MARBLE: Music Audio Representation Benchmark for Universal Evaluation

Ruibin Yuan[1,2,3]* **Yinghao Ma**[1,4]* **Yizhi Li**[1,5]* **Ge Zhang**[1,2,6]* **Xingran Chen**[7] **Hanzhi Yin**[3]
Le Zhuo[8] **Yiqi Liu**[5] **Jiawen Huang**[4] **Zeyue Tian**[9] **Binyue Deng**[10] **Ningzhi Wang**[4]
Chenghua Lin[5]† **Emmanouil Benetos**[4] **Anton Ragni**[5] **Norbert Gyenge**[5] **Roger Dannenberg**[3]
Wenhu Chen[6] **Gus Xia**[11,12] **Wei Xue**[9] **Si Liu**[8] **Shi Wang**[13] **Ruibo Liu**[14] **Yike Guo**[9] **Jie Fu**[2]†

[1]Multimodal Art Projection Research Community [2]Beijing Academy of Artificial Intelligence
[3]Carnegie Mellon University [4]Queen Mary University of London [5]University of Sheffield
[6]University of Waterloo [7]University of Michigan Ann Arbor [8]Beihang University
[9]Hong Kong University of Science and Technology [10]Peking University [11]New York University
[12]MBZUAI [13]ICT, Chinese Academy of Sciences [14]Dartmouth College
ruibiny@andrew.cmu.edu yinghao.ma@qmul.ac.uk
{yizhi.li,c.lin}@sheffield.ac.uk gezhang@umich.edu fujie@baai.ac.cn

Abstract

In the era of extensive intersection between art and Artificial Intelligence (AI), such as image generation and fiction co-creation, AI for music remains relatively nascent, particularly in music understanding. This is evident in the limited work on deep music representations, the scarcity of large-scale datasets, and the absence of a universal and community-driven benchmark. To address this issue, we introduce the **M**usic **A**udio **R**epresentation **B**enchmark for universa**L** **E**valuation, termed MARBLE. It aims to provide a benchmark for various Music Information Retrieval (MIR) tasks by defining a comprehensive taxonomy with four hierarchy levels, including acoustic, performance, score, and high-level description. We then establish a unified protocol based on 14 tasks on 8 public-available datasets, providing a fair and standard assessment of representations of all open-sourced pre-trained models developed on music recordings as baselines. Besides, MARBLE offers an easy-to-use, extendable, and reproducible suite for the community, with a clear statement on copyright issues on datasets. Results suggest recently proposed large-scale pre-trained musical language models perform the best in most tasks, with room for further improvement. The leaderboard and toolkit repository are published[3] to promote future music AI research.

1 Introduction

Despite Artificial Intelligence (AI) rapid advancement in the field of art, it has not yet made significant progress in music, particularly in music understanding. To address this, researchers are studying the interdisciplinary field of Music Information Retrieval (MIR) to develop a general music understanding model. MIR focuses on automatically extracting information from raw music audio [33], which enables a variety of tasks such as music classification, emotion recognition, pitch estimation, and the analysis of musical features such as rhythm, melody, and harmony. Due to issues such as copyright and annotation costs, labelled music datasets are usually small, which limits the performances of supervised models. Given that self-supervised learning (SSL) is useful for various tasks (*e.g.*, NLP [17, 13, 41] and CV [34]) with limited annotated datasets, there have been works on SSL-based audio

---

*The authors contributed equally to this work.

†Corresponding authors.

[3]https://marble-bm.shef.ac.uk

representation learning [20, 27, 26, 2, 15, 39, 48] and music pre-trained models [35, 28, 56, 50, 25, 10, 23, 43, 53, 30]. The existing benchmarks, GLUE [47], SuperGLUE [46], and ERASER [9] in NLP, along with VTAB [55] and VISSL [12] in CV, all play an active role in promoting the development of SSL-related research topics in the corresponding domains. However, there are only scattered and fragmented evaluations of the existing music models rather than comprehensive benchmarks, making it difficult to objectively compare and draw insights across techniques.

In the current context, the SSL music systems are evaluated with downstream task datasets, including genre classification [20, 28, 56, 50, 25, 7, 23, 53, 30], emotion classification [28, 25, 7, 23, 30], instrument classification [28, 50, 39, 30], music tagging [56, 25, 7, 43, 35, 23, 53, 30], key detection [25, 7, 23, 30], music detection [39] and cover song detection [53]. Existing works usually conduct evaluations with different experimental setups, and few of them explore sequential tasks such as beat tracking and source separation. Although in similar domains, SUPERB [52] and HEAR [44] are proposed to facilitate unified analysis of the learned representations of speech and sound events, the distribution of musical audio is significantly different. Thus, there is an urgent need to construct comparable, extensive, and easy-to-use benchmarks to enhance the development of music SSL.

In this paper, we propose a Music Audio Representation Benchmark for universaL Evaluation (MARBLE) to address this problem. MARBLE aims to examine the full spectrum of model capabilities, and thus proposes a taxonomy adapted from Dai et al. [8] to categorise MIR tasks, including acoustic, performance, score, and high-level description. The four-level hierarchy aligned to musician consensus serves as a guideline to further organise the datasets and helps to identify a diversified set of downstream tasks. We select popular tasks in the (now defunct) Music Information Retrieval Evaluation eXchange (MIREX) Challenge[4], and use the corresponding public datasets with limited annotations. As demonstrated in Tab. 1, the current version of MARBLE contains 14 downstream tasks, spread over 10 task categories on 8 publicly or commercially available datasets. Except for the common classification tasks, we also integrate the missing piece of the puzzle – sequence labelling tasks that require frame-wise prediction, including source separation and beat tracking. The datasets used in MARBLE are ensured easy-to-access: all datasets are available for download directly from the official repository or an external website for downloading a specific version.

In addition, we design a unified protocol and build tool-kits to evaluate the generalisation ability of the models. In MARBLE protocol, the models are regarded as backbones to provide universal representations for all tasks, and task-specific prediction heads are concatenated to further trained under *unconstrained*, *semi-constrained*, and *constrained* settings, which is defined by whether the training hyperparameters are restricted and whether the backbone model is frozen (cf. § 3.2). The evaluation suite provides codes for dataset preprocessing and examples of evaluating existing popular SSL models in the benchmark. We select 7 representative music SSL models as our baselines (cf. § 3.1) and release the evaluation results at our publicly available leaderboards[5] as a reference.

Our key contributions are listed as follows: (1) providing a diversified music understanding benchmark with well-defined taxonomy of the MIR tasks; (2) incorporating and organising a wide range of datasets to facilitate comprehensive music model evaluation; (3) designing a unified assessment protocol and building corresponding evaluation suites for processing, training, and benchmarking.

2 Benchmark Tasks

As demonstrated in Tab. 1, we collect datasets in MARBLE to provide the community with a standard, general-purpose, easy-to-use benchmark for various tasks covering all aspects of music. Generally, music processing involves discriminative and generative tasks. The discriminative tasks either classify or regress musical recordings as a whole or use a seq2seq model to make frame-by-frame decisions on entire sequences. The generative tasks include audio synthesis and music composition. For the initial release of MARBLE, we focus on discriminative tasks, and generative tasks are currently outside our scope. The task collection is guided by the principles of (1) receiving a high level of interest in the MIR community, (2) having publicly available datasets allowing everyone to participate, and (3) limited labelled data to effectively measure the universality of the model. Four aspects of music are studied through 14 proposed tasks: **High-level description tasks** including key detection, music tagging,

---

[4]https://www.music-ir.org/mirex/wiki/MIREX_HOME

[5]Considering potential legal constraints, MARBLE allows to submit results on the tasks partially (*e.g.*, tasks on commercially available datasets) for the future participants.

classification gender, and emotion recognition; **Score-level tasks** including estimating the pitch of a musical note and tracking beats; **Performance-level tasks** including detecting musical ornaments or techniques; and **Acoustic-level tasks** including singer identification, instrument classification, and source separation that focus more on raw audio information.

Table 1: The Dataset, Commercial License, and Prediction Head of Each Task Used for the MARBLE Benchmark. SDR refers Source-to-distortion Ratio.

Taxonomy	Task Type	Task & Annotation	Prediction Type	Evaluation Metrics	Commercially Available
High-level Description	Key Detection	Giantsteps key [19]	Multi-class	Weight Score [36]	Yes
	Music Tagging	MagnaTagATune [22]	Multi-label	ROC-AUC & PR-AUC/AP	-
		MTG Top50 [6]	Multi-label	ROC-AUC & PR-AUC/AP	-
	Genre Classification	GTZAN [45]	Multi-class	Accuracy	-
		MTG Genre [6]	Multi-label	ROC-AUC & PR-AUC/AP	-
	Emotion Detection	Emomusic [42]	Regression	$R2^{Valence}$ & $R2^{Arousal}$	-
		MTG MoodTheme [6]	Multi-label	ROC-AUC & PR-AUC/AP	-
Score-level	Pitch Classification	Nsynth [11]	Multi-class	Accuracy	Yes
	Beat Tracking	GTZAN Rhythm [45]	Seq2Seq, Binary-class	F-measure (Threshold 20ms)	-
Performance-level	Vocal Technique Detection	VocalSet [49]	Multi-class	Accuracy	Yes
Acoustic-level	Singer Identification	VocalSet [49]	Multi-class	Accuracy	Yes
	Instrument Classification	Nsynth [11]	Multi-class	Accuracy	Yes
		MTG Instrument [6]	Multi-label	ROC-AUC & PR-AUC/AP	-
	Source Separation	MUSDB18 [37]	Seq2Seq, Regression	SDR	-

2.1 High-level Description Tasks

Key detection involves predicting the scale and key pitch levels of a song. MARBLE solves this task using the Giantsteps [19] and a subset of the Giantsteps-MTG-keys dataset [21]. Giantsteps dataset contains 604 songs and is taken as our dedicated test set. Additionally, we leverage a subset of the Giantsteps-MTG-keys dataset, which contains 1077 music pieces with single-key annotations, for training and validation. Since no standardised split is available for Giantsteps-MTG, we adopt the dataset split strategy employed in [7]. Both datasets contain 2 minutes of electronic dance music covering all 12 pitch classes in major and minor, resulting in a 24-class classification task. For performance evaluation, we employ accuracy with an error tolerance metric, a weighted score metric. This metric grants partial credit for reasonable errors, such as predicting relative secondary keys when the primary key is the ground truth [36].

Music Tagging refers to assigning a predefined set of tags to a given song. These tags encompass various aspects such as genre, instrumentation, mood, and tempo (*e.g.*, fast), making music tagging somewhat overlap with genre classification, emotion recognition, and instrument classification. To conduct our study, we utilise two extensive datasets: MagnaTagATune (MTT) [22] and MTG-Jamendo (MTG) [6]. The MTT dataset comprises 30-second audio clips with manual annotations for tags. It consists of 25.9k clips, amounting to a total duration of 170 hours. For MARBLE, we use the Top50 tags, and adopt a conventional (12:1:3) training, validation, and test split, aligning with all baseline approaches' practices. Besides, the MTG dataset contains 55k clips, corresponding to nearly 2k hours of music. As the audio clips in this dataset may exceed 30 seconds in length, we compute multiple embeddings using a sliding window of 30 seconds and then average them to obtain an overall embedding representation. While both datasets encompass a large number of tags, we follow the customary to limit the vocabulary to the 50 most common tags in each dataset. The evaluation metrics employed for this task are the macro-average of all tag ROC-AUCs (receiver operating characteristic - area under the curve) and the average precision (AP) / PR-AUC (precision-recall - area under the curve). These metrics provide comprehensive insights into the model's performance across all tags.

Genre classification aims to assign each song the most suitable genre label. This study uses two distinct datasets: GTZAN [45] and MTG-Genre. GTZAN consists of 30-second audio clips from 10 genres, making it suitable for a multi-class classification task. To assess the performance of this dataset, we report the accuracy metric. To ensure consistent evaluation, we utilise the "fail-filtered" split as described in [18] for GTZAN. The filtered dataset comprises 930 audio tracks corresponding to approximately 8 hours of music. Besides, MTG-Genre, derived from MTG-Jamendo, contains

55k tracks but focuses solely on 95 genre tags, resulting in a multi-label classification problem. We employ the ROC and AP metrics to evaluate the performance of MTG-Genre.

Emotion Recognition in music aims to determine the emotional content of music pieces. In our study, we utilise two distinct datasets to evaluate the performance of emotion recognition: Emomusic [42] and MTG-MoodTheme [6]. Emomusic contains 744 pieces of 45-second music clips and is annotated with valence and arousal scores. The valence represents the positivity of emotional responses, while arousal indicates emotional intensity. The official evaluation metrics for this dataset is the determination coefficient (r^2) between the model's regression results and human annotations of arousal and valence [42]. During inference, we split the 45-second clips into 5-second sliding windows and computed the average prediction probability as the final prediction. Since no standard dataset split is available for Emomusic, we adopt the same partitioning as [7]. It is important to note that direct comparison of the SoTA model's results with the benchmark may be challenging due to the different dataset splits. Additionally, we utilise MTG-MoodTheme, a subset of MTG-Jamendo consisting of 18.5k audio tracks annotated with 59 human emotion labels. This is a multi-label task with ROC and AP as evaluation metrics.

2.2 Score-level Tasks

Pitch Classification in Music (Monophonic) involves determining the appropriate pitch category for a given audio sample, ranging from MIDI note numbers 0 to 127 on a semitone scale. In this study, we perform pitch classification using the Nsynth dataset [11] within the music information retrieval benchmark. It comprises 340 hours of music, with each excerpt lasting 4 seconds. Since the audio recordings in this dataset are monophonic, the pitch classification task is formulated as a 128-class classification problem, covering all possible MIDI pitch categories (fundamental frequencies from 8Hz to 12.5kHz). The evaluation metric used for this task is the accuracy achieved across all audio clips.

Beat Tracking determines the presence of a beat and a downbeat in each frame of a given music piece. In this benchmark, we only focus on beat tracking, making it a binary-classification task[6]. An offline approach is employed for beat tracking, allowing the model to utilise frame-level information during inference. The model generates frame-by-frame predictions at a specific frequency, which are then post-processed using a dynamic Bayesian network (DBN) [5] implemented with `madmom` [4] to obtain the final result. The GTZAN Rhythm dataset [29] is used in this study. The dataset provides frame-level annotations for each music clip in GTZAN. To enhance model performance and ensure a fair comparison with the spin model, adjacent frames of each beat label are also labelled as beats using a label smoothing technique commonly employed in beat tracking. The model is evaluated using the `f_measure` metric implemented in `mir_eval` [36]. A prediction is considered correct if the difference between the predicted event and the ground truth does not exceed 20ms. It is important to note that while some models may have been trained on other datasets, the GTZAN-train subset is used as the training set, and GTZAN-test is used as the test set for all MARBLE submissions.

2.3 Performance-level Tasks

Vocal Technique Detection task involves identifying different singing techniques within an audio clip. For this task, the MARBLE benchmark utilises the VocalSet dataset [49], the sole publicly available dataset specifically designed for studying singing techniques. This dataset comprises recordings of 20 professional singers (9 female and 11 male) performing 17 distinct singing techniques in various contexts, amounting to a total duration of 10.1 hours. Given that the audio clips are segmented into 3-second intervals, the task focuses on determining the type of technique (*e.g.*Vibrato, Straight) rather than the precise start and end times. To evaluate the performance of models, we employ Accuracy as the evaluation metric. We use a subset of 10 different singing techniques used in Yamamoto et al. [51], which contains 15 singers in the training and validation set, and 5 for the test set. Since there is no predetermined division between the training and validation sets, we assign 9 singers to the training set and 6 singers to the validation set. It is important to note that all 3-second segments originate from the same audio recording file within the same part of the split, such as being exclusively part of the training set. Detailed data partitioning can be found in our provided code.

---

[6]Due to the limitation of time and the size of the dataset, tracking the time signature (*e.g.*, 4/4 metre) and downbeat is deferred to future versions with other datasets.

4

2.4 Acoustic-level Tasks

Instrument Classification refers to the multi-label or multi-class identification of instruments present in a given audio recording. In the MARBLE benchmark, we utilise two datasets: Nsynth and MTG-instrument. The Nsynth dataset comprises 306,000 audio tracks, each corresponding to one of 11 different instruments. The evaluation metric for this dataset is accuracy. On the other hand, MTG-instrument is a subset of MTG-Jamendo, containing 25,000 audio tracks and 41 instrument tags. Each track can have multiple instrument tags and is evaluated based on ROC and AP.

Singer Identification involves recognizing the singer or vocal performer from an audio recording. In previous work on Singer Identification using the VocalSet dataset [49], different splits are employed. For the MARBLE benchmark, we randomly split the dataset into training, validation, and test sets, maintaining a ratio of 12:8:5. All sets contain the same 20 singers. The specific data divisions can be found in the provided code.

Source Separation aims to separate different components of a music recording, such as vocals, drums, bass, and others. In MARBLE, we adopt the widely-used MUSDB18 dataset [37] for this task. MUSDB18 consists of 150 full-length music tracks, totaling approximately 10 hours of audio and multiple isolated stems. Our training set consists of 86 tracks, the validation set contains 14 tracks, and the evaluation set comprises 50 tracks, following the official MUSDB18 setting. During training, we randomly sample 6-second segments and apply random track mixing for data augmentation. Due to the complexity of this task, we utilise the baseline architecture from the Music Demixing Challenge (MDX) 2021 [31]. This architecture consists of three linear layers and three bi-directional LSTM layers. The optimization is performed by directly computing the l2-loss between the predicted and ground-truth spectrograms. The evaluation metric for this task is the Source-to-Distortion Ratio (SDR) as defined in [31], which is calculated as the mean across the SDR scores of all songs.

3 Evaluation Framework

We aim to explore the generality and standardisation of the framework. Therefore, we freeze the parameters of the pre-trained model to extract pre-trained features as fixed depth embeddings fed to each downstream task-specific prediction head. This allows for as lightweight a solution as possible for all tasks, thus testing whether the representations are easily reusable across different downstream tasks. We describe pre-trained baseline models, downstream models, and protocols in the following sections.

3.1 Pre-trained baseline systems

The audio pre-training models explored in this paper are summarised in Table. 2. Note that we do not cover models designed entirely for speech or not open source models. We also examine all the open-source SSL systems specifically designed from music audio, in total 9 different versions of 7 pre-trained features; see Table. 2 for information on pre-trained models.

MusiCNN [35] is a convolutional model pre-trained on the music audio tagging task using the MSD dataset [3]. We use the default configuration of the method, which is to concatenate the mean pooling of the CNN features for a 3-second input with the output of the maximum pool.

Contrastive learning of musical representations (CLMR) [43] leverages a 9-layer 1-D convolutional kernel as the feature extractor, employing a number of data augmentation, and is trained on both MSD and MTT. Both are trained with a contrastive learning approach. The model extracts an embedding every 2.69 seconds.

Jukebox [10] is a music generation model trained using codified audio language modelling (CALM). It is trained on 1.2 million private songs, and the size of the training set is difficult to estimate the exact number of hours. However, assuming an average song length of 3-6 minutes, the total length could be 60k-120k hours, which is large and diverse to allow Jukebox to learn patterns and structures of different musical genres and styles. We use the same mid-layer representation as [7] to improve computational efficiency. Unlike other representations that run on short context windows, JUKEBOX is trained on a long window of 8192 sample points (23.78 seconds) of audio. We use the same strategy as [7] to extract the audio features on the downstream dataset.

Table 2: Information of Baseline Systems.

| Method | MusiCNN MSD-big | CLMR | Jukebox | MULE | MAP-Music2Vec | MAP-MERT-v0 | | MAP-MERT-v1 | |
						base	base-public	base	large
Network	CNN	9-Conv	3-Conv, 36-Trans	22-Conv, 2-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans
#Params	8M	2.5M	5B	62.4M	95M	95M	95M	95M	330M
Input	log-mel	waveform	waveform	log-mel	waveform	waveform	waveform	waveform	waveform
Stride	3s	2.69s	23.78s	2s	20ms	20ms	20ms	13.3ms	13.3ms
Context Length	3s	2.69s	23.78s	3s	30s	5s	5s	5s	5s
Data (hour)	10~20k	1.7k	60~120k	117.5k	1k	1k	0.9k	17k	160k
Pre-training Task	Music Tagging	Contrastive Learning	CALM	Contrastive Learning	MLM Boostrapping	MLM Clustering	MLM Clustering	MLM Clustering	MLM Clustering

MULE (Musicnet-ULarge) [30] is a SSL system based on **SF NFNet-F0** [48], SlowFast Normalizer-Free ResNet. It combines a SlowFast (SF) part (including a slower pathway that captures spatial information and a faster pathway that captures temporal information) with a more efficient and scalable variant of the Normalizer-Free ResNet (NFNet). MULE is contrastively pre-trained on the whole MusicSet dataset [30] and provides promising results on classification tasks. The model extracts an embedding with a 3-second window length and a 2-second hop length.

MAP-Music2Vec [25] is a self-supervised learning (SSL) model specifically based on a bootstrapping mask prediction pre-training strategy. It consists of two main components: the student and teacher models. Both share the same architecture with 12 transformer layers, with the teacher model's parameters being exponential moving averages of the student model's parameters. The student model takes in masked input, and during training, it aims to learn deep features from the teacher model based on the output of the unmasked input. Specifically, it computes the average of the top 8 layers of the Transformer's output in the teacher model. To train the MAP-Music2Vec model, a private dataset comprising approximately 1,000 hours of music data was used. The input length of the MAP-Music2Vec model is set to 30 seconds, producing 50 embeddings per second. These embeddings capture essential features of the music data and can be utilised for various downstream tasks, including sequential tasks such as source separation and beat tracking.

MAP-MERT-v0, also referred to as MERT-95M$^{\text{K-means}}$ in the work by Li et al. [23], is a pre-trained model built upon the speech self-supervised learning (SSL) system HUBERT [15]. It undergoes pre-training for masked prediction, with discrete pseudo-labels obtained from K-Means clustering on music features. The pre-training task of MAP-MERT-v0 involves two pseudo-labels based on logmel and Chroma, along with a CQT reconstruction task that emphasises pitch information. Two versions of the MAP-MERT-v0 model are included: MAP-MERT-v0[7], trained on a private dataset of 1,000 hours, and MAP-MERT-v0-public[8], trained on Music4ALL [40]. The input length of the MAP-MERT-v0 model is set to 5 seconds, generating 50 embeddings per second. This design facilitates fine-tuning for sequential tasks, enabling efficient and effective processing of music data.

MAP-MERT-v1 encompasses two variants: (MAP-)MERT-v1-base[9] and (MAP-)MERT-v1-large[10]. These models, also known as MERT-95M$^{\text{RVQ-VAE}}$ and MERT-330M$^{\text{RVQ-VAE}}$ in the work by Li *et al*. [23], employ EnCodec, a pre-trained discrete deep feature, as a replacement for the K-means feature. This modification facilitates the scaling up of the model. Similar to MAP-MERT-v0, the input length of the MAP-MERT-v1 models is 5 seconds, but they produce 75 embeddings per second. This configuration enables effective fine-tuning for sequential tasks, making the models suitable for processing music data in a variety of applications.

---

[7]https://huggingface.co/m-a-p/MERT-v0
[8]https://huggingface.co/m-a-p/MERT-v0-public
[9]https://huggingface.co/m-a-p/MERT-v1-95M
[10]https://huggingface.co/m-a-p/MERT-v1-330M

3.2 Downstreams and Training Strategies

To evaluate the relevance of representations for downstream MIR tasks, we design evaluation frameworks: the *unconstrained* track, *semi-constrained* track and the *constrained* track. In the unconstrained track, researchers are invited to submit their systems with any hyperparameter and structure configuration, including the option to fine-tune pre-trained models. This track encourages flexibility and exploration, enabling researchers to investigate a wide range of approaches. On the other hand, the semi-constrained track requires the submissions to use frozen pre-trained backbones. Finally, the constrained track employs a standardised setting with limited hyper-parameter search space (cf. Appendix A), where frozen models are used as feature extractors for training a one-layer 512-unit MLP (or 3-layer 512-unit LSTM for source separation) on each task. In addition, we set a computational wall for MARBLE. The systems need to finish each task within a week on our machine equipped with a single consumer GPU (RTX3090). By offering these three evaluation tracks, we aim to provide researchers with a comprehensive platform to assess the performance and relevance of representations in MIR tasks, encouraging innovative approaches and fostering advancements in the field. For the same task with a uniform dataset, if there are different evaluation metrics (*e.g.*, emotion regression, source separation, and tagging), we will average the two evaluation metrics. We select the checkpoints regarding to the best validation results for final testing and submission.

Table 3: Performances of Baselines Evaluated on MARBLE with constrained settings (1/2). We include previous SOTAs for reference. Note that MARBLE imposes strict constraints on downstream structures and hyper-parameter search spaces, while previous SOTAs are not subject to such limitations. Best scores on MARBLE are **bold**, and best scores among all systems are underlined.

Dataset Task	MTT Tagging		GS Key	GTZAN Genre	GTZAN Rhythm	EMO Emotion		Nsynth Instrument	Nsynth Pitch	VocalSet Tech	VocalSet Singer
Metrics	ROC	AP	Acc^{Refined}	Acc	F1^{beat}	R2^V	R2^A	Acc	Acc	Acc	Acc
MusiCNN [35]	90.3	37.8	14.4	73.5	-	44.4	68.8	72.6	64.1	70.3	57.0
CLMR [43]	89.5	36.0	14.8	65.2	-	44.4	70.3	67.9	47.0	58.1	49.9
Jukebox-5B [7, 54]	**91.4**	**40.6**	63.8	**77.9**	-	57.0	73.0	70.4	91.6	76.7	82.6
MULE [30]	91.2	40.1	64.9	75.5	-	**60.7**	73.1	**74.6**	88.5	75.5	__87.5__
MAP-Music2Vec [25]	90.0	36.2	50.6	74.1	68.2	52.1	71.0	69.3	93.1	71.1	81.4
MAP-MERT-v0-95M [24]	90.7	38.2	64.1	74.8	__88.3__	52.9	69.9	70.4	92.3	73.6	77.0
MAP-MERT-v0-95M-public [24]	90.7	38.4	**67.3**	72.8	88.1	59.1	72.8	70.4	92.3	75.6	78.0
MAP-MERT-v1-95M [23]	91.0	39.3	63.5	74.8	__88.3__	55.5	**76.3**	70.7	92.6	74.2	83.7
MAP-MERT-v1-330M [23]	91.1	39.5	61.7	77.6	87.9	59.0	75.8	72.6	__94.4__	**76.9**	87.1
Previous SOTA	92.0 [16]	41.4 [7]	74.3 [21]	83.5 [30]	80.6 [14]	61.7	72.1 [7]	78.2 [48]	89.2 [30]	65.6 [51]	80.3 [32]

Table 4: Performances of Baselines Evaluated on MARBLE with constrained settings (2/2). The overall average scores are calculated on the systems applicable to all tasks. Note that we denote the scores of *Jukebox-5B* on *MTG* tasks with asterisks(*), because it hit the computational wall of MARBLE, meaning that the system was unable to complete the corresponding task within a week on our machine equipped with a single consumer GPU (RTX3090).

Dataset Task	MTG Instrument		MTG MoodTheme		MTG Genre		MTG Top50		MUSDB Source Separation				Avg.
Metrics	ROC	AP	ROC	AP	ROC	AP	ROC	AP	SDR^{vocals}	SDR^{drums}	SDR^{bass}	SDR^{other}	
MusiCNN [35]	74.0	17.2	74.0	12.6	86.0	17.5	82.0	27.5	-	-	-	-	-
CLMR [43]	73.5	17.0	73.5	12.6	84.6	16.2	81.3	26.4	-	-	-	-	-
Jukebox-5B [7, 54]	**78.5***	**22.0***	77.6*	15.3*	**88.0***	**20.5***	83.7*	30.6*	-	-	-	-	-
MULE [30]	76.6	19.2	**78.0**	15.4	**88.0**	20.4	**83.7**	**30.6**	-	-	-	-	-
MAP-Music2Vec [25]	76.1	19.2	76.7	14.3	87.1	18.8	83.0	29.2	5.5	5.5	**4.1**	3.0	59.9
MAP-MERT-v0-95M [24]	76.6	18.7	75.9	13.7	86.9	18.5	82.8	28.8	**5.6**	**5.6**	4.0	3.0	62.3
MAP-MERT-v0-95M-public [24]	77.5	19.6	76.2	13.3	87.2	18.8	83.0	28.9	5.5	5.5	3.7	3.0	63.0
MAP-MERT-v1-95M [23]	77.5	19.4	76.4	13.4	87.1	18.8	83.0	29.0	5.5	5.5	3.8	**3.1**	63.3
MAP-MERT-v1-330M [23]	78.1	19.8	76.5	14.0	86.7	18.6	83.4	29.9	5.3	**5.6**	3.6	3.0	**64.2**
Previous SOTA	78.8	20.2 [1]	78.6	16.1 [30]	87.7	20.3 [1]	84.3	32.1 [30]	9.3	10.8	10.4	6.4 [38]	64.5

4 Results and Discussion

According to Table 3, 4 and Fig 1, all pre-trained baseline representations on MARBLE have achieved decent results. Despite strict constraints on downstream structures and hyper-parameter search spaces, they are able to approach, if not surpass, the previous state-of-the-art (SOTA) in many tasks. For instance, the best performance on NSynth Pitch classification have achieved up to 94.4% accuracy. Nonetheless, the majority of tasks are still far from being solved, including music tagging and source separation tasks. Notably, the performance on MUSDB18 is merely half of the previous SOTAs.
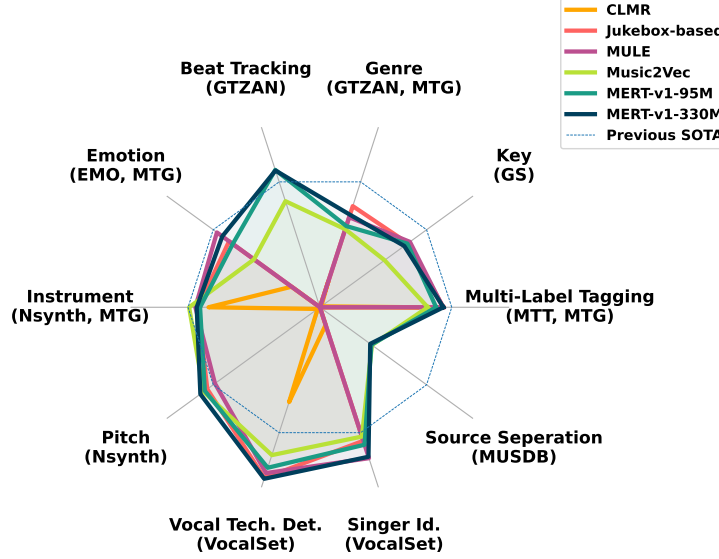
Figure 1: SSL Baselines Compared to previous SOTA. The performances of the tasks are merged according to the task types demonstrated in Tab. 1. Results not applicable are set to 0.

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

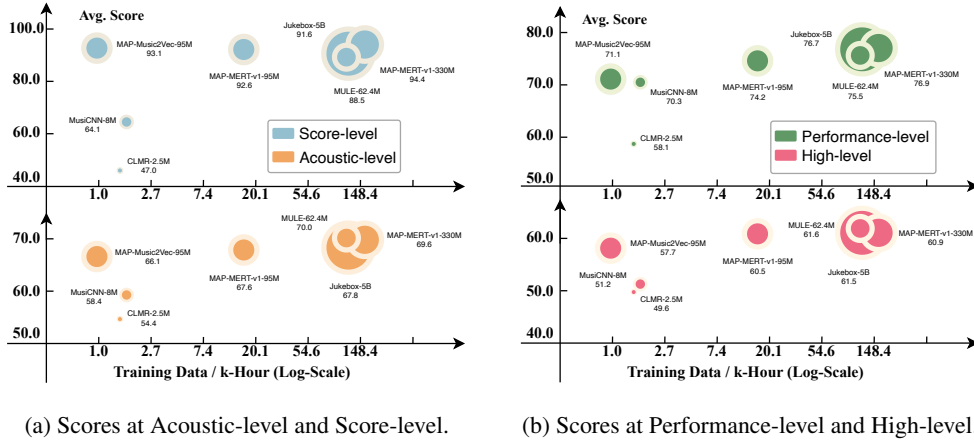(b) Scores at Performance-level and High-level.

Figure 2: Results Analysis Regarding to Training Data Size. Since some models are not applicable to the sequence labelling tasks, the performances of *source separation* and *beat tracking* tasks are excluded on acoustic-level and score-level average score calculation correspondingly. The radii of the scatter points are isometrically log scaling with the parameter sizes.

The MAP family achieves balanced results, successfully performing tasks including sequence labelling, which other models fail to accomplish (as they do not provide frame-level representations or are too cumbersome to train). This series of models excel at multiple taxonomy levels. On certain tasks, MAP-MERTs achieve results close to or surpass the previous state-of-the-art. However, music tagging tasks are dominated by Jukebox-5B and MULE. Jukebox may benefit from its massive parameter size and generative modelling of detailed information, as well as the introduction of metadata during the pre-training period. Conversely, MULE benefits from its proprietary large-scale, high-quality dataset, MusicSet, and the highly discriminative representations learned by contrastive pre-training.

Based on Fig. 2a and 2b, excluding sequence labelling tasks (as some baselines do not support them), we observe a general trend: as the volume of data and the size of model parameters increase, the performance of tasks across four levels correspondingly improves. The choice of pre-training method and model size significantly influences the performance. For instance, MAP-Music2Vec-95M,

utilizing only 1k hours of data for self-supervised learning, outperforms both supervised pre-trained MusiCNN-8M and contrastive pre-trained CLMR-2.5M on the same scale of data. More analysis could be referred to Appendix B.

5 Conclusion

In this work, we introduce the Music Audio Representation Benchmark for universaL Evaluation (MARBLE) as a comprehensive benchmark for evaluating pre-trained music features. It encompasses a hierarchy taxonomy that covers acoustic, performance, score, and high-level description levels, and utilises publicly available datasets for 14 MIR tasks. We establish a standardised preprocessing and data splitting protocol, along with a unified evaluation framework, to ensure fair and reproducible assessment. We report the results of all 9 open-sourced pre-trained models developed on music recordings, showcasing their performance across multiple tasks. The results demonstrate that several pre-trained models achieve comparable or even superior performance to the state-of-the-art models on various tasks within MARBLE. However, there is still ample room for improvement, particularly in music tagging and source separation. With the release of the toolkit, we hope to facilitate future research by providing easy access, reproducibility, and fair comparison of SSL pre-trained models for music understanding. We encourage engagement from researchers in the audio and AI communities to contribute to the advancement of representation learning for music information retrieval.

Limitation and Future Work

Our benchmark has some shortcomings that can be further improved. To begin with, some of the tasks, such as beat tracking and piano transcription, typically use multiple evaluation metrics, but we only include one or two for each of the tasks due to the copyright issues preventing many datasets from being publicly available, lack of standard pre-processing or maintenance, and the limitation of time. Although the selected metric is fundamental and a good indicator, an average of all the metrics might be a better choice. Besides, some of the datasets are not sufficient for a single task. For example, the GTZAN dataset does not have a commercially-available license, and it only includes less than 10 hours of music recordings, making the evaluation more subject to bias. We will include more commercially-available larger datasets on the same tasks. Moreover, we do not include some MIR tasks that lack a common dataset currently, such as cover song detection and query-by-humming. In the future version, we will include more datasets and tasks. Last but not least, MIR on symbolic music is not included in the first version of our benchmark as well.

Apart from the traditional MIR tasks, some interesting tasks deserve more attention for benchmark development in the computer music and AIGC communities. With the benchmark and pre-trained models in MIR, developing an evaluation score on music generation and synthesis might be possible. There may not exist a perfect solution on the subject metrics for music generation to build a benchmark; otherwise, composing musical art will simply search for the waveform with the highest scores. But one can expect such a benchmark can be helpful for the music industry or music education to preclude some bad music generation. Besides, multi-modal approaches that combine music audio with symbolic music and language (*e.g.*, lyrics and music description) also deserve a benchmark.

References

[1] Alonso-Jiménez, P., Serra, X., and Bogdanov, D. (2022). Music representation learning based on editorial metadata from discogs.

[2] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

[3] Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset.

[4] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., and Widmer, G. (2016a). madmom: a new Python Audio and Music Signal Processing Library. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 1174–1178, Amsterdam, The Netherlands.

[5] Böck, S., Krebs, F., and Widmer, G. (2016b). Joint beat and downbeat tracking with recurrent neural networks. In *ISMIR*, pages 255–261. New York City.

[6] Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. (2019). The mtg-jamendo dataset for automatic music tagging. In *International Conference on Machine Learning*. ICML.

[7] Castellon, R., Donahue, C., and Liang, P. (2021). Codified audio language modeling learns useful representations for music information retrieval. *arXiv preprint arXiv:2107.05677*.

[8] Dai, S., Zhang, Z., and Xia, G. G. (2018). Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841*.

[9] DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. (2020). Eraser: A benchmark to evaluate rationalized nlp models. *Transactions of the Association for Computational Linguistics*.

[10] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

[11] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. (2017). Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR.

[12] Goyal, P., Mahajan, D., Gupta, A., and Misra, I. (2019). Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6391–6400.

[13] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

[14] Heydari, M., Cwitkowitz, F., and Duan, Z. (2021). Beatnet: Crnn and particle filtering for online joint beat downbeat and meter tracking. *arXiv preprint arXiv:2108.03576*.

[15] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

[16] Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. (2022). Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*.

[17] Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

[18] Kereliuk, C., Sturm, B. L., and Larsen, J. (2015). Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071.

[19] Knees, P., Faraldo Pérez, Á., Boyer, H., Vogl, R., Böck, S., Hörschläger, F., Le Goff, M., et al. (2015). Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR); 2015 Oct 26-30; Málaga, Spain.[Málaga]: International Society for Music Information Retrieval, 2015. p. 364-70.* International Society for Music Information Retrieval (ISMIR).

[20] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

[21] Korzeniowski, F. and Widmer, G. (2017). End-to-end musical key estimation using a convolutional neural network. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 966–970. IEEE.

[22] Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392. Citeseer.

[23] Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Guo, Y., and Fu, J. (2023). Mert: Acoustic music understanding model with large-scale self-supervised training.

[24] Li, Y., Yuan, R., Zhang, G., Ma, Y., Lin, C., Chen, X., Ragni, A., Yin, H., Hu, Z., He, H., et al. (2022a). Large-scale pretrained model for self-supervised music audio representation learning.

[25] Li, Y., Yuan, R., Zhang, G., MA, Y., Lin, C., Chen, X., Ragni, A., Yin, H., Hu, Z., He, H., et al. (2022b). Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. In *Ismir 2022 Hybrid Conference*.

[26] Ling, S., Liu, Y., Salazar, J., and Kirchhoff, K. (2020). Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6429–6433. IEEE.

[27] Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

[28] Manco, I., Benetos, E., Quinton, E., and Fazekas, G. (2022). Learning music audio representations via weak language supervision. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 456–460. IEEE.

[29] Marchand, U. and Peeters, G. (2015). Swing ratio estimation. In *Digital Audio Effects 2015 (Dafx15)*.

[30] McCallum, M. C., Korzeniowski, F., Oramas, S., Gouyon, F., and Ehmann, A. F. (2022). Supervised and unsupervised learning of audio representations for music understanding. *Ismir 2022 Hybrid Conference*.

[31] Mitsufuji, Y., Fabbro, G., Uhlich, S., Stöter, F.-R., Défossez, A., Kim, M., Choi, W., Yu, C.-Y., and Cheuk, K.-W. (2022). Music demixing challenge 2021. *Frontiers in Signal Processing*, 1:18.

[32] Modrzejewski, M., Szachewicz, P., and Rokita, P. (2023). Transfer learning with deep neural embeddings for music classification tasks. In *Artificial Intelligence and Soft Computing: 21st International Conference, ICAISC 2022, Zakopane, Poland, June 19–23, 2022, Proceedings, Part I*, pages 72–81. Springer.

[33] Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer.

[34] Newell, A. and Deng, J. (2020). How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7354.

[35] Pons, J. and Serra, X. (2019). musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*.

[36] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P., and Raffel, C. C. (2014). Mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, pages 367–372.

[37] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. (2017). The MUSDB18 corpus for music separation.

[38] Rouard, S., Massa, F., and Défossez, A. (2023). Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[39] Saeed, A., Grangier, D., and Zeghidour, N. (2021). Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE.

[40] Santana, I. A. P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R. B., Feltrim, V. D., Domingues, M. A., et al. (2020). Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404. IEEE.

[41] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., and Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

[42] Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., and Yang, Y.-H. (2013). 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 1–6.

[43] Spijkervet, J. and Burgoyne, J. A. (2021). Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*.

[44] Turian, J., Shier, J., Khan, H. R., Raj, B., Schuller, B. W., Steinmetz, C. J., Malloy, C., Tzanetakis, G., Velarde, G., McNally, K., et al. (2022). Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

[45] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.

[46] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

[47] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

[48] Wang, L., Luc, P., Wu, Y., Recasens, A., Smaira, L., Brock, A., Jaegle, A., Alayrac, J.-B., Dieleman, S., Carreira, J., et al. (2022). Towards learning universal audio representations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4593–4597. IEEE.

[49] Wilkins, J., Seetharaman, P., Wahl, A., and Pardo, B. (2018). Vocalset: A singing voice dataset. In *ISMIR*, pages 468–474.

[50] Wu, H.-H., Kao, C.-C., Tang, Q., Sun, M., McFee, B., Bello, J. P., and Wang, C. (2021). Multi-task self-supervised pre-training for music classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 556–560. IEEE.

[51] Yamamoto, Y., Nam, J., and Terasawa, H. (2022). Deformable cnn and imbalance-aware feature learning for singing technique classification. *arXiv preprint arXiv:2206.12230*.

[52] Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., et al. (2021). Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

[53] Yao, D., Zhao, Z., Zhang, S., Zhu, J., Zhu, Y., Zhang, R., and He, X. (2022). Contrastive learning with positive-negative frame mask for music representation. In *Proceedings of the ACM Web Conference 2022*, pages 2906–2915.

[54] Zai El Amri, W., Tautz, O., Ritter, H., and Melnik, A. (2022). Transfer learning with jukebox for music source separation. In *Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part II*, pages 426–433. Springer.

[55] Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. (2019). The visual task adaptation benchmark.

[56] Zhao, Y. and Guo, J. (2021). Musicoder: A universal music-acoustic encoder based on transformer. In *International Conference on Multimedia Modeling*, pages 417–429. Springer.

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, 3-layer 512-unit LSTM (source separation 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.
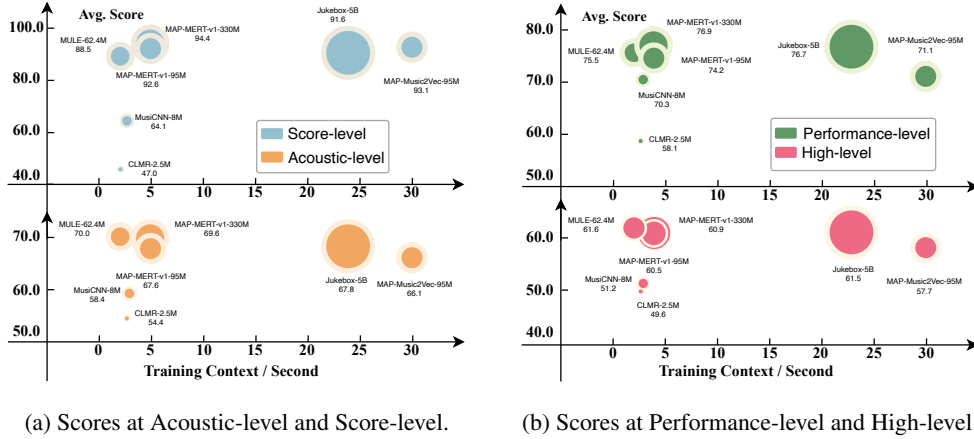
Figure 3: Results Analysis Regarding Training Context Length. The performances of *source separation* and *beat tracking* tasks are ignored similar to Fig. 2.

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 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 guideline. 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.
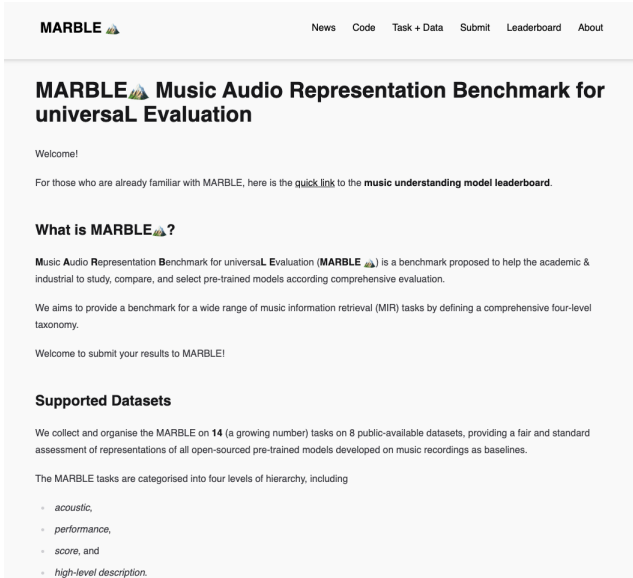
Figure 4: Website for the Proposed MARBLE Benchmark.

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