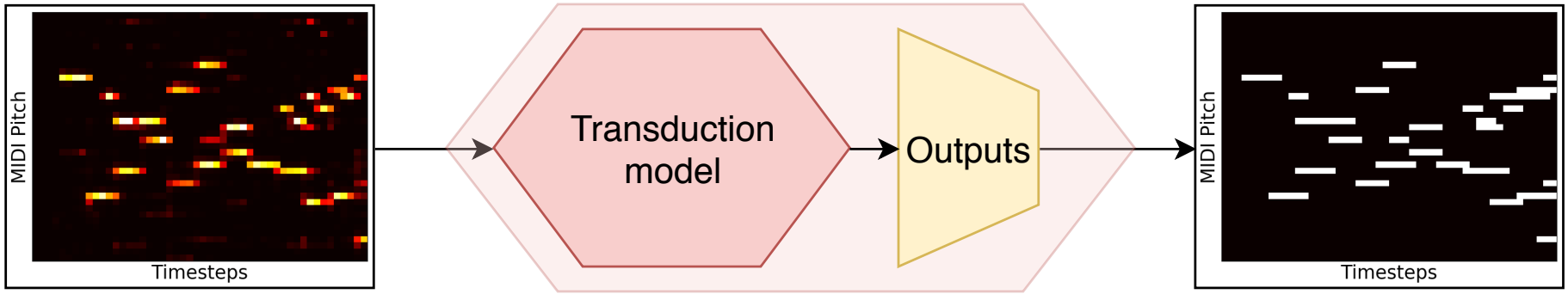
Adrien Ycart, **Daniel Stoller**, Emmanouil Benetos

Automatic Music Transcription (AMT)



- Objective: obtain a binary piano roll from audio
- Intermediate step: obtain a non-binary posteriogram
- We compare various neural-network approaches to learn a mapping between posteriogram and piano-roll

Comparison: 16 configurations

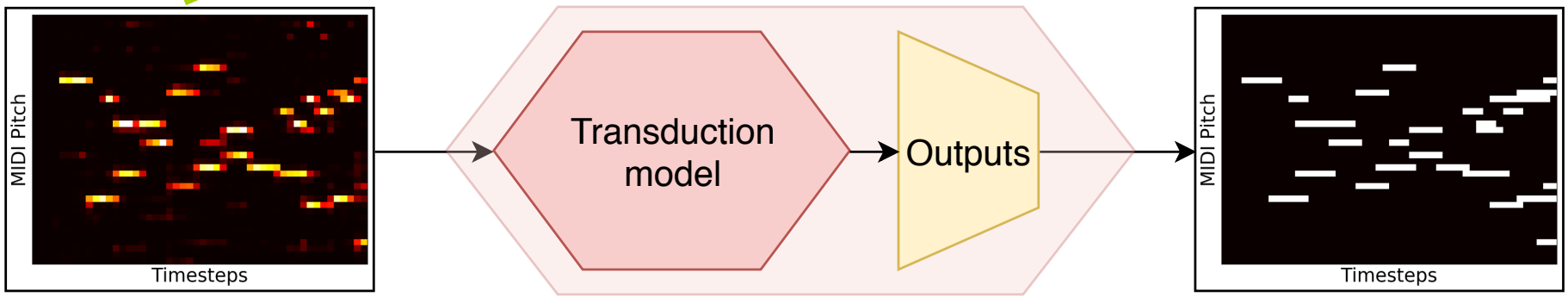


Posteriogram

Binary piano roll

Comparison: 16 configurations

- Acoustic models:
  - *Kelz et al. (2016):* piano-specific CNN
  - *Bittner et al. (2017)*: general-purpose multi-pitch detection CNN



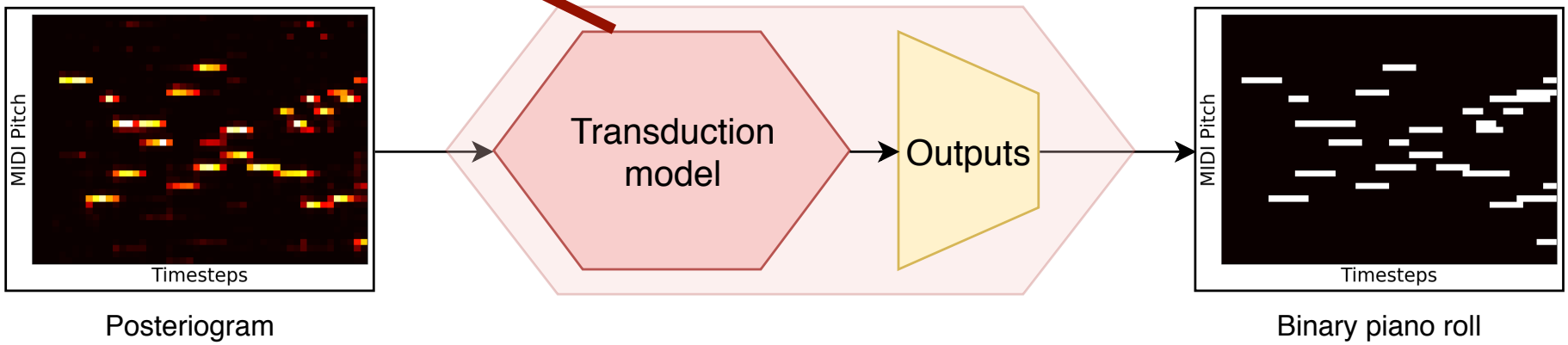
Posteriogram

Binary piano roll

Comparison: 16 configurations

- Transduction models:
  - *Ycart et al. (2018)*: LSTM
  - Newly-proposed CNN

} Same number of parameters

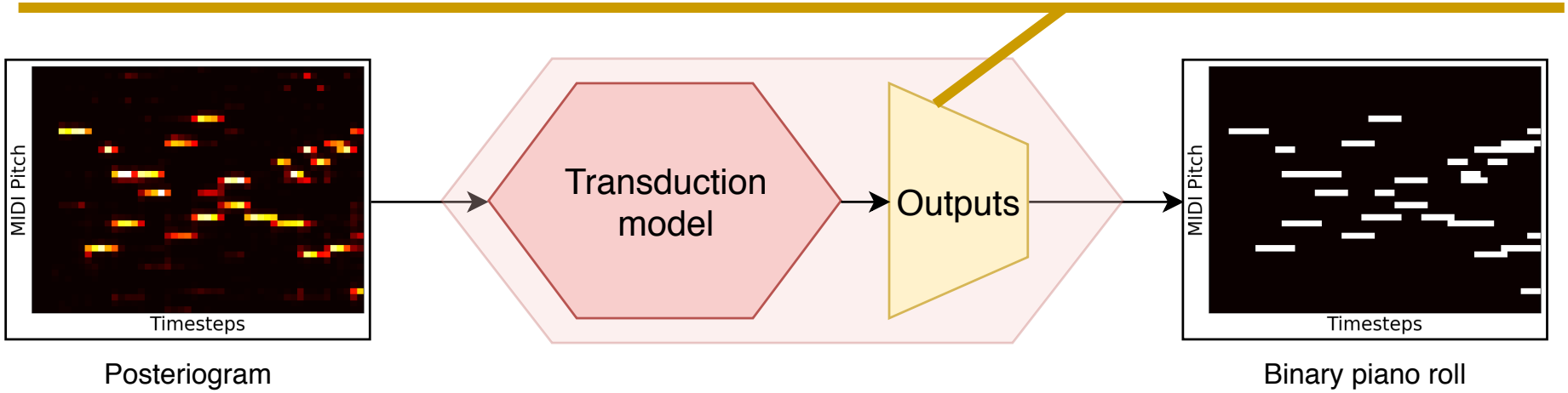


Posteriogram

Binary piano roll

Comparison: 16 configurations

- Outputs:
  - Sigmoid outputs + threshold post-processing
  - *Dong et al. (2018)*: Binary neurons

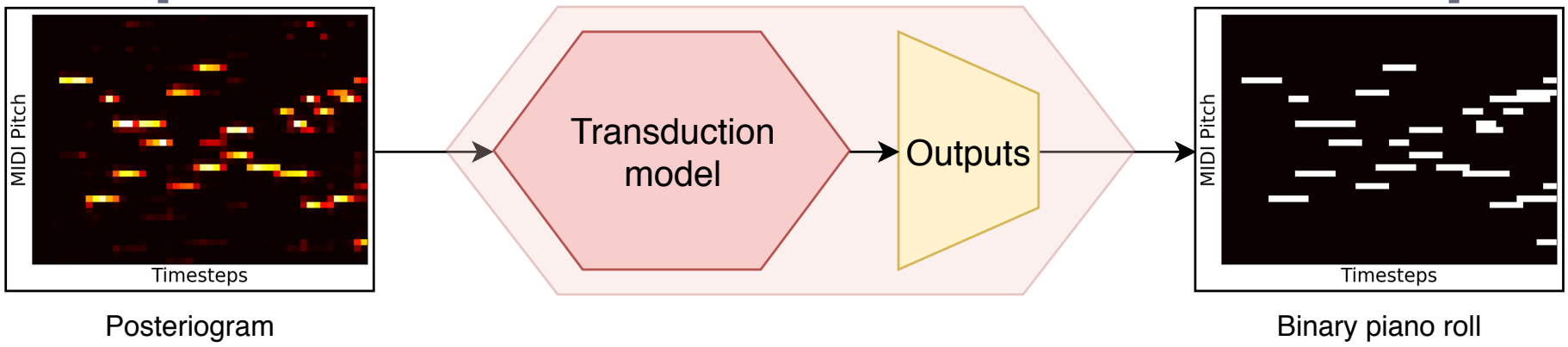


Posteriogram

Binary piano roll

Comparison: 16 configurations

- Training loss:
  - Frame-based: Cross-entropy (non-binary) / F-measure (binary)
  - Adversarial (CNN conditional discriminator)



Posteriogram

Binary piano roll

Main Results

- Overall results:
  - Best-performing model with both acoustic models:
    CNN, sigmoid outputs, cross-entropy loss
  - Outperforms 2 baselines: thresholding / HMM decoding *(Poliner et al., 2016)*
- Transduction model:
  - LSTM often performs worse than simple thresholding
    → Overfitting on specific pianos in the training set
- Training loss and outputs:
  - Training with GAN loss is worse than cross-entropy
  - Binary neurons generally do not improve results with GAN loss
  - F-measure loss similar or slightly worse than cross-entropy