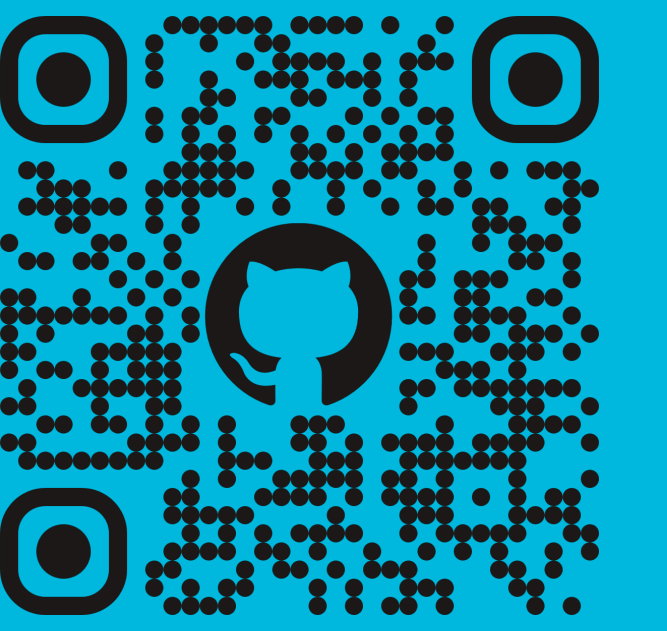




TL;DR

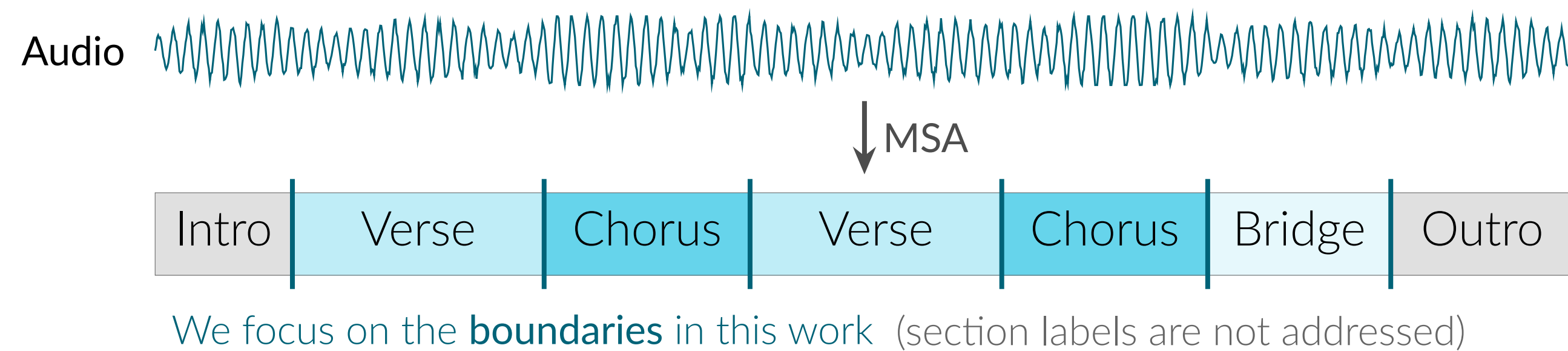
- **Training-free MSA:** generic, frozen deep audio embeddings + unsupervised segmentation (at the bar scale). No annotation needed.

- **Deep models > spectrograms, but not always:** MATPAC++ [1] is best; $\sim 1/3$ of the models fall below a Log-Mel baseline. **CBM** [2] is the best segmenter (33/36).

- **A case for trimming as default:** untrimmed scores are inflated by trivial first/last boundaries $\Rightarrow$ **adopt** (double) **trimming**.

Music Structure Analysis (MSA)

Goal: partition a song into non-overlapping sections [3] (here: *only boundaries*). State-of-the-art: **Supervised deep learning models** (All-in-One [4]).



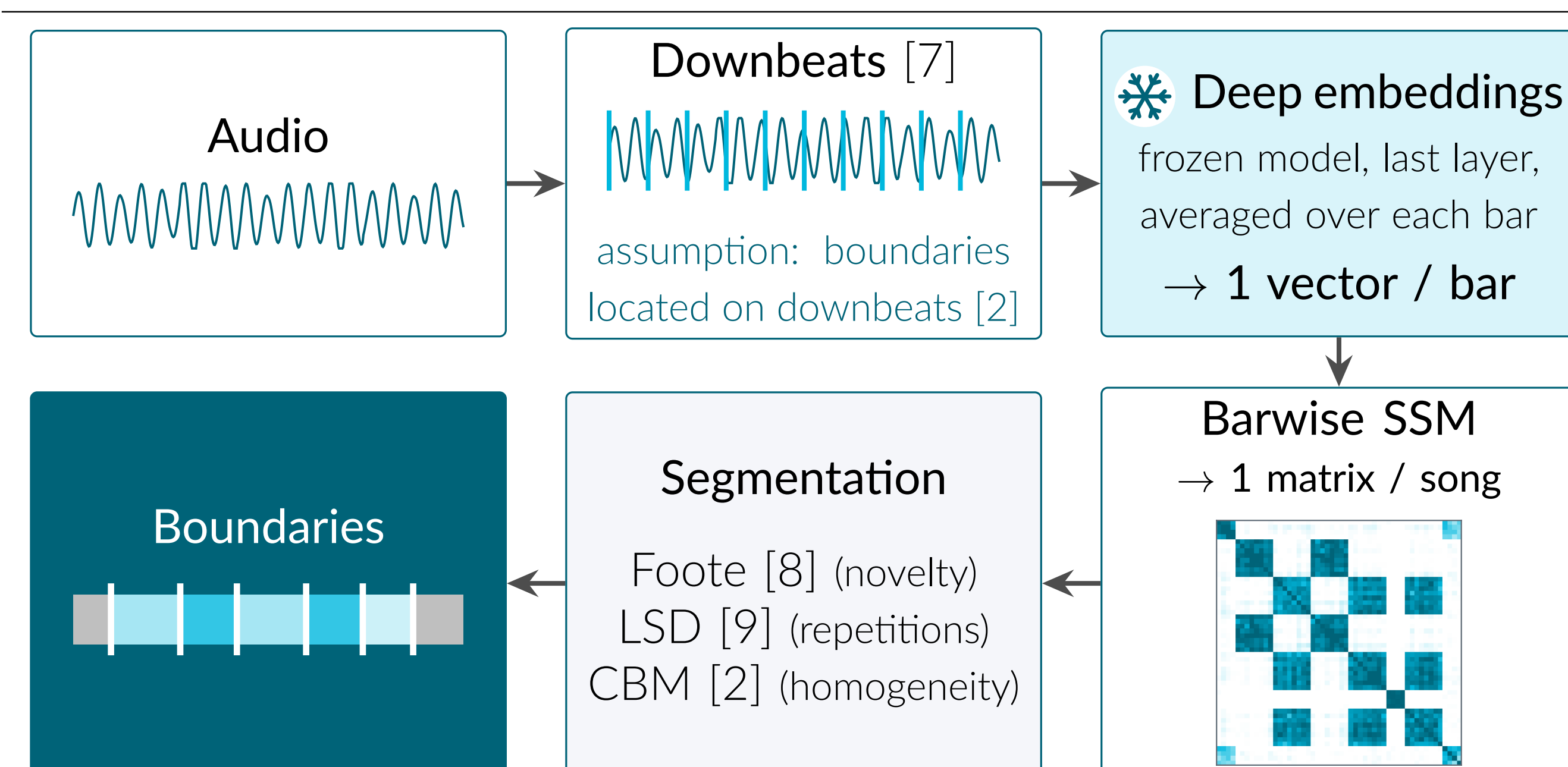
2 problems with supervision: 1) requires costly annotations and 2) structure is **ambiguous**: one annotation discards other valid interpretations.
 $\Rightarrow$ We address the task using **un-/self-supervised** learning!

Research Question

Do generic audio models encode structural information from pre-training?

- $\rightarrow$ MSA-dedicated SSL works [5], but needs a tailored model and training.
- $\rightarrow$ Toyama *et al.* [6] evaluate generic models with a **linear probe**: still *supervised*.
- $\rightarrow$ **Here:** frozen embeddings + strictly **unsupervised** segmentation.

Workflow



Protocol

- **Datasets:** RWC-Pop (100 songs), SALAMI (884), Harmonix (912).
- **Metrics:** boundary hit-rate F-measure at 0.5 s and 3 s ($F_{0.5s}$, F_{3s}).
- **Baseline:** same pipeline with Barwise TF [2] (Log-Mel per bar).
- **Hyperparameters:** grid search; best *per dataset* or one setting *across datasets*.

Deep Audio Models

9 open-source models, 12 embedding spaces.

Pre-training	Models
Masked acoustic modeling	AudioMAE, MERT*, MusicFM*, MuQ*
Masked latent prediction	M2D [†] , MATPAC++ [1]
Compression autoencoding (codecs)	DAC, CoDiCodec [‡]
Audio-text contrastive	CLAP

* trained on music only. [†] audio & multimodal spaces. [‡] continuous & discrete spaces.

Downstream Segmentation Algorithms

Algorithm	Principle
Foote [8]	Checkerboard kernel along the diagonal
LSD [9]	Laplacian eigenvectors + k -means
CBM [2]	Dynamic programming on similarity blocks

Which Embeddings Encode Structure?

	$F_{0.5s}$ (%)			F_{3s} (%)		
MATPAC++	55.30^C	69.26^C	44.04^C	73.79^C	82.13^C	59.47^C
M2D (multimodal)	51.27 ^C	65.14 ^C	41.78 ^C	71.14 ^C	80.79 ^C	58.27 ^C
CoDiCodec (continuous)	53.00 ^C	67.10 ^C	39.98 ^C	72.64 ^C	80.34 ^C	55.30 ^C
M2D (audio)	52.05 ^C	64.75 ^C	38.65 ^C	70.48 ^C	79.00 ^C	53.65 ^C
MuQ (audio)	52.31 ^C	65.92 ^C	38.07 ^F	71.04 ^C	78.17 ^C	52.36 ^F
MERT	52.43 ^C	64.53 ^C	39.81 ^C	70.11 ^C	77.14 ^C	53.31 ^C
CLAP	49.32 ^C	61.60 ^C	40.28 ^C	69.23 ^C	77.84 ^C	56.69 ^C
CoDiCodec (discrete)	50.28 ^C	62.78 ^C	38.75 ^C	70.74 ^C	78.04 ^C	54.30 ^C
Baseline (Barwise TF)	46.95 ^F	61.48 ^C	40.02 ^C	72.39 ^F	76.14 ^C	55.98 ^C
MusicFM	49.88 ^F	62.71 ^C	38.48 ^F	72.23 ^F	75.45 ^C	52.52 ^F
MuQ (multimodal)	45.41 ^C	62.46 ^C	39.13 ^C	67.94 ^C	79.34 ^C	55.77 ^C
AudioMAE	31.75 ^C	38.62 ^C	32.32 ^C	56.92 ^C	59.63 ^C	44.62 ^C
DAC	24.55 ^C	34.37 ^C	31.83 ^C	47.69 ^C	54.43 ^C	43.34 ^C
	Harmonix	RWC-Pop	SALAMI	Harmonix	RWC-Pop	SALAMI

Figure 1. Best score per embedding (%), per-dataset hyperparameters. Superscript: best algorithm (CBM / Foote).

- $\rightarrow$ **Masked latent prediction** wins; **discrete spaces** lose; **multimodal**: mixed.
- $\rightarrow$ **CBM** is the best segmenter in **33/36** conditions.

Comparison with the Literature

F_{3s} (%)	Harmonix	RWC-Pop	SALAMI
MATPAC++ (ours)	73.79	82.13	53.74
Barwise TF [2] [§]	68.84	71.27	55.98
<i>Linear probing [6] $\rightarrow$ ours (same model, Harmonix; CLAP & DAC not shown)</i>			
AudioMAE	58.11 $\rightarrow$ 56.92		
MERT	60.99 $\rightarrow$ 70.11		
MusicFM	63.91 $\rightarrow$ 70.52		
MSA-specific SSL [5] [‡]	80.8	–	70.1
All-In-One [4] (supervised) [§]	n/a (train/test leakage)	80.07	60.06

Ours: CBM, cross-dataset hyperparameters. [§] recomputed here. [‡] trimmed, oracle level.

- $\rightarrow$ Our pipeline **beats linear probing** for 4 of the 5 common models.
- $\rightarrow$ MATPAC++ **matches All-In-One on RWC-Pop**, without supervision.

Trimming: Towards Rigorous Evaluation

Trimming: drop first & last segments. **Double:** also drop edge silences.

	RWC-Pop			SALAMI		
F_{3s} (%)	None	Trim	Double	None	Trim	Double
MATPAC++	82.13	79.87	77.88	53.74	47.12	43.33
Barwise TF	71.27	67.59	64.08	55.98	47.65	42.38
All-In-One	80.07	77.27	74.61	60.06	52.23	46.73

$\Rightarrow$ Up to **–13.6 points**: trimming should become the **default**.

Conclusion & Future Work

- Generic deep audio models do encode structural information! $\rightarrow$ Barwise processing & CBM into MSA-specific SSL models.
- Trimmed scores should become the community standard. $\rightarrow$ Extend the current framework to segment labelling.

References

- [1] A. Quenec et al., "MATPAC++: Enhanced masked latent prediction for self-supervised audio representation learning," *arXiv preprint arXiv:2508.12709*, 2025.
- [2] A. Marmoret et al., "Barwise music structure analysis with the correlation block-matching segmentation algorithm," *TISMIR*, 2023.
- [3] O. Nieto et al., "Audio-based music structure analysis: Current trends, open challenges, and applications," *TISMIR*, 2020.
- [4] T. Kim and J. Nam, "All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio," in *WASPAA*, 2023.
- [5] M. Buisson et al., "Self-supervised learning of multi-level audio representations for music segmentation," *IEEE/ACM Trans. Audio, Speech and Language Processing*, 2024.
- [6] K. Toyama et al., "Do foundational audio encoders understand music structure?," in *ICASSP*, 2026.
- [7] F. Foscarin et al., "Beat this! accurate beat tracking without DBN postprocessing," in *ISMIR*, 2024.
- [8] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *ICME*, IEEE, 2000.
- [9] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering," in *ISMIR*, 2014.