Appendix A Evaluation Protocol Details

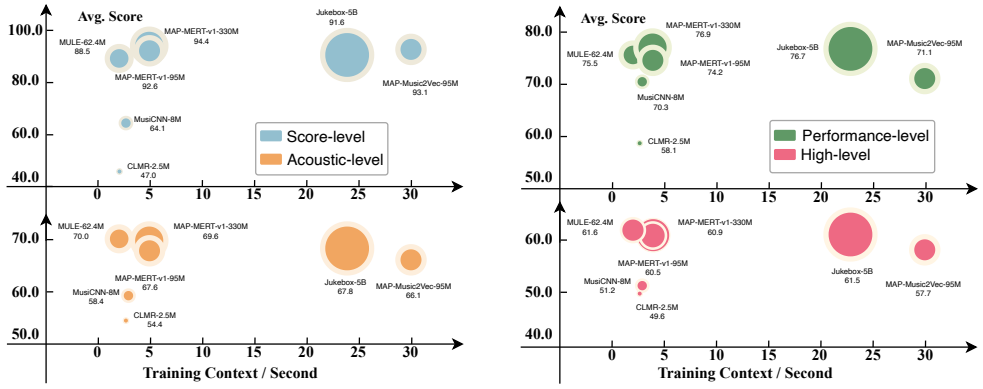
The hyper-parameter search range of the constrained evaluation track is given as follow:

1. **Layer**: {every single layer, weighted sum}
2. **Model**: {one-layer 512-units MLP, one-layer 512-unit LSTM (melody extraction only), 3-layer 512-unit LSTM (source separation only), 3-encoder-3-decoder layers transformer (lyrics transcription only)}
3. **Batch size**: {64}
4. **Learning rate**: {5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}
5. **Dropout probability**: {0.2}

Appendix B Detail Analysis

What have the music audio pre-trained representations learned? We observe that all the representations have learned multiple levels of knowledge in Fig. 1. Most of the selected baselines are particularly good at high-level music description tasks, such as genre classification and emotion recognition. However, when pre-trained with a full supervision paradigm, the representations may not be able to model pitch and key well, as they could overfit the supervision signal less relevant to pitch-related information. On the contrary, SSL methods usually mitigate this issue by providing more generalisable representations. Some representations do not support frame-level representations, which makes it difficult to evaluate their performance on tasks such as source-separation and beat tracking. Therefore, it is unclear how well these models have learned such information.

How can we design better pre-training strategies for music audio representation learning? As mentioned in the above paragraph, we suggest that a good pre-training strategy needs to prevent overfitting the supervision signal, which makes self-supervised learning a more promising approach. Moreover, we argue that an optimal method for music pre-training should be able to scale up to larger data and model size. Based on observations from Figure 2, it appears that larger data and model size have a greater impact on performance than the training paradigm (generative, contrastive, or mask prediction) at the current stage of research. Besides, stacked transformer models are good candidates for future pre-training architecture, as they can be easily scaled up, and usually provide frame-level representations in a well-considered design.



(a) Scores at Acoustic-level and Score-level.

(b) Scores at Performance-level and High-level.

Figure 3: Results Analysis Regarding Training Context Length. The performances of *source separation* and *beat tracking* tasks are ignored similar to Fig. 2.

How does context length affect performance? According to Fig. 3, the relationship between context length and performance exhibits a rather complex and irregular pattern, for which it is currently difficult to draw any conclusive insights. This is due to the limited number of music audio representations available at the moment, coupled with challenges in controlling variables. However, we are able to derive some preliminary observations when considering factors such as data size (D)

and parameter size (N). We observe that within a context length (L) of approximately 3 to 5 seconds, scaling up N and D can be effective, but the performance quickly saturates. Furthermore, according to MAP-Music2Vec-95M, solely increasing the L without scaling the N and D may also lead to performance saturation. Interestingly, when scaling up all three aspects, according to Jukebox-5B with 23 seconds context and 60~120khr data, the performance still saturates. The underlying cause of this saturation may be associated with the training paradigm.

Appendix C Website and Leaderboard

To accompany the MARBLE benchmark with leaderboard data and detailed resources presentation, we build a website, which can be found at https://marble-bm.shef.ac.uk. All the resources and comprehensible introduction of the benchmark and submission guideline are indexed on the homepage as shown in Fig. 4. The participants can easily find the process of submitting their results according to the guidelines. As demonstrated in Fig. 5, we provide a well-organised leaderboard for MARBLE, where the evaluated results can be re-ranked according to different metrics and filtered by tasks.

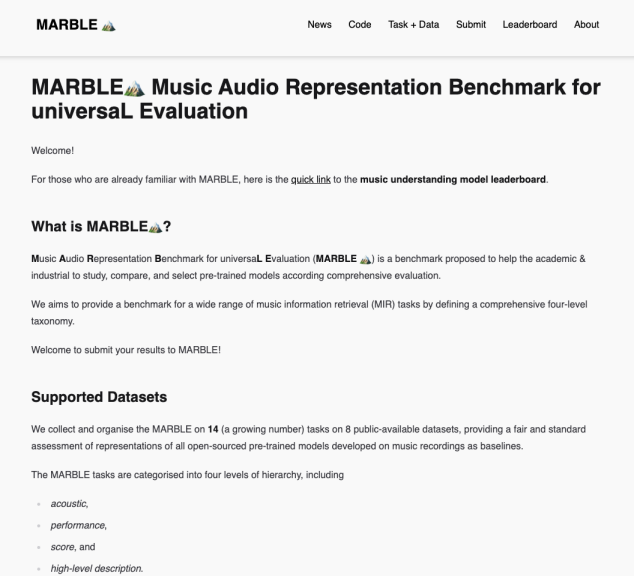


Figure 4: Website for the Proposed MARBLE Benchmark.

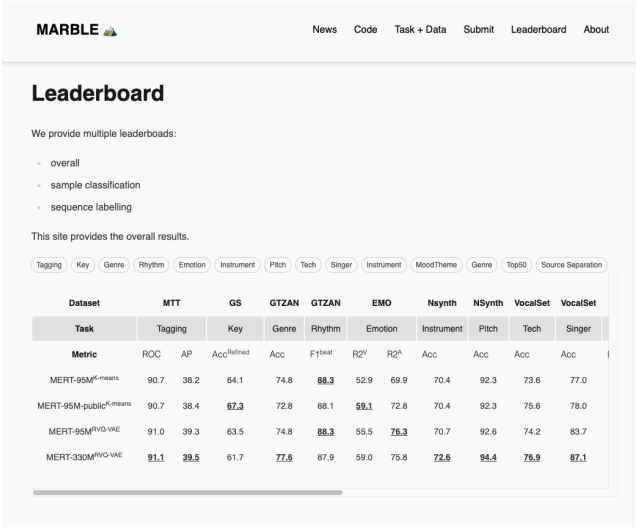


Figure 5: Music Understanding Model Leaderboard Hosted on the MARBLE Website.

Appendix D Details on Chord Estimation

D.1 Chord Vocabulary

Our chord vocabulary includes "none" and 35 different chords on each of the 12 root notes, 421 in total. The root notes are listed as follows: {C Db D Eb E F Gb G Ab A Bb B}. We do not distinguish between equal notes under the twelve equal temperaments. For example, we think that C# and Db are the same note and have the essentially equivalent function in chord prediction. We use sharp in the code implementation for identification but use flat in the following tables.

The following Tables of the 35 types of chords with examples and a number of samples in the datasets.

chord name	maj	min	aug	maj6	min6	7	maj7	min7	dim7	hdim7
example	C:maj	C:min	C:aug	C:maj6	C:min6	C:7	C:maj7	C:min7	C:dim7	C:hdim7
chord tones	1,3,5	1,b3,5	1,3,5	1,3,5,6	1,b3,5,7	1,3,5,b7	1,3,5,7	1,b3,5,b7	1,b3,b5,bb7	1,b3,b5,b7
chord number	1120	368	16	70	12	374	292	204	2	106

chord name	9	maj9	min9	11	sus2	sus4	maj/3	maj/5	min/b3	min/5	7/3	7/5	7/b7
example	C:9	C:maj9	C:min9	C:11	C:sus2	C:sus4	C:maj/3	C:maj/5	C:min/b3	C:min/5	C:7/3	C:7/5	C:7/b7
chord tones	1,3,5,b7,9	1,3,5,7,9	1,b3,5,b7,9	1,3,5,b7,9,11	1,2,5	1,4,5	3,5,1	5,1,3	b3,5,1	5,1,b3	3,5,b7,1	5,b7,1,3	b7,1,3,5
chord number	78	22	48	8	88	44	82	264	10	82	10	44	46

chord name	maj7/3	maj7/5	maj7/7	min7/b3	min7/5	min7/b7	dim7/b3	dim7/b5	dim7/bb7	hdim7/b3	hdim7/b5	hdim7/b7	N
example	C:maj7/3	C:maj7/5	C:maj7/7	C:min7/b3	C:min7/5	C:min7/b7	C:dim7/b3	C:dim7/b5	C:dim7/bb7	C:hdim7/b3	C:hdim7/b5	C:hdim7/b7	No chord
chord tones	3,5,7,1	5,7,1,3	7,1,3,5	b3,5,b7,1	5,b7,1,b3	b7,1,b3,5	b3,b5,bb7,1	b5,bb7,1,b3	bb7,1,b3,b5	b3,b5,b7,1	b5,b7,1,b3	b7,1,b3,b5	No chord
chord number	6	66	14	6	30	42	0	0	0	0	6	2	

These are some special or rare chords in the dataset and we use some Chord Substitutions based on similar chords or the chord annotation for the music score instead of the ground truth chord annotation the musician actually plays.

1. **majmin7** was substituted with **7**: The "majmin7" chord is equivalent to the "7" chord, so we are making a replacement to standardize the notation.
2. **minmaj7** was substituted with **min7**: Both chords share the root, minor third, and perfect fifth. When mapping minmaj7 to min7, the major seventh is altered to a minor seventh, ensuring the "minor" character of both chords remains consistent.
3. **min11** was substituted with **11**: Both chords are minor chords composed of the seventh and eleventh tones. Given their infrequent occurrences, we map "min11" to the "11" chord.
4. **Substitution for out-of-vocabulary colour chords**: The performed chord annotations in the GuitarSet also contain out-of-vocabulary colour chords such as (1,5)/1, (1,5,b7)/1, (5,2,b7,4)/4. For such chords, we identify the corresponding standard chords in the instructed chord annotations and substitute them.
5. **Special Transposition Handling for Standard Chords**: Map to the standard transposition that is closest to the corresponding transposed note.

D.2 Chord Recognition Metric Definition

1. **root**: Evaluating chord recognition algorithms based on the root notes of the identified chords. Only compares the root of the chords.
2. **majmin**: Only compares major, minor, and "no chord" labels. Any other chord types or variations, such as 7th chords, augmented, diminished, and so on, are not considered in this specific evaluation.
3. **mirex**: Compare chords along MIREX rules. A estimated chord is considered correct if it shares at least three pitch classes in common.
4. **thirds**: Chords are compared at the level of major or minor thirds (root and third). For example, both ('A:7', 'A:maj') and ('A:min', 'A:dim') are equivalent, as the third is major and minor in quality, respectively.
5. **traids**: Chords are considered at the level of triads (major, minor, augmented, diminished, suspended). In addition to the root, the quality is only considered through #5th scale degree (for augmented chords). For example, ('A:7', 'A:maj') are equivalent, while ('A:min', 'A:dim') and ('A:aug', 'A:maj') are not.

6. **sevenths**: Compares according to MIREX "sevenths" rules. Only major, major seventh, seventh, minor, minor seventh and no chord labels are compared.
7. **majmin_inv**: Compares major/minor chords, with inversions. The bass note must exist in the triad.
8. **sevenths_inv**: Compares according to MIREX "sevenths" rules, with inversions. The bass note must exist in the chord.

During the evaluation process, frame-level predictions are directly merged to event-level by the `mir_eval` function so we do not apply any post-processing to the prediction.

Appendix E Details on Lyrics Transcription

E.1 MulJam2.0 dataset

MulJam2.0 is derived from MulJam, featuring larger and more refined human annotation on the test set. We select 34 songs from the training set and obtain human lyrics annotation to expand the test set. For each language, 20 songs are randomly selected from the original training set to form the validation set. A few songs are excluded due to poor alignment for obtaining the line-level annotations (For details, please refer to [71]). We also exclude the songs in the training and validation sets that were present in Jamendo (3 songs in training and 1 song in validation), ensuring that the songs in the evaluation datasets remain unseen during training. The numbers of songs by language can be found in Tab. 6.

The human annotation is performed at the song level. We applied similar procedures to obtain line-level annotations, as was done for the training set in MulJam. We use the timestamps provided by Whisper [44], and align the lines predicted by Whisper with the human annotation. As in [71] , lines with unusually high character rates (exceeding 37.5 Hz) are removed. However, for the test set we choose not to filter by the similarity between the aligned text pairs, to prevent introducing excessive bias in favor of Whisper predictions.

Table 6: Number of songs in MulJam2.0 and Jamendo datasets.

Dataset	**MulJam2.0**			**Jamendo**
Split	Train	Valid	Test	Test
English (en)	3557	20	28	20
French (fr)	977	19	19	20
Spanish (es)	584	19	13	20
German (de)	107	20	3	20
Italian (it)	278	20	7	-
Russian (ru)	106	16	4	-
Total	5609	114	74	80

E.2 Language Model and Tokenizer

The language model (LM) is trained using a speechbrain [48] language model recipe [12]. The model comprises of 12 transformer encoder layers, with an attention dimension of 768, 12 attention heads, and a position-wise feed-forward layer dimension of 3072. The LM is trained using cross-entropy loss for 20 epochs, and the model with the lowest loss is selected.

The target character set is the union of the character sets from 6 languages, resulting in a total of 91 tokens: ϵ, `<bos>`, `<eos>`, `<unk>`, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, À, Á, Â, Ä, Æ, Ç, È, É, Ê, Ë, Ì, Í, Î, Ï, Ñ, Ò, Ó, Ô, Ö, Ù, Ú, Û, Ü, Œ, Ÿ, È, А, Б, В, Г, Д, Е, Ж, З, И, Й, К, Л, М, Н, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ы, Ь, Э, Ю, Я.

E.3 Training Details

The beam search used for validation and testing incorporates a combination of CTC probabilities, LM probabilities (applied only at test time), and S2S probabilities. We assign a weight of 0.4 to the CTC probabilities and 0.3 to the LM probabilities. During validation, we utilize a beam size of 10 and

[12] https://github.com/speechbrain/speechbrain/blob/develop/recipes/LibriSpeech/LM/hparams/transformer.yaml

calculate Word Error Rate every 5 epochs to optimize processing efficiency. For thorough evaluation, we scale up the beam size to 40 during the testing phase. The accuracy of the S2S branch output is continually monitored to determine whether early stopping should be triggered and to facilitate model selection.

E.4 Results and Discussion

The results of multilingual lyrics transcription using different pretrained features can be found in Tab. 7. In addition to MulJam, we also present WERs on the Multilingual Jamendo evaluation set [13]. This dataset consists of 80 songs in 4 languages: English, French, Spanish, and German. While Italian and Russian songs are not included, Jamendo's human-annotated line-level annotation aligns well with our evaluation setting. For comparison, we reference the state-of-the-art model Whisper [44], a robust model designed for speech recognition but also performs effectively on singing voice. Whisper has been trained on an extensive corpus of multilingual and multitask supervised data collected from the internet. It is also the foundation of the MulJam dataset.

Lyrics transcription is a challenging task that involves detecting vocal pronunciations in the presence of background music and making the most probable predictions based on linguistic knowledge. The multilingual context makes this task even more demanding. When performing lyrics transcription with SSL features, it is essential that these features capture clear vocal information, and that the backend provides robust inference to generate coherent text from the vocal pronunciations. Achieving this with SSL features is indeed a significant challenge. The results presented in Table 7 indicate that there is room for improvement in this task.

Among the six languages we considered, English, French, and Spanish, which have a larger number of songs than the other three, yield better results. This suggests that there may be an impact from the imbalanced training data. Russian, on the other hand, produces the worst result for two main reasons: 1. Russian employs the Cyrillic writing system, which has its own set of characters. 2. The training data for Russian is insufficient for the model to establish a connection between the pronunciation rules of Cyrillic and Latin alphabets.

The MulJam test set is human-annotated at the song level but relies on the alignment with Whisper results to derive line-level annotations. Therefore, it is worth noting that bias is introduced, as the alignment is reliable only when the human annotation closely matches the Whisper's prediction.

Table 7: Multilingual lyrics transcription results on MulJam and Jamendo.

Language	**English**		**French**		**Spanish**		**German**		**Italian**		**Russian**		**Whole**	
Metric	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
MulJam2.0 test														
MAP-Music2Vec [31]	54.7	79.2	58.3	90.9	43.2	83.7	63.4	99.5	53.0	91.9	101.6	125.6	56.4	87.8
MAP-MERT-v0-95M [30]	48.7	71.2	55.5	85.4	41.0	80.1	65.9	100.9	49.1	86.3	99.5	124.9	52.6	82.3
MAP-MERT-v0-95M-public [30]	49.0	71.2	55.3	85.4	39.0	76.6	63.5	99.9	50.3	90.3	104.7	129.3	52.5	82.7
MAP-MERT-v1-95M [29]	**45.5**	66.5	52.5	81.9	38.2	73.9	58.8	93.2	44.4	81.6	**96.1**	**117.8**	49.4	77.9
MAP-MERT-v1-330M [29]	**45.5**	**65.9**	**50.7**	**79.6**	**35.9**	**71.9**	**58.3**	**93.1**	**42.4**	**80.3**	100.5	125.5	**48.5**	**77.0**
SOTA [44]	33.2	44.8	52.9	70.1	29.9	43.8	36.5	53.0	38.1	58.5	34.7	53.7	39.5	54.8
Jamendo														
MAP-Music2Vec [31]	49.0	73.6	55.3	87.1	50.3	90.7	67.8	108.8	-	-	-	-	55.7	89.6
MAP-MERT-v0-95M [30]	48.5	71.8	54.0	85.1	49.3	87.6	67.6	108.1	-	-	-	-	54.8	87.6
MAP-MERT-v0-95M-public [30]	46.9	71.5	52.0	81.5	44.8	82.8	66.3	106.8	-	-	-	-	52.6	85.2
MAP-MERT-v1-95M [29]	**43.6**	**67.2**	**49.4**	**79.6**	**43.2**	**80.6**	62.1	103.3	-	-	-	-	**49.6**	**82.2**
MAP-MERT-v1-330M [29]	45.7	68.8	50.2	80.1	44.1	82.8	**61.0**	**102.3**	-	-	-	-	50.3	83.1
SOTA [44]	24.9	39.3	29.2	49.9	21.2	41.7	25.8	46.6	-	-	-	-	25.4	44.4