# MULTI-SOURCE CONTRASTIVE LEARNING FROM MUSICAL AUDIO

**Christos Garoufis**[1,2,3], **Athanasia Zlatintsi**[1,2,3], **Petros Maragos**[2,3]
[1]Institute of Language and Speech Proc., Athena Research Center, Athens, Greece
[2]Institute of Robotics, Athena Research Center, Athens, Greece
[3]School of ECE, National Technical University of Athens, Athens, Greece
cgaroufis@mail.ntua.gr, nancy.zlatintsi@athenarc.gr,maragos@cs.ntua.gr

## ABSTRACT

Contrastive learning constitutes an emerging branch of self-supervised learning that leverages large amounts of unlabeled data, by learning a latent space, where pairs of different views of the same sample are associated. In this paper, we propose musical source association as a pair generation strategy in the context of contrastive music representation learning. To this end, we modify COLA, a widely used contrastive learning audio framework, to learn to associate a song excerpt with a stochastically selected and automatically extracted vocal or instrumental source. We further introduce a novel modification to the contrastive loss to incorporate information about the existence or absence of specific sources. Our experimental evaluation in three different downstream tasks (music auto-tagging, instrument classification and music genre classification) using the publicly available Magna-Tag-A-Tune (MTAT) as a pre-training dataset yields competitive results to existing literature methods, as well as faster network convergence. The results also show that this pre-training method can be steered towards specific features, according to the selected musical source, while also being dependent on the quality of the separated sources.

## 1. INTRODUCTION

Self-supervised learning (SSL) [1, 2] has recently emerged as a learning paradigm that can take advantage of large, unlabelled datasets. An approach that has gained traction in multiple domains involves contrastive learning [3–8], where positive pairs of different *views* of samples, created by suitable augmentation pipelines, are projected in a latent space where projections of each pair are enforced to be as close as possible, while deviating from projections of other pairs. This process can be utilized for pre-training neural networks that can then be fine-tuned to a specific task-at-hand, or by using directly the learned representations as feature vectors for supervised downstream tasks.

In the case of music processing, contrastive SSL has proven competitive to fully supervised alternatives in tasks such as music auto-tagging [9, 10], genre recognition [10] and music recommendation [11] by means of either large music collections [10], or sufficiently sized out-of-domain datasets [7,12]. Moreover, a recent development concerns the transfer of contrastive learning in a multimodal setting, combining information between, i.e., audio and text [13, 14].

An important part of every contrastive learning paradigm is the method used to create the positive pairs. Its importance is profound, since it can lead to large variations in the performance of downstream tasks [4, 9]. In audio representation learning, the most basic methodology for generating positive pairs is cropping different segments from the same audio clip [7], which, in most cases, is also included in more sophisticated augmentation pipelines. In the case of time-frequency representations, inspired by SpecAugment [15], time and frequency masking, pitch shifting, and temporal warping can be applied [10, 11, 16], whereas augmentation pipelines employed for raw audio may involve polarity shift, artificial reverberation, pitch shifting, gain reduction, or additive noise [9].

Musical pieces usually consist of multiple musical sources, highly coordinated in terms of tonality and rhythm. Besides the harmonic properties of the co-playing sources, the existence or absence of a particular source in a piece also carries semantic information; classical music rarely contains vocals, with the exception of chorals, drums and bass are more prominent in rock or electronic songs, etc. Finally, each of the co-playing sources conveys distinct information about the genre and rhythmical, or high-level, properties of the musical piece. Motivated by the above, in this work we propose musical source association (MSA) [1] as a pretext task for audio representation learning, where a neural network learns to associate song excerpts with either a stochastically selected (multi-source) or a particular (source-targeted) vocal or instrumental source. The proposed framework also utilizes semantic information, by taking into account the existence or absence of particular sources in each excerpt. This diverges from a number of research works in the literature that utilize a specific source, such as the singing voice [17–19] or percussion [20, 21]. Application of our proposed methodology on top of COLA [7], a popular self-supervised audio representation learning framework indicates that it can yield competitive results to a number of positive pair creation strategies [9, 10] in three downstream tasks. We further show that models trained to associate song excerpts with a particular source can be steered towards recognizing musi-

*Copyright: © 2023 Christos Garoufis et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

[1] Code/models available: https://github.com/cgaroufis/MSCOL_SMC23

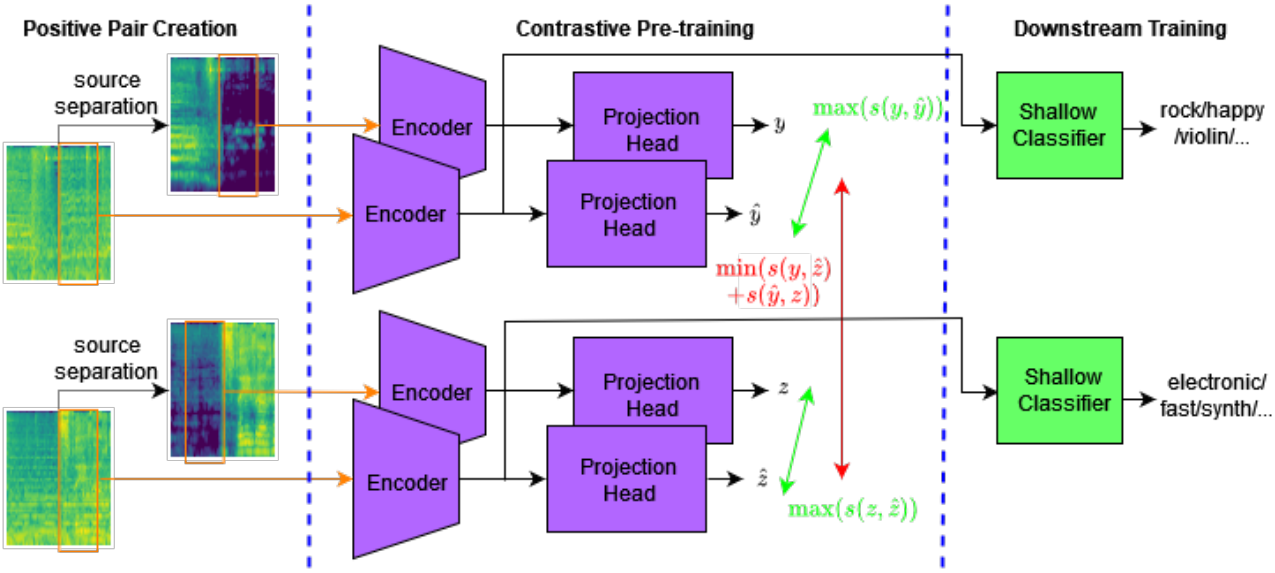


Figure 1. An overview of our proposed framework for learning audio representations, where pairs of song excerpts with a corresponding source are used to pre-train an encoder network through a contrastive loss objective. This enforces the representations of each pair to be close in a latent space, and the representations of different pairs to be further apart.

cal properties tied to this instrumental source. Finally, we repeat the study conducted in [22] concerning the clarity of the separated sources, with the results indicating that in contrast to the case of environmental sounds [22], the quality of the learned representations correlates with the clarity of the separated musical sources.

The rest of the paper is organized as follows: We discuss the related literature in Sec. 2, and present our proposed methodology in Sec. 3. Sec. 4 describes the datasets used in this work, as well as the preprocessing applied upon them, while in Sec. 5 we outline our experimental setup. The results are presented and discussed, for both multi-source and source-targeted models, in Sec. 6, while we draw some final conclusions in Sec. 7.

## 2. RELATED WORK

During recent years, a multitude of frameworks suited for contrastive SSL from musical signals have appeared in the literature. These frameworks can be decomposed into a number of distinct components: i) the *augmentation pipeline* that is used to generate the positive pair of different views, ii) the *backbone encoder* that learns the intermediate representations, iii) the *projection head*, where the resulting representations are projected, and iv) the *contrastive loss function* that is applied upon the projected representations.

COLA [7] constitutes a simple, yet effective framework for contrastive audio representation learning in the time-frequency domain, operating on mel-spectrograms. Anchor-positive pairs are generated by cropping different temporal segments of the same mel-spectrogram, and then processed through an EfficientNet-B0 [23] backbone, to provide an embedding of fixed length. To formulate a learning objective upon these embeddings, the network is tasked to identify the positives that correspond to each anchor within a batch. Embeddings learned from a sufficiently large pre-training dataset [24] showed to have high generalization ability across different speech and music tasks.

Numerous works in music representation learning utilize this core methodology, altering some of its components. For instance, [10] propose utilizing a backbone based on Swin-Transformers [25] in conjunction with a data-driven augmentation pipeline, which involves time cropping, pitch shifting, time warping, and stochastic masking of the spectrograms in both time and frequency axes. Moreover, a number of positive mining strategies and loss function formulations on top of a CNN backbone are compared in [11]. In [9], the authors successfully applied these principles in the waveform domain, by designing a series of augmentations applied in the waveform level, and then a suitable encoder based on SampleCNN [26] for learning the intermediate representations. Finally, utilization of editorial metadata in the form of information about the artist, album, and others, has also been proposed for pair mining in [27], yielding good results in a number of downstream tasks.

Another category of pretext tasks utilized for learning high-level audio representations is based on source separation. In [28], additive mixtures of audio excerpts are used in conjuction with their components as input pairs, which the authors expand on in [22] by using an unsupervised source separation system [29] to extract separated views of the initial audio segment. In music signal processing, similarities between song excerpts and the corresponding vocals have been exploited, using triplet losses [17, 18] or batch-wise contrastive losses [19], usually for the task of artist identification. Furthermore, in [20] the positive pairs are created using the percussive part of a music segment and its non-percussive accompaniment, without temporal cropping – an idea that was expanded in [21]. The features learned this way were shown to outperform previous methods in the task of beat synchrony prediction.

## 3. METHODOLOGY

An overview of our proposed approach is displayed in Fig. 1. Building upon the COLA [7] framework, pairs of song ex-

cerpts and a corresponding source are used to pre-train an encoder via a contrastive loss objective. The encoder is tasked to maximize a similarity measure between learned representations of the song excerpt and its separated source, while minimizing it for song excerpts and separated sources from different input songs. Afterwards, these representations can be used for other downstream tasks.

### 3.1 Positive Pair Creation

We consider the additive mixture model for musical signals, where each signal can be decomposed in $M$ co-playing sources, which can be either active or silent. Thus, we generate positive pairs by matching each song excerpt (further denoted as anchor) with a time-shifted version of an extracted source from the same excerpt (denoted as positive), such as the vocals or one of the instruments. Association of a musical excerpt with an isolated source enables its easier association with sonic or high-level properties tied to a particular source. Furthermore, time-shifting the source excerpt, similar to [7], both allows the network to learn long-term dependencies in the audio and acts as data augmentation, increasing the positive pairs seen during training. This framework can be used to train either models targeted to the association of song excerpts with multiple sources, or a particular, source-targeted one (such as the vocals, the melodic accompaniment, the percussion, etc).

We note that in both cases, each batch contains source excerpts from the same musical source; thus, they have similar timbre, increasing the specificity of the learned representations. Furthermore, in the case of models trained on multiple sources, each batch may either include only non-silent source segments, or source segments sampled from the whole training set, which could contain silence. Sole utilization of the first approach discards a large portion of training data and prevents the model from learning semantic information pertaining to the existence or absence of a source. On the other hand, the second approach on its own yields, due to the appearance of multiple silent segments per batch, a smaller effective batch size, which is detrimental to learning useful representations from SSL [4, 7, 9].

### 3.2 Network Backbone

Similar to [7], the encoder network is based on EfficientNet-B0 [23]. In more detail, it consists of a series of inverted residual depthwise-convolutional blocks with linear bottlenecks [30], grouped in stages, with each stage gradually reducing the resolution of its input. The output tensor of the last stage is flattened via a global max pooling operation, to acquire an one-dimensional feature vector for each input instance, with fixed size equal to 1280. On top of the encoder, we apply a linear projection head to obtain an embedding vector, of dimensionality equal to 512, followed by a Layer Normalization and a $\tanh()$ activation.

### 3.3 Contrastive Loss Objective

Contrastive learning methods apply the loss function upon the projected representations of their inputs within-batch, i.e., they try to identify the instances that correspond to the respective anchors among other instances in the batch. However, in the cases where the selected source spectrogram is silent, the same positive instance will correspond to multiple anchor spectrograms, generating thus false positives. Thus, before applying a similarity function between the embedding pairs, we aggregate all embeddings of silent source instances into a single embedding vector, $\bar{y}$. Comparisons are performed within-batch; for pairs where the positive example corresponds to an active source, the embeddings of the anchor are encouraged to be close to the embeddings of their positive pair, while far from those of the other positive instances in each batch. Similarly, for anchor excerpts with a silent source, the embeddings are attracted to the within-batch centroid of the silent positive embeddings, and repelled from all other instances corresponding to active sources. Following [7], the bilinear similarity is used as the similarity function between two instances $y, \hat{y}$, defined as:

$$s(y, \hat{y}) = y^T W \hat{y}, \tag{1}$$

where $W$ denotes the corresponding bilinear coefficient matrix, implemented as an additional learnable linear layer that is applied upon $\hat{y}$. To formulate a loss objective upon the pair similarities, the softmax function is applied upon the batch similarity matrix, with a temperature parameter $T$, and a binary cross-entropy loss function is estimated from the computed logits:

$$\mathcal{L} = -\sum_{y \in S} \log \frac{\exp(s(y, \hat{y})/T)}{\sum_{z \in S^{*-}} (\exp(s(y, z)/T))}. \tag{2}$$

This loss is similar to the variant of the NT-XEnt loss [31] used in [7], but is modified to accomodate the false positives from including silent source excerpts. In particular, $S$ denotes the set of anchor embeddings in the current batch, but $S^{*-}$ now denotes the modified distractor embeddings, namely, the intact embeddings of the non-silent positives and the centroid of the embeddings of the silent positives.

## 4. DATA AND PREPROCESSING

**Datasets**: As our pre-training dataset, we would ideally want to utilize a large dataset of music tracks, paired with their isolated musical sources. However, no such dataset is publicly avaliable, with the most popular dataset for benchmarking music source separation algorithms [32] having a total duration of around 10 hours. An intriguing alternative stems from the increase in the performance of automatic audio source separation algorithms in recent years [33–36], which allows the creation of large datasets consisting of high-quality automatically separated sources. However, the process of extracting the various sources from large unlabelled music collections is time-consuming. As a result, we utilized the Magna-Tag-A-Tune (MTAT) dataset [37], which includes 25863 song clips, sampled at 16 kHz, with a duration of 30 seconds each, leading to a total duration of 215 hours. This allows for the acquisition of the estimated source excerpts via automatic source separation algorithms in a sensible time period, while being sufficiently large for self-supervised audio pre-training [9, 13].

Each of the song clips in MTAT is associated with a number of tags. MTAT contains a total of 188 unique tags, out of which the top 50 are used as a popular benchmark for music auto-tagging. In this work, we use the filtered version of MTAT presented in [38] which includes only samples associated with any of the 50 most frequent tags.

Concerning the downstream tasks, we re-used MTAT for the task of music auto-tagging. Moreover, to examine the cross-dataset generalization of our approach, we utilized NSynth [39] for the instrument family classification task, and the *small* subset of the FMA dataset [40] for the task of genre classification. NSynth includes approximately 300000 monophonic note instances from 11 instrument families, sampled at 16 kHz, whereas the *small* subset of the FMA dataset contains 8000 song excerpts, each 30 seconds long, sampled at 16 kHz, and equally split into 8 root genres.

**Data Preprocessing**: All audio segments corresponding to MTAT were sliced into 4-sec chunks, and mel-spectrograms were extracted using a window size of 25ms, a hop length of 10ms, and 64 mel bands, following [7]. Since in [7] an input segment length of 1 second is utilized, random segments of 98 frames were cropped, at each training iteration, for both anchor and source samples.

To acquire the audio excerpts corresponding to various musical sources, we used the open-source open-unmix [41] framework for music source separation, which decomposes a musical mixture into the vocal track, the bass track, the drums, and the melodic accompaniment. Open-unmix operates on the time-frequency domain, and specifically in the STFT magnitude, estimating soft masks for each source, which are then applied in the STFT of the mixture to isolate them. The song excerpts were upsampled to 44.1 kHz for compatibility purposes, and the extracted sources were downsampled to 16 kHz before spectrogram computation. As a further pre-processing step, after inspecting the energy of the extracted sources, we replaced all source excerpts with a mean absolute value below 0.01 with silence. We observe that with the exception of the melodic accompaniment ("Other"), the rest of the sources are active for less than half of the audio segments; in particular, 44.67 % of the bass segments, 41.94 % of the drum segments, 87.72 % of the melodic accompaniment segments and 31.50 % of the vocal segments in the training set are active.

## 5. EXPERIMENTAL PROTOCOL

The training procedure was split into two stages: a) self-supervised pre-training of the encoder along with the projection head and b) training of a shallow classifier upon the learned encoded representations (freezing the encoder and dropping the projection head). Models were trained for both source-targeted (single-source) and multi-source cases. In the former case, the positive pairs were created only from one extracted source, while in the latter case, a source was selected at random for each batch. In both cases, the encoder and the projection head were trained via the contrastive loss objective defined in Eq. (2), using $T = 0.2$. The training procedure lasted 10000 steps (approximately 800 epochs for MTAT), with each step consisting of 64 batches; due to GPU limitations, each batch consisted of 128 positive pairs. Adam was used for the self-supervised pre-training with an initial learning rate equal to 0.001, which was halved after 5000 steps. During pre-training, we only use the training subset of MTAT [2], corresponding to unseen testing data. The pretext task performance at the validation set was used for early stopping; we halted pre-training after the running average of the validation loss over 1000 steps did not improve.

Concerning the training of the downstream classifiers, in the case of MTAT, the common 12:1:3 split between training, validation and testing data [9,42] was employed, while for both NSynth and FMA we utilized the default splits between training, validation and testing data: in the case of NSynth, the training, validation and testing sets consist of 289205, 12678, and 4096 audio segments, while for FMA, the data are split in a stratified 8:1:1 ratio into the respective splits. For the cases of MTAT and FMA, we directly used a linear classifier, while for NSynth, we used an intermediate layer with 512 neurons and a ReLU activation function, as in [13, 43]. The classifiers were trained using Adam with a learning rate of 0.0005 for MTAT and FMA and 0.0003 for NSynth, and a batch size of 128, while early stopping was applied with a patience of 5 epochs. Since music auto-tagging constitutes a multi-instance multi-label task, we use the categorical cross-entropy as its loss function, and report on the macro average ROC-AUC and PR-AUC values over all tags. On the other hand, for both instrument family classification and music genre classification, the binary cross-entropy is used as the loss function and the weighted accuracy (WA, %) as the evaluation metric. During evaluation, all models receive 1-sec audio segments, and aggregate the predictions for each instance via averaging the per-segment predictions. We further note that all downstream classification experiments were repeated 5 times, and we report on their average scores. To assess the statistical significance of the results, t-tests were performed between the results of the repeated experiments.

We compared our proposed methodology (MSA, music source association) against the following baselines, which we retrain using the same pre-training dataset:

- The unmodified COLA [7] framework, where the positive pairs were created by choosing different segments of the same audio clip.
- The encoder, projection and contrastive loss used in [7], coupled with the data-driven methodology to create positive pairs described in [10], denoted here as MWS (masking, warping and shifting).
- The proposed methodology, but instead of using a fully trained source extractor for obtaining the positive instances, the various sources are extracted from a randomly initialized, untrained variant of open-unmix [41]. This way, we investigate whether targeted masking of the input helps in learning useful representations as opposed to random soft masking.

Finally, we compare our method to CLMR [9], a waveform-based contrastive learning framework for music, which was also trained on MTAT as a pre-training dataset.

[2] According to the split in https://github.com/jongpillee/music_dataset_split

| SSL Framework | MTAT | | MTAT* | | NSynth | FMA |
|---|---|---|---|---|---|---|
| | ROC-AUC | PR-AUC | ROC-AUC | PR-AUC | WA (%) | WA (%) |
| CLMR [9] | - | - | 0.887 | 0.356 | - | 0.484 |
| COLA [7] | 0.886 | 0.396 | 0.880 | 0.334 | 0.593 | 0.460 |
| COLA [7] + MWS [10] | 0.898 | 0.425 | 0.892 | 0.358 | **0.645** | 0.493 |
| COLA [7] + Random Mask | 0.883 | 0.390 | 0.880 | 0.337 | 0.632 | 0.476 |
| COLA [7] + MSA (ours) | **0.900** | **0.429** | **0.895** | **0.361** | 0.627 | **0.510** |

Table 1. Performance of the pre-trained representations in MTAT in different downstream tasks: music auto-tagging (MTAT and MTAT*), instrument classification (NSynth) and genre classification (FMA), according to the strategy used for generating the positive pairs. The results for CLMR are collected from [9] and [10].

| Configuration | MTAT | | MTAT* | | NSynth | FMA |
|---|---|---|---|---|---|---|
| | ROC-AUC | PR-AUC | ROC-AUC | PR-AUC | WA (%) | WA (%) |
| COLA [7] + MSA | **0.900** | **0.429** | **0.895** | **0.361** | 0.627 | **0.510** |
| NSV1 | 0.896 | 0.418 | 0.891 | 0.351 | 0.648 | 0.501 |
| NSV2 | 0.895 | 0.418 | 0.890 | 0.354 | 0.645 | 0.486 |
| A1 | 0.898 | 0.424 | 0.893 | 0.358 | 0.635 | 0.499 |
| A2 | 0.897 | 0.423 | 0.892 | 0.357 | **0.653** | 0.504 |
| A3 | 0.841 | 0.310 | 0.838 | 0.262 | 0.537 | 0.425 |

Table 2. Ablation study on the effect of the batch composition on the performance in different downstream tasks.

## 6. RESULTS AND DISCUSSION

### 6.1 Multi-Source Models

**Comparison to Baselines**: The performance of our proposed method in all three downstream tasks, in comparison to the baselines mentioned above, is presented in Table 1. For easier comparison with the literature, we report results on both filtered (MTAT) and unfiltered (MTAT*) versions of MTAT. We observe that MSA outperforms both the COLA baseline [7] and application of random soft binary masks (COLA + Random Mask) on both versions of MTAT, at a statistical significance level of $p < 0.01$ for all metrics. In fact, the performance of the random soft masking variant of MSA is similar to COLA ($p > 0.005$ for all metrics after Bonferroni correction), indicating that unstructured masking of the spectrograms on its own cannot provide sufficient information towards learning useful intermediate representations. Furthermore, we note that our method performs comparably ($p > 0.005$ for all metrics post-Bonferroni) to the data-driven MWS methodology for creating the positive pairs, even though we employ a simpler augmentation pipeline – merely matching an audio excerpt to a cropped segment of an extracted source. In the cross-dataset setup, MSA only outperforms COLA ($p < 0.01$) in the task of instrument classification, and both COLA and COLA+Random Mask ($p < 0.01$) in genre recognition, with no statistically significant difference with MWS ($p > 0.05$). In comparison to CLMR [9], we observe that we obtain better results on FMA, while achieving comparable performance on MTAT*. Finally, we note that the performance gap between MTAT and MTAT* is consistent with the literature [27].

**Ablation Study**: In order to investigate the properties of our proposed framework, we perform an extensive ablation study. The effect of incorporating silent source excerpts during training is examined through the following variants:

- Non-Silent Variant #1 (NSV1): All batches contain non-silent positive examples from the same source, and silent source excerpts are discarded.
- Non-Silent Variant #2 (NSV2): Similar to NSV1, but the source segments in each batch may correspond to different sources.

Furthermore, the importance of performing time-shifting to the source excerpts, modifying the loss function to incorporate silent segments, and constructing training batches containing a single source per batch, is examined through the following ablated variants, using the full training dataset:

- Ablation #1 (A1): Batches consist of positive pairs that contain multiple, different sources in each batch.
- Ablation #2 (A2): The loss function of [7] is used without the modification applied to accomodate the false positives in the case of silent segments.
- Ablation #3 (A3): The source excerpts are temporally aligned with the corresponding song excerpt.

The results of this ablation study are displayed in Table 2. We note that the proposed framework significantly outperforms all non-silent variants at the $p < 0.001$ level (after Bonferroni correction) in music auto-tagging, and NSV2 ($p < 0.01$ post-Bonferroni) in genre recognition, indicating that incorporating silent segments assists in learning useful features for these tasks. Compared to A1 and A2, MSA achieves slightly higher (but not significant at the $p < 0.05$ level) performance in these tasks, implying a slight positive effect of the examined parameters. However, this trend is reversed in instrument classification, with all examined variants, with the exception of A3, scoring at least comparably to MSA, and A2 outperforming it at the $p < 0.05$ level, indicating that the proposed loss is not suitable for this task. Finally, A3 performs significantly ($p < 0.001$) worse than MSA in all downstream tasks, showing that time-shifting of the source excerpts plays a critical role in learning informative representations.

**Qualitiative Analysis**: The evolution of the ROC-AUC and PR-AUC scores in the downstream task of music auto-

| Source | MTAT | | MTAT* | | NSynth | FMA |
|---|---|---|---|---|---|---|
| | ROC-AUC | PR-AUC | ROC-AUC | PR-AUC | WA (%) | WA (%) |
| Bass | 0.885 | 0.392 | 0.880 | 0.332 | 0.624 | 0.487 |
| Drums | 0.881 | 0.385 | 0.876 | 0.328 | 0.611 | **0.503** |
| Accomp. | 0.892 | 0.410 | 0.888 | **0.348** | **0.636** | 0.480 |
| Vocals | **0.896** | **0.412** | **0.891** | **0.348** | 0.632 | 0.481 |
| None (COLA) | 0.886 | 0.396 | 0.880 | 0.334 | 0.593 | 0.460 |
| Multi-Source | 0.900 | 0.429 | 0.895 | 0.361 | 0.627 | 0.510 |

Table 3. Performance of the pre-trained representations (from source-targeted models) in MTAT on different downstream tasks, according to the source used for generating the positive pairs.

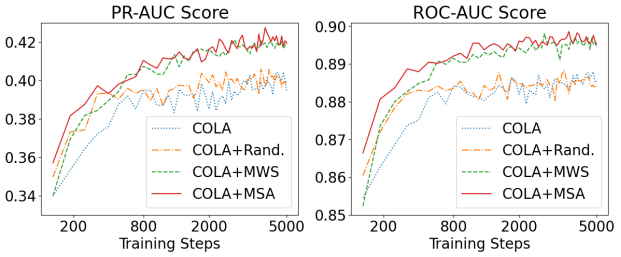


Figure 2. Evolution of the PR-AUC and ROC-AUC scores on MTAT in the course of the self-supervised pre-training.

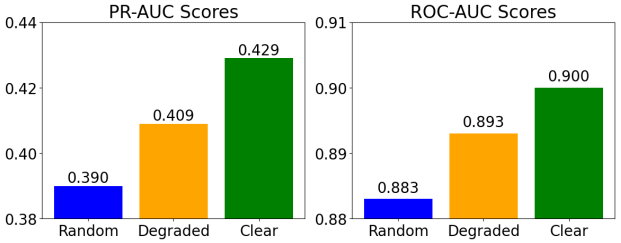


Figure 3. PR-AUC (left) and ROC-AUC (right) scores in MTAT, for positive pairs using random (blue), degraded (orange) and clear (green) source excerpts.

tagging in MTAT, as evaluated using embeddings from various checkpoints of the pre-training process, is shown in Fig. 2. Interestingly, we note that our proposed methodology (red line) increases its performance faster than its data-driven counterpart [10] (green dashed line) at the early stages of pre-training, reaching a PR-AUC score of 0.40 within 300 training steps. We assume that the musically motivated pretext task of music source association allows the network to quickly learn meaningful intermediate representations. However, the multitude of generated pairs of the data-driven MWS strategy, due to the inherent stochasticity of the masking, pitch shifting and time warping operations (whereas only 4 different spectrogram excerpts, one for each source, can be cropped in MSA), allows its performance to catch up as pre-training progresses. This could imply that the two strategies can be combined, by utilizing them either in tandem or in succession.

To further investigate the necessity of utilizing separated sources of high quality, we conducted an experiment similar to [22] to examine the performance of MSA under degrading sources. Thus, for each source, we trained an instance of open-unmix [41] for just 2 epochs, using the musdb18 [32] dataset. The ROC-AUC and PR-AUC metrics in MTAT when using the noisy source excerpts, compared to both the completely random masks (COLA + Random Mask) and the clear sources extracted from the fully trained one, are displayed in Fig. 3. We observe that in contrast to the case of unsupervised source separation of environmental sounds [22], the pretext task of source association in musical signals is aided by high quality source excerpts; even though noisy sources lead to worse results ($p < 0.001$) than the output of the fully-trained open-unmix, they still outperform ($p < 0.01$) random masking.

### 6.2 Source-Targeted Models

The results of the models trained to learn the association between a song excerpt and a particular source are displayed in Table 3. We observe that, on average, the models trained using the vocals and the melodic accompaniment perform the best, with a PR-AUC score of 0.412 and 0.410 on MTAT, respectively, outperforming the COLA baseline at the $p < 0.05$ level. On the other hand, the bass- and drums-based models do not reach this performance, recording scores slightly lower than the COLA baseline in MTT and slightly higher in instrument classification, with the exception of the genre recognition task, where the drums-based model outperforms (albeit not on the $p < 0.05$ statistical significance level) all other source-targeted models. However, with the exception of the instrument classification task, none of these targeted models reach the performance of the multi-source models, indicating that utilization of multiple sources can be useful for the identification of various low or high-level musical traits.

Indeed, all sources contribute to the performance of the multi-source model, since particular spectral, timbral or higher-level characteristics are primarily associated with each source. To visualize this, in Fig. 4, we display, for the 20 most frequently appearing tags in MTAT, the differential in the per-tag PR-AUC scores between COLA [7] and the models trained with each of all 4 extracted sources: bass, drums, the rest of the melodic accompaniment and vocals. We observe that source-targeted models display a specialization, according to the utilized source. For instance, despite their overall low performance, the embeddings trained to associate song excerpts with the corresponding percussion achieve relatively good performance in rhythmic tags, such as *fast* and *drums*, as well as related genres such as *techno*. In the same vein, embeddings trained on the melodic accompaniment display the highest tagging metrics in the recognition of accompanying instruments, such as *guitar*, *piano*, and *violin*, while those trained using the corresponding vocals perform the best in the related tags *vocal*, *male*, *female* and *singing*. Interestingly, none of the source-targeted models outperform the

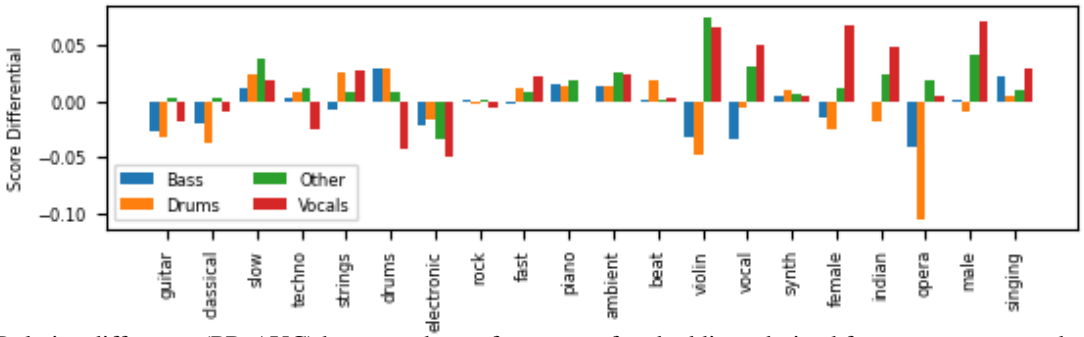


Figure 4. Relative difference (PR-AUC) between the performance of embeddings derived from source-targeted models and COLA-derived embeddings, in the 20 most frequent tags from the MTAT dataset.

COLA baseline in a number of tags, such as *electronic*.

## 7. CONCLUSIONS

In this paper, we investigated the suitability of musical source association as a pretext task in the context of contrastive music representation learning. Results on three different downstream tasks indicate that our framework can yield competitive results compared to recent literature methods for creating positive pairs under the same pre-training dataset. Furthermore, we show that it can be used to steer downstream classifiers toward specific sonic properties, by using a single (vocal or instrumental) source for the association task. In the future, we would like to explore the scalability of our findings in larger music collections, as well as experiment with other musically motivated pretext tasks, such as mining positive pairs from playlists. Another interesting direction concerns examining the efficiency of our approach in producing genre or high-level tags, using single-track audio, i.e. the vocal track.

### Acknowledgments

This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project Number: 7773). We would also like to thank the anonymous reviewers, whose constructive comments helped in improving the quality of this paper.

## 8. REFERENCES

[1] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative Unsupervised Feature Learning with Convolutional Neural Networks," in *Proc. NeurIPS 2014*, Montreal, Canada, 2014.

[2] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond *et al.*, "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning," in *Proc. NeurIPS 2020*, online, 2020.

[3] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," in *Proc. NeurIPS 2018*, Montreal, QU, Canada, 2018.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. ICML 2020*, online, 2020.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proc. CVPR 2020*, online, 2020.

[6] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *Proc. CVPR 2021*, online, 2021.

[7] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive Learning of General-Purpose Audio Representations," in *Proc. ICASSP 2021*, online, 2021.

[8] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proc. EMNLP 2021*, Punta Cana, Dominican Republic, 2021.

[9] J. Spijkervet and J. A. Burgoyne, "Contrastive Learning of Musical Representations," in *Proc. ISMIR 2021*, online, 2021.

[10] H. Zhao, C. Zhang, B. Zhu, Z. Ma, and K. Zhang, "S3T: Self-Supervised Pre-training with Swin Transformer for Music Classification," in *Proc. ICASSP 2022*, Singapore, Singapore, 2022.

[11] J. Choi, S. Jang, H. Cho, and S. Chung, "Towards Proper Contrastive Self-Supervised Learning Strategies for Music Audio Representation," in *Proc. ICME 2022*, Taipei, Taiwan, 2022.

[12] S. Srivastava, Y. Wang, A. Tjandra, A. Kumar, C. Liu *et al.*, "Conformer-Based Self-Supervised Learning for Non-Speech Audio Tasks," in *Proc. ICASSP 2022*, Singapore, Singapore, 2022.

[13] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Learning Music Audio Representations via Weak Language Supervision," in *Proc. ICASSP 2022*, Singapore, Singapore, 2022.

[14] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, "MuLan: A Joint Embedding of Music Audio and Natural Language," in *Proc. ISMIR 2022*, Bengaluru, India, 2022.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, Graz, Austria, 2019.

[16] M. Emami, D. Tran, and K. Koishida, "Augmented Contrastive Self-Supervised Learning for Audio Invariant Representations," *arXiv preprint arXiv:2112.10950*, 2021.

[17] K. Lee and J. Nam, "Learning a Joint Embedding Space of Monophonic and Mixed Music Signals for Singing Voice," in *Proc. ISMIR 2019*, Delft, the Netherlands, 2019.

[18] K. L. Kim, J. Lee, S. Kum, and J. Nam, "Learning a Cross-Domain Embedding Space of Vocal and Mixed Audio with a Structure-Preserving Triplet Loss," in *Proc. ISMIR 2021*, online, 2021.

[19] H. Yakura, K. Watanabe, and M. Goto, "Self-Supervised Contrastive Learning for Singing Voices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1614–1623, 2022.

[20] D. Desblancs, R. Hennequin, and V. Lostanlen, "Zero-Note Samba: Self-Supervised Beat Tracking," *https://hal.science/hal-03669865/document*, 2022.

[21] M. Heydari and Z. Duan, "Singing Beat Tracking with Self-Supervised Front-end and Linear Transformers," in *Proc. ISMIR 2022*, Bengaluru, India, 2022.

[22] E. Fonseca, A. Jansen, D. P. Ellis, S. Wisdom, M. Tagliasacchi *et al.*, "Self-Supervised Learning from Automatically Separated Sound Scenes," in *Proc. WASPAA 2021*, New Waltz, USA, 2021.

[23] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. ICML 2019*, Long Beach, CA, USA, 2019.

[24] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence *et al.*, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in *Proc. ICASSP 2017*, New Orleans, LA, USA, 2017.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei *et al.*, "SWin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. CVPR 2021*, online, 2021.

[26] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-end Deep Convolutional Neural Networks using Very Small Filters for Music Classification," *Applied Sciences*, vol. 8, no. 1, p. 150, 2018.

[27] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music Representation Learning Based on Editorial Metadata from Discogs," in *Proc. ISMIR 2022*, Bengaluru, India, 2022.

[28] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised Contrastive Learning of Sound Event Representations," in *Proc. ICASSP 2021*, online, 2021.

[29] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised Sound Separation using Mixture Invariant Training," in *Proc. NeurIPS 2020*, online, 2020.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetv2: Inverted Residuals and Linear Bottlenecks," in *Proc. CVPR 2018*, Salt Lake City, UT, USA, 2018.

[31] K. Sohn, "Improved Deep Metric Learning with Multi-Class n-Pair Loss Objective," in *Proc. NeurIPS 2016*, Barcelona, Spain, 2016.

[32] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 Corpus for Music Separation," https://doi.org/10.5281/zenodo.1117372, 2017.

[33] Y. Luo and J. Yu, "Music Source Separation with Band-Split RNN," *arXiv preprint arXiv:2209.15174*, 2022.

[34] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling Magnitude and Phase Estimation with Deep Res-U-Net for Music Source Separation," in *Proc. ISMIR 2021*, online, 2021.

[35] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez *et al.*, "Music Demixing Challenge 2021," *Frontiers in Signal Processing*, vol. 1, p. 18, 2022.

[36] E. Manilow, C. Hawthorne, C.-Z. A. Huang, B. Pardo, and J. Engel, "Improving Source Separation by Explicitly Modeling Dependencies Between Sources," in *Proc. ICASSP 2022*, Singapore,Singapore, 2022.

[37] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of Algorithms using Games: The Case of Music Tagging," in *Proc. ISMIR 2009*, Kobe, Japan, 2009.

[38] M. Won, S. Chun, O. Nieto, and X. Serrc, "Data-Driven Harmonic Filters for Audio Representation Learning," in *Proc. ICASSP 2020*, online, 2020.

[39] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi *et al.*, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," in *Proc. ICML 2017*, Sydney, Australia, 2017.

[40] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A Dataset for Music Analysis," in *Proc. ISMIR 2017*, Suzhou, China, 2017.

[41] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A Reference Implementation for Music Source Separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[42] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-Based Automatic Music Tagging Models," in *Proc. SMC 2020*, online, 2020.

[43] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations," in *Proc. ICML 2020*, online, 2020.