SEMI-SUPERVISED CONTRASTIVE LEARNING OF MUSICAL REPRESENTATIONS

Julien Guinot[1,2] **Elio Quinton**[2] **György Fazekas**[1]
[1] Centre for Digital Music, Queen Mary University of London, U.K.
[2] Music & Audio Machine Learning Lab, Universal Music Group, London, U.K.
j.guinot@qmul.ac.uk

ABSTRACT

Despite the success of contrastive learning in Music Information Retrieval, the inherent ambiguity of contrastive self-supervision presents a challenge. Relying solely on augmentation chains and self-supervised positive sampling strategies can lead to a pretraining objective that does not capture key musical information for downstream tasks. We introduce semi-supervised contrastive learning (SemiSupCon), a simple method for leveraging musically informed labeled data (supervision signals) in the contrastive learning of musical representations. Our approach introduces musically relevant supervision signals into self-supervised contrastive learning by combining supervised and self-supervised contrastive objectives in a simpler framework than previous approaches. This framework improves downstream performance and robustness to audio corruptions on a range of downstream MIR tasks with moderate amounts of labeled data. Our approach enables shaping the learned similarity metric through the choice of labeled data that (1) infuses the representations with musical domain knowledge and (2) improves out-of-domain performance with minimal general downstream performance loss. We show strong transfer learning performance on musically related yet not trivially similar tasks - such as pitch and key estimation. Additionally, our approach shows performance improvement on automatic tagging over self-supervised approaches with only 5% of available labels included in pretraining.

1. INTRODUCTION

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning structured representations of data without the need for costly and time-consuming labeling. SSL approaches have achieved competitive performance on downstream tasks with minimal labeled data in many domains [1–8]. In the field of Music Information Retrieval (MIR), the complexity of labeling for many tasks - due to the high technicality and subjectivity - underscores the importance of such self-supervised methods [5,8–14]. Instance-discriminative SSL specifically, such as contrastive learning, has proven to be effective in learning meaningful representations for a multitude of downstream tasks [8, 9, 15]. However, major design choices such as positive mining strategies and augmentations are crucial to downstream performance [8, 16–19], and selecting a strategy for a given task remains a challenge, prompting the reintroduction of supervision within the SSL framework. In MIR, the key notion of "similarity" in contrastive learning can derive from a variety of musical attributes. Guiding the model towards a musically informed similarity metric is an objective that may be achieved by leveraging supervised labeled data, i.e. *supervision signals*.

In this work, we propose a novel semi-supervised contrastive learning method, SemiSupCon. Our method leverages both unlabeled and labeled data for contrastive learning, an extension of Contrastive Learning of Musical Representations (CLMR) in the music domain [8] and SupCon [20] in Computer Vision. Our approach differs from previous attempts at combining self-supervised contrastive learning with an auxiliary supervision signal in that it is the first to our knowledge to implement a fully-contrastive semi-supervised learning pipeline. The simple machinery of this method allows for leveraging new supervision signals beyond labels within the contrastive objective.

Briefly, the contributions of this work are the following: (1) We propose an architecturally simple extension of self-supervised and supervised contrastive learning to the semi-supervised case with the ability to make use of a variety of supervision signals. (2) We show the ability of our method to shape the representations according to the support supervision signal used for the learning task with minimal performance loss on other tasks. (3) We propose a representation learning framework with low-data regime potential and higher robustness to data corruption. Our implementation and experiments are made publicly available at https://github.com/Pliploop/SemiSupCon

2. RELATED WORK

Self-supervised learning aims to learn representations that capture the semantic structure of data without labels in order to utilize these representations on downstream tasks. Among self-supervised learning approaches, Contrastive

Learning teaches a model to identify augmented samples originating from the same data point amongst distractor negative samples [1, 8]. Beyond its success in neighboring fields, MIR and audio representation learning have largely benefited from Contrastive Learning approaches [2, 8, 9, 21–23]. From the implementation of CLMR, several works have expanded on contrastive learning for music, with competitive results on many downstream tasks and in multiple modalities [9, 10, 13, 24, 25]. One of the key challenges of contrastive learning is establishing an effective positive mining strategy to select positive and negative samples [16–18]. Previous studies show that both the positive mining strategy and the augmentation chain are crucial toward the performance on a given downstream task [16–19] - an inappropriate sampling strategy can lead to treating similar samples as negatives, to the detriment of downstream performance [26–28]. In MIR specifically, even the temporal proximity of two positive segments within an audio clip is influential on downstream performance depending on the task, as shown in [18]. Previous works have attempted to design domain-appropriate strategies for music and audio contrastive learning, including auxiliary similarity metrics [24, 29–31], weak supervision [15, 32–34], as well as music-specific preprocessing and augmentations [8, 10, 25].

Self-supervision is inherently limited by the ability of the positive mining strategy to select semantically relevant positives. Some approaches have attempted to reintroduce supervision signals for positive mining within the contrastive objective to reduce noise induced by self-supervised pseudolabels. SupCon [20] introduces supervised contrastive learning, which uses class labels to mine positives. Other approaches have extended contrastive learning to the semi-supervised regime by leveraging both labeled and unlabeled data. However, these approaches often use complex machinery, such as auxiliary classification modules or multiple losses [29, 35–37], making them inflexible and difficult to balance with regard to the supervision signal. Recently, in MIR, *Akama et. al* [29] employ contrastive learning as an auxiliary loss for automatic tagging, with improved results over supervision alone.

3. METHODS

3.1 Self-Supervised contrastive learning

In the SSL setting for contrastive learning [1, 8], each sample in a N-sample batch is augmented into two views through a stochastic augmentation chain. Let B be a batch of these augmented views x_i. Indices $i \in I = \{1, 2...2N\}$ represent the index of a data point in the batch (anchor). $p(i)$ is the index of the augmented data point originating from the same original sample as the anchor (positive sample). $N(i)$ is the set of negatives: data points in the augmented batch excluding the anchor and positives: $N(i) = I \setminus \{i, p(i)\}$. Let z_i be the embedded representation of the data point by an encoder $E : x \mapsto E(x) \in \mathbb{R}^{d_E}$ and a projection head $g : E(x) \mapsto g(E(x)) = z_i \in \mathbb{R}^{d_g}$ into the contrastive latent space. In the SSL setting, the objective function for the contrastive method is the normalised temperature-scaled cross-entropy loss [1] between samples i and $p(i)$ for all pairs in the batch:

$$\mathcal{L}^i_{ssl} = -\log \frac{\exp(sim(z_i, z_{p(i)})/\tau)}{\sum_{n \in N(i) \cup \{p(i)\}} \exp(sim(z_i, z_n/\tau))} \quad (1)$$

Where τ is a temperature hyperparameter, sim is a similarity function - usually, cosine similarity [1, 8]. For the sake of brevity we notate $\sigma_{i,j} = \exp(sim(z_i, z_j)/\tau)$ in the rest of this work.

3.2 Supervised contrastive learning

In the supervised setting [20], the set of *supervised* positives $P_s(i)$ are now defined by the label information in the set of labels y_i: $P_s(i) = \{p \in I | y_p = y_i\} \setminus i$. As in [20], the supervised contrastive loss objective is given by:

$$\mathcal{L}^i_{sl} = \frac{-1}{|P_s(i)|} \sum_{p \in P_s(i)} \log \frac{\sigma_{i,p}}{\sum_{n \in N(i) \cup P_s(i)} \sigma_{i,n}} \quad (2)$$

The contrastive matrix $\mathbf{M}$ is constructed by leveraging class information obtained by mining the labels, i.e. if two samples x_i and x_j are in the same category then $\mathbf{M}_{i,j} = 1$.

3.3 Semi-supervised Contrastive Learning

Let $\mathcal{U}$ be a set of unlabeled samples, and $\mathcal{S}^*$ be a set of labeled samples. We sample a proportion p_s of the labeled dataset for training such that $|\mathcal{S}| = p_s|\mathcal{S}^*|$. Let $\mathcal{A} = \mathcal{U} \cup \mathcal{S}$ be the set of all data points seen during training. During training, we use both labeled and unlabeled data points by sampling batches B comprised of proportions b_s (resp. $1 - b_s$) of labeled (resp. unlabeled) samples. $P_s(i) = \varnothing$ if i is the index of an unlabeled data point. We now define our semi-supervised contrastive loss, with $P_A(i) = P_s(i) \cup P_u(i)$, where $P_u(i)$ is the set of self-supervised positives $\big(\{p(i)\}$ in Eq. 1$\big)$:

$$\mathcal{L}^i_{sem} = \frac{-1}{|P_A(i)|} \sum_{p \in P_A(i)} \log \left(\frac{\sigma_{i,p}}{\sum_{n \in N(i) \cup P_A(i)} \sigma_{i,n}} \right) \quad (3)$$

With the inclusion of both sets of positives, we generalize to both labeled and unlabeled data in our representation learning task: Note that if $\mathcal{U} = \varnothing$ or $\mathcal{S} = \varnothing$, the semi-supervised contrastive loss reverts back to the fully-supervised loss or the fully self-supervised loss (as $P_s(i) = \varnothing$) respectively. The approach is shown Figure 1.

This approach differs from simply adding the supervised and self-supervised contrastive losses together, as our objective maintains the number of samples to discriminate against in the self-supervised setting by leveraging labeled data as negatives for the self-supervised samples.

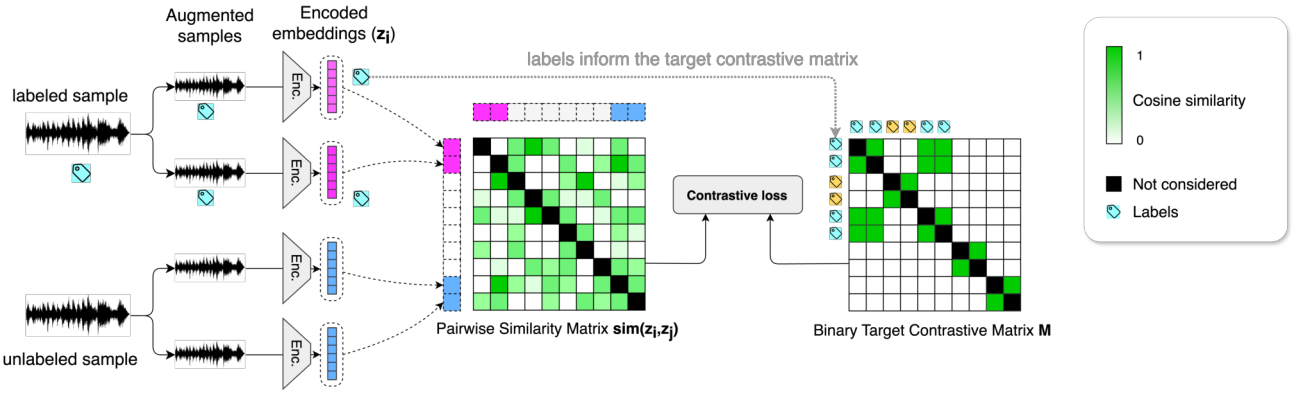


Figure 1: Semi-Supervised Contrastive Learning. The sparsely labeled dataset contains a mix of unlabeled data and labeled data. Given a batch, available labels (blue and yellow tags) are used to augment **M**. Unlabeled samples degenerate back to the self-supervised case. Loss is computed between the pairwise similarity matrix from the encoded embeddings and the target matrix using Equation 3

3.3.1 Extension to other supervision signals

The range of supervision signals this method can leverage are limited only by the ability to construct the target contrastive matrix. In this, SemiSupCon can make use of support data beyond single label multiclass tasks. To demonstrate this, we devise two strategies for training on MagnaTagATune [38], which are studied in Section 5.4. For a multi-label signal, if $C \in \mathbb{N}$ labels coincide between two samples, we set the the corresponding index in the target contrastive matrix $\mathbf{M}_{i,j} = 1$. The criterion C is a hyperparameter which is studied in Section 5.4. By default we use $C = 1$, i.e., if any labels coincide between two samples they are considered as positives.

Further, we can construct a target continuous similarity metric factor $\alpha_{i,j}$ which denotes the degree of "semantic similarity" between the samples by weighing the common classes by the total number of labels:

$$\alpha_{i,j} = \frac{2C_{i,j}}{(C_i + C_j)}$$

$C_{i,j}$ is the number of common classes for x_i and x_j, C_i and C_j are the number of classes of x_i and x_j. The similarity term $\sigma_{i,j}$ is then weighted by $\alpha_{i,j}$ in Eq. 3.

4. EXPERIMENTS AND RESULTS

4.1 Datasets

For our experiments, we use The Free Music Archive (FMA) dataset [39] as a self-supervised dataset, i.e., we do not use its labels. To match the scale of the supervised datasets, we elect to use the *medium* subset, containing 25000 clips of 30 seconds of audio.

We utilize several labeled datasets as support labeled data for training and for evaluation to demonstrate the cross-domain usefulness of SemiSupCon. For automatic tagging and most of our experiments, we use MagnaTagATune (MTAT) [38] as labeled data as a proxy evaluation of general music understanding. We reproduce the canonical 12:3:1 train-test-validation splits [8]. We use MTG-Jamendo (all subsets, including the top 50 tags, genre, mood/theme, and instrument) [40] as another tagging dataset. We use NSynth [41] for pitch and instrument classification of short snippets, and MedleyDB [42, 43] for instrument classification with longer audio clips than NSyth. We use Giantsteps [44] as a key classification dataset - as in [45], we use the original dataset as our training set and the MTG-Giantsteps dataset as our test set. For genre classification, we use the fault-filtered GTZAN dataset [46, 47]. We use the VocalSet dataset [48] for two additional tasks: singer identification and technique classification. Finally, we regress Arousal (A) and Valence (V) on EmoMusic [49] as a downstream evaluation task only, with the same train-test split as [45].

4.2 Model input, augmentation chain

As in [8, 9], we crop 2.7 second segments of mono 22050kHz audio as input to the encoders, SampleCNN [50] or TUNe+ [9]. We sample and augment 2 adjacent non-overlapping segments as positives. The dimensions of the encoders and the 2-layer ReLU-nonlinear projection head are $d_E = 512$ and $d_g = 128$. We implement a stochastic augmentation chain similar to CLMR [8], TUNe [9], and [10]. In order, we apply (Table 1):

Augmentation	**probability**	**parameter**	**Min/Max**	**unit**
Gain	0.4	Gain	-15‡ / 5†	dB
Polarity inv.	0.6	-	-	-
Colored Noise	0.6	Signal/noise ratio	3‡ / 30‡	dB
		Spectral decay	-2‡ / 2†	dB/octave
Filtering	(One of)			
Low pass	0.3	Cutoff	0.15‡ / 7‡	kHz
High pass	0.3	Cutoff	0.2† / 2.4†	kHz
Band pass	0.3	center frequency	0.2 / 4 †	kHz
		Bandwidth fraction	0.5† / 2	-
Band cut	0.3	center frequency	0.2 / 4†	kHz
		Bandwidth fraction	0.5† / 2	-
Pitch shifting	0.6	transpose	-4‡ / 4†	semitones
Delay	0.6	reflection time	100‡ / 500	ms
		reflections	1† / 3†	-
		attenuation	-6† / -3†	dB/reflection
		wet/dry factor	0.25† / 1	-

Table 1: Training augmentation chain. Only one amongst the four frequency filters is applied at once. Ranges denoted with † (resp. ‡) are subject to increasing (resp decreasing) in Subsection 5.3

Ours	$b_s = p_s$	AUROC ↑ SampleCNN (‡)	AUROC ↑ TUNe+ (★)	AP ↑ ‡	AP ↑ ★
Self-Supervised	0	88.8	88.9	41.6	41.6
Semi-Supervised	0.05	89.4	89.4	42.5	42.1
	0.1	89.5	89.4	42.2	42.2
	0.25	89.5	89.4	42.5	42.8
	0.5	89.7	89.5	42.9	43.3
	0.75	89.9	89.8	43.3	43.5
	0.5/1	89.8	89.8	43.1	43.0
Supervised	1	**90.3**	90.1	44.3	**44.6**
Literature					
SampleCNN [8]	-	89.3* (88.6† [8])		41.2* (34.4† [8])	
CLMR_{FMA} [8]	-	86.6†		31.2†	
TUNe+ [9]	-	89.2†		36.6†	
MERT [5]	-	91.0†		39.3†	

Table 2: Performance on automatic tagging. Results denoted by † are reported in their original paper. In our experiment, we constrain $p_s = b_s$ except for one run where $p_s = 1, b_s = 0.5$. We trained our own end-to-end supervised SampleCNN with the same compute budget as SemiSupCon and report results with *.

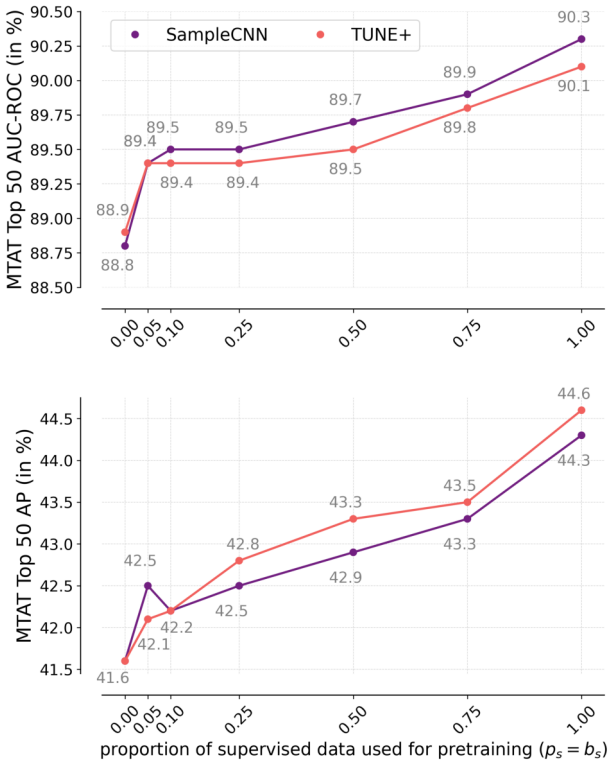


Figure 2: Evolution of AUROC and AP on MTAT probing with proportion of supervised MTAT data used for training.

4.3 Training and evaluation details

For our baseline models, we adopt a training setup similar to TUNe [9] and CLMR [8]. Models are trained for 200k steps on semi-supervised batches sampled from MagnaTagATune as labeled data and FMA-Medium as unlabeled data [38, 39] using the Pytorch Adam optimiser with a learning rate of $1e^{-4}$. For ablation and variation studies, we train our models for 50k steps unless otherwise stated. All models are trained with $\tau = 0.1$ with a non-augmented batch size of 96 on a single RTX A5000 GPU unless otherwise specified. We report steps instead of epochs to standardise the amount of data seen during training.

To evaluate pretrained models, we freeze the encoder and discard the projection head. Frozen representations are fed into a 2-layer ReLU-nonlinear MLP for probing on downstream tasks. For probing, we use the Adam optimizer with a learning rate 0.0003 and an early stopping mechanism conditioned on validation loss. For automatic tagging tasks, we report area under receiver-operator curve (AUROC) and mean Average Precision (AP). For classification tasks, we report top-1 accuracy except for key classification: the metric for this task is a weighted score taking into account reasonable errors [45] - We use the `mir_eval` [51] implementation for evaluation . For emotion regression we report R^2 values between predicted and actual values.

5. RESULTS

5.1 Automatic tagging with semi-supervised contrastive learning

We train a self-supervised baseline, a supervised contrastive baseline with and without augmentations, an end-to-end supervised baseline using the sampleCNN architecture, and five variants of our semi-supervised approach with different proportions of labeled data ($p_s \in [0.01, 0.05, 0.1, 0.2, 0.5]$) for Automatic Tagging on MTAT. MTAT labels augment the contrastive matrix **M** with positives in the case of supervised or semi-supervised pretraining. We vary the in-batch and global proportion of supervised data b_s and p_s simultaneously. We report results on the same task in the literature in Table 2 for comparable datasets and training scales.

When trained for 200k steps, the supervised contrastive model is competitive with larger self-supervised approaches. Furthermore, it outperforms both our implementation and the results claimed in CLMR for self-supervised contrastive and end-to-end supervised models. Figure 2 shows the influence of $p_s = b_s$ on AUROC and AP. As the proportion of supervised data increases, so does the performance on the downstream evaluation. Including only 5% of labeled data leads to an increase in performance from 88.8 to 89.4 in AUROC. For our experiment with $p_s = 1$ and $b_s = 0.5$, both architectures perform worse than $p_s = b_s = 0.75$, as the model has seen 100k steps of supervised data versus 150k.

5.2 Influence of pretraining labeled dataset

In this experiment, we pre-train multiple semi-supervised models using datasets described in Section 4.1 as support labeled data and FMA as unlabeled data - one model per dataset, each for 50000 steps. We then freeze all models and train shallow MLP probes on all downstream tasks for each model. We train a self-supervised baseline for comparison. Semi-supervised approaches are trained with $b_s = 0.5$ and $p_s = 1$. Table 3 shows these results.

Semi-supervised training on the target dataset always surpasses the self-supervised baseline by a significant mar-

Target Dataset		MTAT		Jamendo				NSynth		Giantsteps	GTZAN	VocalSet		MedleyDB	Emo
Subset		50	All	50	Genre	Mood	Inst.	Pitch	Inst.	Key	Genre	Tech.	Singer	Inst.	A/V
Metrics		AUROC						Acc.		Acc_w	Acc.				R^2_V / R^2_A
Self-Supervised															
FMA		88.4	86.2	80.1	83.3	74.0	71.6	36.8	51.7	13.5	65.5	53.8	71.1	56.5	46.7/71.5
Semi-Supervised $b_s = 0.5$															
MTAT	50	**89.3**	86.8	80.0	83.4	73.8	73.3	34.5	46.9	11.3	65.5	53.2	70.0	67.3	44.3/65.9
	All	89.1	**87.5**	80.3	83.2	74.1	73.0	34.0	51.0	14.9	68.2	52.4	72.9	**72.8**	41.6/76.2
Jamendo	50	88.6	86.6	**81.5**	83.4	74.6	72.5	33.8	50.0	14.7	**74.1**	52.1	71.7	62.0	50.1/**77.9**
	Genre	88.6	86.3	80.5	**84.0**	74.6	71.5	33.4	50.2	14.6	72.8	52.0	74.6	66.3	48.2/70.3
	Mood	88.3	86.6	81.0	83.0	**74.7**	72.3	38.2	47.7	14.9	71.3	53.5	71.4	60.9	48.0/73.0
	Instrument	88.4	86.3	80.8	83.1	74.0	71.6	37.2	52.5	14.9	69.3	54.5	67.9	63.0	**52.4**/70.6
NSynth	Pitch†	88.3	86.3	79.7	82.6	73.5	72.0	**79.0**	48.6	20.1	65.5	56.9	75.6	64.1	37.5/66.6
	Inst.	88.6	85.7	79.6	82.7	73.3	71.7	26.6	**59.6**	16.0	67.2	57.3	72.3	66.3	40.3/75.0
GiantSteps	Key†	87.7	85.0	79.0	82.1	73.0	70.5	50.8	51.3	**22.3**	69.3	54.1	71.4	61.2	39.6/63.6
GTZAN	Genre	88.8	86.8	80.9	83.9	74.1	71.5	38.6	46.9	16.3	74.0	53.4	71.7	66.3	28.7/56.4
VocalSet	Technique	88.7	86.7	79.6	82.5	73.3	71.0	46.0	53.5	12.1	63.5	**63.0**	77.8	67.3	41.5/70.1
	Singer	88.9	86.2	80.1	82.6	73.6	72.8	45.2	52.4	15.3	67.2	54.0	**87.1**	69.6	54.3/74.6
MedleyDB	Instrument	88.6	87.0	80.2	82.6	73.6	**73.8**	32.0	48.8	13.2	62.1	58.6	74.3	62.0	41.6/74.8
SOTA		92.7 [12]	95.4 [34]	84.3 [13]	88.0 [14]	78.6 [13]	78.8 [52]	94.4 [5]	78.2 [53]	74.3 [54]	86.9 [55]	76.9 [5,45]	87.5 [5,45]	-	61.7/76.3 [5,14]
CLMR [45]		89.5		81.3	84.6	73.5	73.5	47.0	67.9	14.8	65.2	58.1	49.9	-	44.4/70.3

Table 3: Results for cross-task evaluation. Models are trained for 50k steps on FMA [39] as the self-supervised dataset and support supervised datasets (rows), and evaluated on target datasets (columns). Giantsteps†, NSynth† are trained without pitch shifting augmentation. Results in bold are the best results obtained for evaluation on a target dataset. SOTA results are included for illustration purposes, but do not necessarily leverage comparable methodologies.

gin when evaluating on the same dataset - with minimal loss of performance on other downstream tasks.

Some complementary tasks improve performance on other downstream datasets, proving semi-supervised contrastive learning a viable transfer learning strategy. Expectedly, training on genre tagging data increases out-of-domain performance on genre classification, instrument tagging on instrument classification, etc. Training on mood data from MTG-Jamendo provides a performance boost on emotion regression. A notable example is the improvement in performance on NSynth pitch when training with key data as support labeled data and *vice versa*. This demonstrates an improvement in the understanding of *pitch* by the model on tasks which are musically related but not trivial transfer learning instances. Most importantly, this occurs without performance loss on general music understanding, *i.e.* automatic tagging. Other musically grounded examples are pitch pretraining improving instrument classification performance and instrument pretraining improving emotion regression performance.

5.3 Robustness to in-domain data corruption

In this section, we evaluate the robustness of our semi-supervised, supervised, and self-supervised contrastive approaches to audio corruptions compared to the end-to-end baseline. We train the probing head without augmentation until convergence and evaluate the model *with* augmentations applied. We design different severity degrees of our augmentation chain (See Subsection 4.2) by applying a modifier to the application probabilities: for severity $s \in [0, 1...4]$, we scale probabilities of application of each augmentation by $s/2$ such that $s = 2$ is the chain applied during training. We sensibly multiply or divide the min and max values of each augmentation hyperparameter (see Table 1) by $s/2$. We then evaluate all models with these augmentation chains on MagnaTagATune. The results are shown in Figure 3. Contrastive approaches are more robust to in-domain corruption than end-to-end approaches - hypothetically because we train contrastive models to be invariant to such transformations through the augmentation chain - which is not an objective of the end-to-end supervised approach.

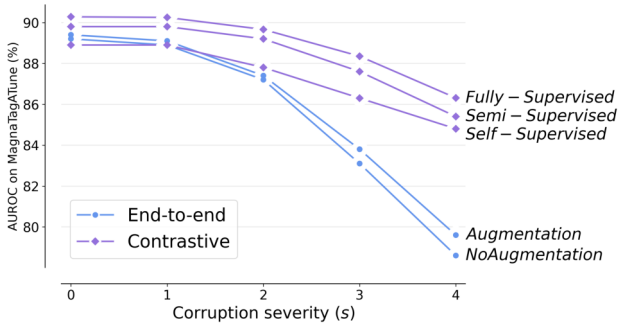


Figure 3: Effect of corruption severity on downstream performance. Contrastive models are more robust than Cross-entropy trained models.

5.4 Multilabel positive mining strategy

In this experiment, we test multiple label-based positive mining strategies. First by varying the number of common labels for mining positives - i.e. $C \in \{1, 2, 4, 6\}$. Further, we explore the "semantic weighing" strategy described in Section 3.3.1, in which the target similarity between two tracks is weighed by the number of common labels and the total number of labels. We test these strategies on both semi-supervised and supervised contrastive models. Results are reported in Table 4.

For supervised approaches, the continuous target produced by semantic weighing produces the best results, on par with 4x training steps with a criterion $C = 1$ (as shown in Table 2). In the supervised case, as the criterion in-

Positive strategy	Supervised		Semi-Supervised	
Class criterion	AUROC	AP	AUROC	AP
$C = 1$	90.1	44.2	**89.3**	41.3
$C = 2$	90.1	43.9	89.0	**41.6**
$C = 4$	89.3	42.8	89.0	41.3
$C = 6$	88.9	42.3	89.0	41.5
Weighing	**90.6**	**45.3**	88.9	**41.6**

Table 4: Multilabel positive mining strategy as described in Section 3.3.1.

creases, performance deteriorates. We hypothesise that this could be because it is an *easier* task for the model to discern that two tracks with many common tags are similar (higher C), as they likely share many attributes, therefore providing a weaker training signal. Understanding what links two tracks from a single tag is more challenging and appears to yield more robust representations. The continuous "relative similarity" target created by the weighing strategy is a more nuanced task and appears to be a stronger supervision signal. This guides the model towards more robust representations, which explains the higher performance. In the semi-supervised case, we speculate that the binary self-supervision signal overpowers the continuous target as a less nuanced objective with harsher penalties for failure. These penalties could overpower softer penalties from the continuous target in the loss, preventing optimal convergence. Future work should focus on understanding and reconciling these aspects of the semi-supervised approach to leverage other continuous signals.

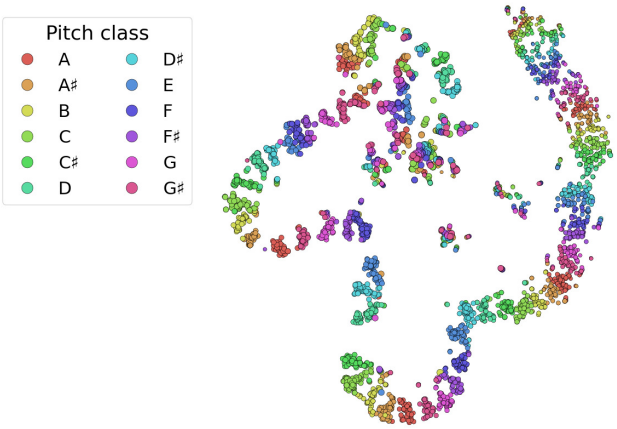
5.5 Qualitative analysis

The results reported in Section 5.2 show that performance on downstream tasks improves when labels from a related task are used for model training, with minimal loss of performance on other tasks. We hypothesise that the internal latent representations are given structure relative to the supervision signal while maintaining the semantic structure given by the self-supervision signal. To illustrate this, we perform t-SNE dimension reduction on embeddings produced by the semi-supervised model from Table 3 trained with NSynth (Figure 4a) as support labeled data and fully self-supervised (Figure 4b) evaluated on the test set of NSynth-pitch.

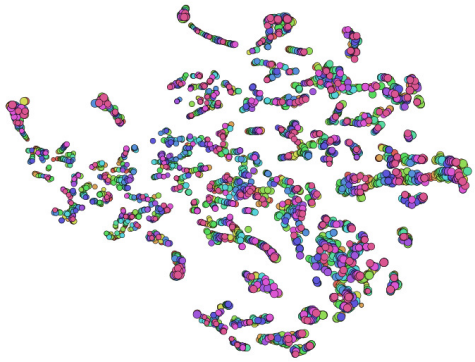
In the set of Figures 4, the latent spaces for the NSynth test set produced by these two models are shown. When pretrained on NSynth-pitch, the latent space is highly organized. Separability by class is much clearer than when pretrained on FMA. We notice that several musical structures emerge in this latent space. Notably, octaves go from low to high clockwise. Pitches that are "similar" are close together, i.e., semitones and octaves of the same pitch class.

6. CONCLUSION AND FUTURE WORK

We presented SemiSupCon, a simple method for leveraging both supervision and self-supervision signals in contrastive representation learning. By leveraging reduced amounts of labeled data during pretraining, SemiSupCon outperforms end-to-end comparable supervised baselines



(a) Latent embeddings of the NSynth-pitch test set from a semi-supervised model trained on FMA+NSynth-pitch



(b) Latent embeddings of the NSynth-pitch test set from a self-supervised model trained on FMA

Figure 4: Exploration of the NSynth pitch latent space. Octaves are denoted by size and pitch class by the color of the dot. Each dot is a full audio sample

on downstream tasks. We find that SemiSupCon is more robust to data corruption at inference compared to end-to-end supervised methods. Additionally, SemiSupCon can utilize various supervision signals with minimal performance loss on out-of-domain tasks and achieve performance transfer on similar tasks. While performance gains might seem moderate on automatic tagging for instance, other downstream tasks show more distinct improvements. Furthermore, the contrastive objective can lead to explicitly structured latent spaces with emergent musical structures - enhancing the musical interpretability of latent spaces by design of the support supervision signal - i.e. labeling small amounts of data.

Future work will focus on exploring additional supervision signals and tasks such as perceptual metrics, tempo estimation, and chord estimation. Other avenues include leveraging the low-data proficiency of SemiSupCon for human-in-the-loop representation learning. The architecture of SemiSupCon being very flexible, it can be further adapted to multimodal approaches or hierarchical representation learning. A more comprehensive exploration of the influence of the proportion of labeled data and the exact effect of labels and contrastive matrix sparsity on downstream performance will also be undertaken.

7. ACKNOWLEDGMENT

This work is supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (EP/S022694/1) and Universal Music Group.

8. REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi *et al.*, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning (ICML)*. PMLR, 2020, pp. 1597–1607.

[2] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[3] J.-B. Grill, F. Strub, F. Altché *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[4] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 15 750–15 758.

[5] L. Yizhi, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "MERT: Acoustic music understanding model with large-scale self-supervised training," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2023.

[6] Y. Gong, C.-I. Lai, Y.-A. Chung *et al.*, "SSAST: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

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

[8] J. Spijkervet and J. A. Burgoyne, "Contrastive Learning of Musical Representations," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2021, pp. 673–681.

[9] M. A. V. Vásquez and J. A. Burgoyne, "Tailed U-Net: Multi-scale music representation learning." in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2022, pp. 67–75.

[10] H. Zhao, C. Zhang, B. Zhu *et al.*, "S3T: Self-supervised pre-training with swin transformer for music classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 606–610.

[11] Y. Li, R. Yuan, G. Zhang, Y. MA, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He *et al.*, "MAP-Music2Vec: A simple and effective baseline for self-supervised music audio representation learning," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2022.

[12] M. Won, Y.-N. Hung, and D. Le, "A foundation model for music informatics," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1226–1230.

[13] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, and A. Ehmann, "Supervised and unsupervised learning of audio representations for music understanding," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2022, pp. 256–263.

[14] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2021, pp. 88–96.

[15] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2022, pp. 640–649.

[16] T. Xiao, X. Wang, A. A. Efros, and T. Darrell, "What should not be contrastive in contrastive learning," in *The 9th International Conference on Learning Representations (ICLR)*, 2021.

[17] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What Makes for Good Views for Contrastive Learning?" in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 6827–6839.

[18] J. Choi, S. Jang, H. Cho *et al.*, "Towards proper contrastive self-supervised learning strategies for music audio representation," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[19] M. C. McCallum, M. E. Davies, F. Henkel, J. Kim, and S. E. Sandberg, "On the effect of data-augmentation on local embedding properties in the contrastive learning of music audio representations," *arXiv preprint arXiv:2401.08889*, 2024.

[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[21] H. Al-Tahan and Y. Mohsenzadeh, "CLAR: Contrastive learning of auditory representations," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2530–2538.

[22] Y.-A. Chung, Y. Zhang, W. Han *et al.*, "W2V-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[23] E. Fonseca, D. Ortego, K. McGuinness *et al.*, "Unsupervised contrastive learning of sound event representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 371–375.

[24] D. Yao, Z. Zhao, S. Zhang *et al.*, "Contrastive Learning with Positive-Negative Frame Mask for Music Representation," in *Proceedings of the ACM Web Conference 2022*, Apr. 2022, pp. 2906–2915.

[25] C. Garoufis, A. Zlatintsi, and P. Maragos, "Multi-Source Contrastive Learning from Musical Audio," no. arXiv:2302.07077. arXiv, May 2023.

[26] H. Guo and L. Shi, "Ultimate Negative Sampling for Contrastive Learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5.

[27] S. Ge, S. Mishra, C.-L. Li, H. Wang, and D. Jacobs, "Robust Contrastive Learning Using Negative Samples with Diminished Semantics," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 27 356–27 368.

[28] T. Huynh, S. Kornblith, M. R. Walter, M. Maire, and M. Khademi, "Boosting Contrastive Self-Supervised Learning with False Negative Cancellation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 986–996.

[29] T. Akama, H. Kitano, K. Takematsu *et al.*, "Auxiliary self-supervision to metric learning for music similarity-based retrieval and auto-tagging," *PLOS ONE*, vol. 18, no. 11, p. e0294643, Nov. 2023.

[30] P. Manocha, Z. Jin, R. Zhang *et al.*, "CDPAM: Contrastive Learning for Perceptual Audio Similarity," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 196–200.

[31] P. Alonso-Jiménez, X. Favory, H. Foroughmand, G. Bourdalas, X. Serra, T. Lidy, and D. Bogdanov, "Pre-training strategies using contrastive learning and playlist information for music classification and similarity," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[32] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "Coala: Co-aligned autoencoders for learning semantically enriched audio representations," in *Self-supervision in Audio and Speech Workshop, International Conference on Machine Learning (ICML)*, 2020.

[33] A. Ferraro, X. Favory, K. Drossos *et al.*, "Enriched Music Representations with Multiple Cross-modal Contrastive Learning," *IEEE Signal Processing Letters*, vol. 28, pp. 733–737, 2021.

[34] 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. of International Society for Music Information Retrieval (ISMIR)*, 2022.

[35] B. Kim, J. Choo, Y.-D. Kwon, S. Joe, S. Min, and Y. Gwon, "Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning," in *34th conference on Neural Information Processing Systems (NEURIPS) Workshop*, 2021.

[36] Y. Zhang, X. Zhang, J. Li, R. Qiu, H. Xu, and Q. Tian, "Semi-supervised contrastive learning with similarity co-calibration," *IEEE Transactions on Multimedia*, 2022.

[37] F. Yang, K. Wu, S. Zhang, G. Jiang, Y. Liu, F. Zheng, W. Zhang, C. Wang, and L. Zeng, "Class-aware contrastive semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 14 421–14 430.

[38] 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. of International Society for Music Information Retrieval (ISMIR)*, 2009, pp. 387–392.

[39] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 316–323.

[40] D. Bogdanov, M. Won, P. Tovstogan *et al.*, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, 2019.

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

[42] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 155–160.

[43] R. M. Bittner, J. Wilkins, H. Yip, and J. P. Bello, "Medleydb 2.0: New data and a system for sustainable data collection," *ISMIR Late Breaking and Demo Papers*, p. 36, 2016.

[44] P. Knees, Á. Faraldo Pérez, H. Boyer, R. Vogl, S. Böck, F. Hörschläger, M. Le Goff *et al.*, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2015, pp. 364–70.

[45] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, Y. Liu, J. Huang, Z. Tian, B. Deng *et al.*, "Marble: Music audio representation benchmark for universal evaluation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[46] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[47] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[48] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset." in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2018, pp. 468–474.

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

[50] 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.

[51] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common mir metrics." in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2014, p. 2014.

[52] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from discogs," in *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2022, pp. 825–33.

[53] L. Wang, P. Luc, Y. Wu *et al.*, "Towards Learning Universal Audio Representations," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 4593–4597.

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

[55] P. Alonso-Jiménez, L. Pepino, R. Batlle-Roca, P. Zinemanas, D. Bogdanov, X. Serra, and M. Rocamora, "Leveraging pre-trained autoencoders for interpretable prototype learning of music audio," *arXiv preprint arXiv:2402.09318*, 2024.