Self-Supervised and Human-in-the-loop Learning for Musical Audio: Towards Expert and Interpretable Musical Representations

Author:
Julien Guinot

Supervisor:
Prof. George Fazekas[1]

Industry Supervisor:
Dr. Elio Quinton[2]

Co-supervisor:
Prof. Emmanouil Benetos

Independent Assessor:
Prof. Johan Pauwels

Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London

[1]Queen Mary University of London
[2]Universal Music Group

Abstract

The field of Music Information Retrieval (MIR) has long suffered from the absence of large labeled datasets, which are costly and time-consuming to annotate due to the subjectivity and complexity of the data. Self-Supervised learning has emerged as a potent solution to circumvent this bottleneck in the development of music Machine Learning applications, with great success in recent years. However, the learned representations from these self-supervised approaches are still largely non-specialized to music, difficult to explain, and lack interactivity and adaptability for experts to interact with them. This stunts the potential for both performance and relevance of these models.

This report presents the background and planned research for a Ph.D. research project aiming to improve learned representations for music by machine learning models in these key areas. A set of research questions aiming to solve the problems that current representation learning methods for music are suffering is devised, and a research plan including novel architectures, methodologies, and tasks, as well as viable avenues of research to tie these explorations together is proposed to answer these questions. Current preliminary results are briefly explored, and following these results a detailed plan for the next year of the PhD as well as a high-level overview of the plan for the remainder of the project is laid out.

Contents

Chapter 1

Introduction

1.1 Motivation

Machine learning tasks are infamously reliant on large amounts of data in the supervised setting. Although in some cases large corpora of labeled or captioned data are available, more often than not the scarcity of annotated data is a bottleneck in the development of machine learning applications. Furthermore, the inclusion of rare instances and the development of machine learning model vocabularies are hampered by the nature of existing datasets, which include only a small repertoire of possible outcomes. This stunts the generalization potential and the inclusivity of supervised machine learning models. To counteract this natural bottleneck for machine learning models, techniques and methods have been developed distancing the learning process from the need for large fully annotated datasets. Self-supervised learning, Few/zero-shot learning, multimodal weak supervision, and Human-in-the-loop are but a few methods that have helped alleviate the need for labeled data. These developments have been leveraged to great lengths and have notably led to recent leaps in the field of computer vision and Natural Language Processing.

These methods, seeking to learn a representation of the data through tasks that remove the need for large corpora of labeled data, are essential for music and the global music industry. "Better" representations of music benefit listeners, users, creators, producers, mixing engineers, and global companies. Machine learning tasks for music which previously relied on relatively small corpora of labeled data (automatic tagging, generation, automatic chord estimation, source separation, etc.) would then be able to exploit the large amount of **unlabeled** data, which would both benefit the performance of these tasks if the representations are meaningful, and improve upon ethical considerations for AI in music such as cultural inclusivity and explainability in model outputs.

Despite the glaring importance of the development of these techniques for musical audio, such research has not seen comparable explosive development to CV and NLP, or even general audio. Though initiatives towards self-supervised learning for musical audio have emerged, most are relatively recent. Multiple reasons can be attributed to this slower development: Firstly, music is a very specific sub-field of audio with structure, hierarchy, and perceptual links to emotion, creativity, and culture. Furthermore, music possesses a subset of structural "rules" very specific to the medium relating to its structure and the hierarchical nature of its ontology and the ontology of its linked modalities. These rules might not be captured by advances in SSL for general audio, or even by simply training general audio models on "only" musical data. This, the existence of copyrighting laws, and the extensive list of modalities that music is linked to renders access to large, diverse, and multimodal corpora of music difficult.

These hurdles induce music representation learning approaches lagging behind similar approaches to other domains. So far, representation learning approaches for music are mostly transferable applications from other domains without specialization to the structure and nature of music. Learned latent spaces are still mostly opaque and difficult to navigate. Multimodal datasets lack specialized and expert annotations and do not suit use-cases beyond a "novice" user / listener. Finally, despite the success of human-in-the-loop approaches for both generation and analysis in the image and natural language domains, such approaches have been slow to gain traction in the field of music, where expert annotations and feedback are often the most valuable. These shortcomings unveil a clear research path towards improving representation learning for music, to facilitate ML models realize their full potential in the field of music. This is supported by budding work toward making representation learning more appropriate for music, understandable, and relying on human feedback in the recent literature that begins to address these shortcomings.

1.2 Aim

The research proposed in this PhD project aims to move beyond the need for supervised training for music audio while keeping in mind these motivations. Through this research, I plan to develop music-specific and music-appropriate architectures and self-supervised learning methods that will improve the explainability, interpretation, and navigability of musical latent spaces. I will further leverage the support of additional modalities with the aim of adapting the representations learned by these modalities for musicians and nonnovice stakeholders of the music industry. I will aim to re-center representation learning for music around the music and the humans interacting with the music, and improve the learned representations from supervised and self-supervised tasks through human-in-the-loop approaches.

In doing so, the goal is to shift the research direction for music representation learning - a crucial paradigm for the development of all ways of applications of AI to music - towards a more music-centric and human-centric viewpoint. From this viewpoint, the representation learning methods and modalities used to learn a structured space for music representations are instilled by design and curation with music and signal processing domain knowledge, and these representations are further improved by interaction with human users, which help in making them more explainable and in improving their performance on downstream tasks. The objectives of this research are intertwined and feed into each other. As such, this proposal aims towards a multifaceted and rich avenue of research with promise in both feasibility, interest to the community, and downstream applications.

To this end, I will leverage existing methods of self-supervised learning and re-design them with musically coherent objectives in mind, or design new methodologies with the same objectives, both in their tasks and architecture. I will rely and improve upon curated data of additional modalities for music to further disentangle these representations and validate the usefulness of previously-mentioned architectures. Finally, I will develop novel human-in-the-loop applications - so far under-explored - for music in supervised, self-supervised, and multimodal settings.

This report is organized as follows. Section 2 covers previous related work in self-supervision for the audio and music domain, multimodal approaches for music, and human-in-the-loop approaches in both a broad sense and applied to music. This prompts a specification of the problem this research is trying to solve, which was briefly covered above but will be explicited in section 3. The research questions that emerge from this analysis and potential research directions will be covered in Section 4. Section **??** Introduces preliminary work in the early months of the Ph.D., with more details in Appendices**??**. I then give a research plan with potential topics of interest in the short-term, long-term, as well as a dissemination plan in Section 5.

Chapter 2

Related Work

This section covers an overview of current work on representation learning for music, including self-supervised learning, use of extra modalities as weak supervision signals, and human-in-the-loop approaches. For each of these areas, I first conduct a broad overview of the governing principles, works of interest in other domains, and finally existing studies on musical audio leveraging these methods.

2.1 A general overview of self supervised learning

Supervised methods in Deep Learning (**DL**) require large amounts of *labeled* data. Depending on the task, these data can be easy to acquire as they can be prohibitive. As supervised learning hit a bottleneck due to this hurdle, strategies to use data outside the target data set to train models on a specific task in a meaningful fashion became a crucial research direction. Early strategies included transfer learning and knowledge distillation, in which a model is trained on a **pretext** dataset and fine-tuned on a smaller **downstream** dataset. The intuition is that the model should learn inductive biases covering both datasets. However, domain adaptation is not trivial. Transfer learning in this fashion still requires a large labeled dataset to pretrain, so the main issue with labeled datasets is not solved.

Self-Supervised Learning (**SSL**) is the next step in that direction. The core idea of Self-Supervised Learning is to train a model on a pretext task requiring some knowledge of the underlying semantic structure of the data, *i.e.* **pseudo-labels**, created by the task itself. In doing so, the model learns a latent space that is representative of said semantic structure. The pretraining can serve as a task in and of itself (e.g image colorization, text infilling) or the learned structured representation can be used as an input for a downstream task with a reduced amount of labels by either fine-tuning (the entire model or parts of it) or probing (see Figure 2.1).

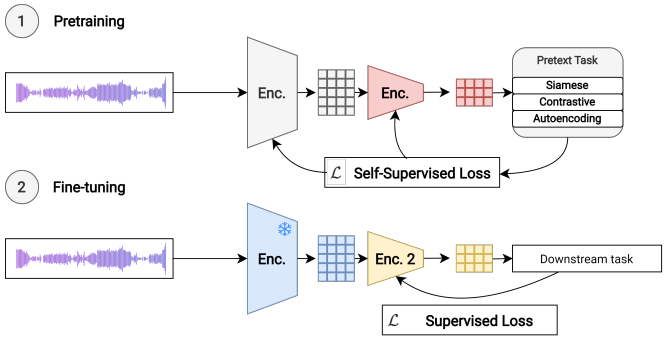
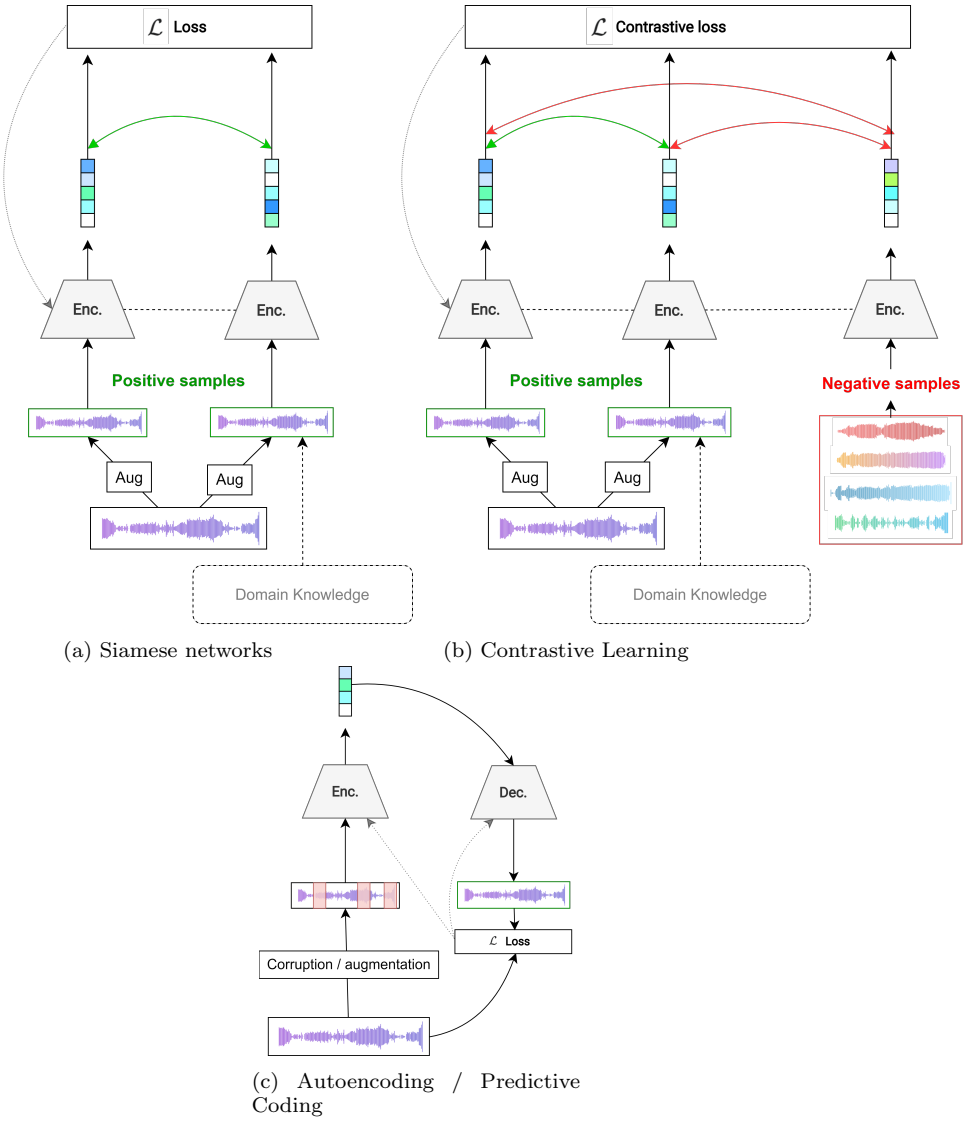


Figure 2.1: SSL pipeline overview

Pretext tasks are highly dependent on the medium, and their selection is crucial for the performance of models on the downstream task. We consider these tasks as belonging to broad categories which correspond to a broad taxonomy of learning strategies: **Siamese Networks, Contrastive Learning, and Auto-encoding** (See figure 2.2).

Siamese networks [1] rely on the assumption that different views of the same item are semantically similar and should be close in the learned latent space. Siamese networks are composed of two



(a) Siamese networks

(b) Contrastive Learning

(c) Autoencoding / Predictive Coding

Figure 2.2: An overview of SSL methods

symmetrical network branches with shared weights and parameters and aim to maximize the similarity of different views of the same sample (see Figure 2.2a). One of the conjoined networks creates a pseudo-label which the second network must maximize similarity with.

These models are prone to a phenomenon known as representation collapse, in which the task of maximizing similarity is trivial if the model always outputs the same representation. BYOL [2] deals with this issue by de-synchronizing the training of the sister models (similar also to SimSiam [3]). Barlow Twins [4] simplifies the training procedure by re-symetrizing the branches and maximizing the cross-correlation matrix of distorted views of the same sample.

Contrastive Learning [5] Directly Addresses the hurdle of representation collapse that Siamese networks face by incorporating negative samples within the training loop. The assumption that different views of the same sample are similar holds true, and in addition, negative samples (i.e samples that are not from the anchor) should be pushed away in the latent space (see figure 2.2b). Contrastive learning has had much success in that no complex machinery is needed to prevent representation collapse. SimCLR [5], [6] introduced contrastive learning by augmenting all samples in a given mini batch, yielding 2 examples of each sample. They introduce the NT-Xent Loss, pushing apart negative samples and pushing positive samples together with a similarity metric (usually cosine similarity [5], [7], [8]). Momentum Contrast encoding (MoCo[7]–[9]), uses a second momentum encoder. Representations of samples from previous batches are also used as negative samples - a technique which is discarded in

MoCo V3 [9].

Other works using contrastive learning have been met with great success in their respective fields [5], [9]–[11], with linear probing strategies approaching or matching the state of the art.

While contrastive Learning and Siamese Networks rely on similar mechanisms to learn the underlying distribution of data, **Auto-encoding** relies on a model reconstructing the input sample, shown figure 2.2c. By decoding from an information bottleneck, the bottleneck must hold meaningful representations about the data toi be able to reconstruct a representation of the input data. To avoid learning 1-to-1 mapping functions and improve generalization of the latent space in case of high-dimensional latents, denoising autoencoders [12], [13] learn to reconstruct a "clean" sample from a corrupted input. Other advances in autoencoding include Variational Autoencoders (VAEs) [14], Vector-Quantized VAEs (VQ-VAEs) constrain learned representations to a discrete codebook, reducing the issue of posterior collapse in VAEs [15]–[17]. Motivated by the large amount of image and text data available online, auto-encoding methods were popularized for NLP and CV. While originally popularized as training strategies for Language Modeling, these techniques have since been adapted to images, audio, and music.

Masked modeling, shown in figure 2.3, operates by corrupting (masking) parts of the input and having the model predict the masked content.

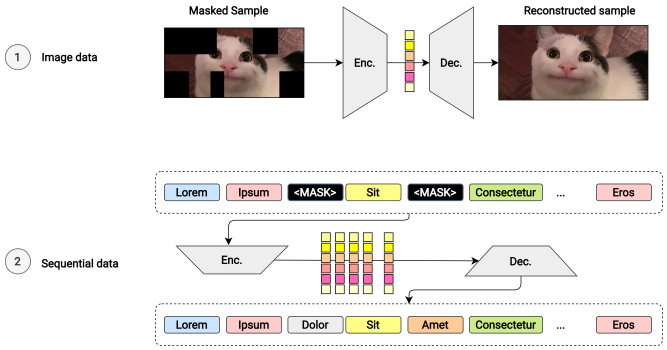


Figure 2.3: Masked Modeling

This technique was originally proposed for BERT and its variants [12], [18], but was then expanded into numerous applications [19]–[24], specifically since the advent of vision transformers [25]. Data2Vec [26], Wav2Vec2 [27] and Word2Vec [28] have all used masked modeling towards generalizing representation learning in multiple modalities. Recent developments in generative models, such as denoising diffusion probabilistic models (DDPM [29]) and generative adversarial networks (GANs [30]) have also been used as masked autoencoders [31]–[33]

Autoregressive Predictive Coding is a training strategy used in the original transformer [34] and GPT models [35] , in which a sequential model learns to reconstruct input token sequences shifted by one token into the future. In doing so, it learns a structured representation of the tokens that precede it. The autoregressive approach only makes sense when the future is not available to the model, and has had fewer applications than Masked Modeling in the vision domain [36], [37].

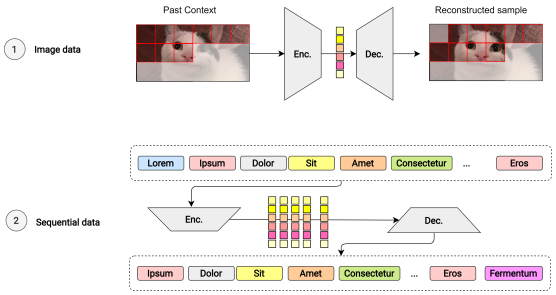


Figure 2.4: Autoregressive Predictive Coding

2.2 Self-supervised learning for general audio

In this section, we undertake a more in-depth overview of Self-Supervised Learning strategies applied to audio. Many of these applications are extensions of previously explored NLP and CV SSL techniques applied to audio, with relative success, while others have been specifically developed for Audio SSL. Although work on general-purpose audio representation has been well undertaken and highly facilitated by larger, newer audio datasets [38]–[44] - even prompting holistic benchmarks of audio representations [45], [46], a comparatively large proportion of studies for general audio are oriented towards speech due to the interest in speech-to-text and text-to-speech applications.

2.2.1 Siamese Networks

Most of the main architectures of Siamese networks (see section 2.1) were reproduced for the purpose of audio representation learning despite contrastive learning gaining traction at about the same time as siamese networks reached the field of audio representation learning. Siamese networks [1] were repurposed for content-based audio similarity learning in [47]. BYOL-A (audio) [48] implements an augmentation pipeline for BYOL [2] and achieves competitive performance with the most recent contrastive methods in the learning of audio representation. Audio Barlow Twins [49] Also achieves competitive performance on the HEAR speech and environmental sound tasks.

Outside of general audio representations, siamese networks have had success in the field of speech processing, for both representation learning [50]–[52] and speaker verification tasks [53]–[55].

2.2.2 Contrastive Learning

Early contrastive methods for general audio differ in their architecture and input formats, but rely on a similar approach to SimCLR. COLA [56], CLAR [57], and CL-SER [58] use encoded raw waveforms or mel-spectrograms for their contrastive learning task, with relative success in detecting sound events compared to supervised baselines. MultiCLAR [59] experiments with the use of both, demonstrating the usefulness of combining both representations. Some studies have adapted the Audio Spectrogram Transformer **(AST)** [60], showing promising results in the learning of general audio representation. Contrastive learning has also been used to learn deep similarity metrics, an example of which is CDPAM, which tunes a contrastive similarity metric using human feedback for perceptual evaluation [61]. Recently, it has been shown that perceptual metrics serve as robust loss functions for self-supervized learning, even on unstructured data [62].

Contrastive Predictive Coding [63], [64] introduces a new method of using a front-end encoder to compress audio waveform frames into a lower sample rate embedding sequence and applying a context model to model a global context distribution of the sequence. This idea of acoustic sequence embedding has been fundamental in the development of new audio representation learning methods.

Applications of contrastive learning on such embedded audio sequences include CPC [63], which uses a sequential model to extract a summarized representation of past sequence embeddings, considered the anchor. Positive samples are drawn from embeddings of future frames and negative samples are drawn from past frames, thus maximizing the agreement between present and future. Wav2Vec [65] and wav2vec2 [27] tweak the method from [63] slightly with multiple lookahead steps [65] and a transformer feature aggregator [27]. Vector Quantization applied to ASM converts these continuous sequences into a discrete audio tokens (we refer to this process as tokenized audio sequence modeling), which achieves significant data compression with minimal information loss.

2.2.3 AutoEncoding

SSL advances for general audio using autoencoding architectures were first explored with autoregressive predictive coding, which is sometimes referred to as Generative Pretraining [66]. These approaches have shown promise in learning speech representations [67], [68] but have not been explored much for general audio representation learning - more on such approaches applied to music in Section 2.3

Approaches using autoencoding have included leveraging internal representations of Neural Audio Codecs for representation learning. Neural Audio Codecs [69], [70] compress audio representations with the goal of compressing the signal while providing the best possible reconstruction accuracy - effectively autoencoding. Recent developments have shown the potential of Residual Vector Quantization (**RVQ**) based Neural Codecs in learning general representations [71], [72]. RVQ is a development of VQ where the signal is quantized into multiple hierarchal codebooks, recursively constructing a hierarchical discrete representation of the signal. (Figure 2.5).

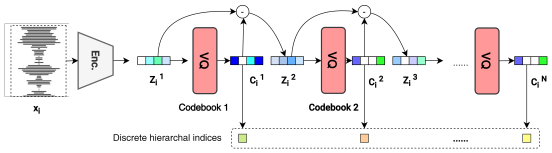


Figure 2.5: Residual Vector quantization (RVQ) process : a single frame is processed into N codebook entries

Two main approaches to denoising autoencoders (Masked autoencoders) have taken hold in recent approaches : leveraging recent advances in Audio Spectrogram Transformers (AST [60]), or using tokenized audio representations. Multiple studies use AST to perform masked image modeling as a pretext task. SSAST [73] and [74] use similar architectures with a masked autoencoder BERT-like setup to reconstruct audio spectrograms, with MSP-SSAP usign a dual masked modeling / contrastive objective, where one patch is contrasted with other patches in the image. Audio-MAE and MAE-AST [75], [76] only encodes unmasked patches for memory savings and regularization. [76] combines this with a contrastive loss on decoded embeddings only. Both these methods show that encoding information from masked patches is costly in both performance and memory. Multi-Window-MAE [77] and Multi-scale AST (MAST) [78] explore using features from different scales to aggregate information from multiple views. ASiT [79] uses a student-teacher setup in which the students' representations only serve as embeddings for a reconstruction head and only evaluate the loss on masked tokens. BEATs [80] uses vector quantization on spectrogram patches to conduct Masked modeling in a discrete space. Recently, A-JEPA has adapted an idea from the image domain [81] in learning to predict a target chunk of an audio spectrogram encoded by a target encoder based from an encoded representation of a disjoint context, using vision transformers (effectively AST) as encoders and decoders with a curriculum masking strategy.

Methods using ASM and TASM have become a second branch of Denoising Autoencoders that has gained popularity. Early works include Mockingjay [82], in which masked encoded frames are directly reconstructed with an L1 loss. HuBERT [20] is another successful approach, where acoustic tokens prediscovered using clustering techniques on extracted features are used as ground truth labels for a masked modeling task. HuBERT is later distilled with discrete tokenization in CoBERT [19]. W2V-BERT [83], DiscreteBERT [84] and VQ-Wav2Vec [17] predict discretized audio tokens in a masked modeling task. *Very* recent developments have shown that Neural Audio Codecs (see above) are also able to extract meaningful representations for dowstream tasks.

2.2.4 Multitask SSL for general audio

Multitask learning methods have also progressively evolved for SSL for audio, wherein the optimization criterion is decided by two or more tasks instead of one. Usually, these tasks are a combination of masked autoencoding and contrastive learning. Audio-MAE mentions that adding a contrastive objective to masked auto-encoding is actually detrimental to model performance, but other studies report otherwise. MAE-AST [76] and SSAST [73] use two heads on the decoder outputs, where the anchor is the decoded embedding and the positive and negative samples are encoded emebeddings. ASiT [79] uses contrastive learning on sequence and class tokens from the teacher and student network on top of a reconstruction loss from the student network. HuBERT [20] combines its reconstruction objective with a contrastive loss between logits and clusters. AudioFormer [72] introduces multi-instance contrastive loss on top of MAM pretraining for RVQ sequences and shows improvement upon pure MAM. multimodal, multitask approaches include COALA [85], [86] and [21]. CLSR [87] uses multiple contrastive losses, latent MSE reconstructive loss, and a semantic alignment loss to co-align modalities.

In this section, we have covered recent and key developments in self-supervised methods for audio representation learning. Table 2.1 summarizes the contributions to self-supervised learning for general audio that are foundational to advances in SSL for music.

2.3 SSL for musical audio

Machine learning studies for musical audio have historically lagged slightly behind applications to general audio. Multiple factors can be attributed to this : The level of academic interest in the various subfields, the lack of annotated *or* non-annotated datasets to support the development of DL models, but more specifically, the highly specific nature of musical audio. Though a subclass of general audio,

Model	*SSL method*	*Loss*
CPC [63]	Contrastive	NT-Xent
wav2vec [65]	Contrastive	NT-Xent
COLA [56]	Contrastive	NT-Xent
CLAR [57]	Contrastive	NT-Xent
CL-SER [58]	Contrastive	NT-Xent
W2V-BERT [83]	Masked modeling, Contrastive	Cross-Entropy, NT-Xent
HuBERT [20]	Masked modeling	Cross-entropy
BYOL-A [48]	Siamese Network	BYOL loss
VQ-Wav2Vec [17]	MM **THEN** Cont.	NTX **then** CE
ASiT [79]	Contrastive	NT-Xent
SSAST [73]	MM, Cont.	MSE, NT-Xent
Audio Barlow [49]	Siamese Network	Cross-Correlation
AudioMAE [75]	Contrastive	NT-Xent
MAST [78]	Contrastive	NT-Xent

Table 2.1: Summary of key SSL methods for general audio

music can almost be considered a different medium to general audio, with very specific structure, challenges, and rules - some dictated by centuries of cultural development and music history, some because of the harmonic and time-structured nature itself of music. Furthermore, music is a medium with a high number of linked modalities - which makes learning comprehensive and interpretable representations of the musical medium a vast multimodal challenge. This discrepancy has been observed in the development of supervised methods for music, and SSL for musical audio has not been an exception. Indeed, while SSL methods for audio have gradually emerged, they have done so in the wake of comparable methods for general audio.

While general audio representation learning generalizes well to music in some cases, a recent study has shown that the applicability of pre-trained speech models on musical tasks is mostly hit-and-miss [88]. However, newer studies departing from already explored approaches in general audio applied to musical audio have begun to emerge. These studies show a trend of beneficially applying domain and structural knowledge of musical audio - supported by growing datasets and additional modalities - to self-supervised tasks with great success. This growing number of successful music-specific studies is relatively recent.

2.3.1 Siamese Networks for Musical Audio

The applications of siamese networks for the purpose of similarity maximization as in their original implementation is rather limited in the literature for musical audio. Some examples include content similarity for cold start recommendation [89], plagiarism detection [90], and audio-to-score alignment [91]. However, a recent trend in music representation learning, equivariant self-supervision, has used Siamese networks to learn highly interpretable representations, as musical information is explicitly encoded into the learned latent space. A network g is equivariant to a transformation q if $g(q(x)) = q(g(x))$. Contrastive learning relies on invariance to learn representations, by training the network to be invariant to these transformations to learn representations that are robust towards these augmentations. Equivariance learning, on the other hand, hard-bakes a representation of the transformation into the latent space. In equivariance learning, representation collapse is prevented by design.

In music, a first iteration of equivariance learning was SPICE [92], designed for unsupervised pitch estimation, in which the relative difference between two pitches is predicted. PESTO [93], its successor, uses a simple equivariance constraint by designing an affine transformation where the latent space of pitch classes is transposed when the input is also transposed. An equivariant loss then learns to predict the parameters of said affine transformation , which prevents representation collapse. Equivariant Self-Supervision for Tempo Estimation [94] leverages the same idea and uses time-stretching, a transformation which tempo is equivariant to.

While equivariance in and of itself is a self-supervised approach to so far already quite well explored issues, its main contribution is that by design, the latent spaces contain musical information. Indeed, because of the equivariance to a given transformation, applying the transformation within the latent space is equivalent to applying it in the audio domain, opening the way to highly explainable latent space navigation with multiple musical attributes being explicitly encoded into said space. Equivariance is a promising approach to learn interpretable latent representations, the challenge lies within finding equivariant transformations for nontrivial musical attributes, making this a compelling field to explore.

2.3.2 Contrastive learning for Musical Audio

Contrastive Learning for Musical Audio [95] was quickly adapted from SimCLR [5] with great success in the task of automatic music tagging, beating out supervised counterparts such as MusicNN [96] and SampleCNN [97] as well as CPC [63] with a relatively small model size and minimal fine-tuning. This success was due to two contributions which outline current research directions for contrastive learning for musical audio.

Firstly, CLMR outlines a new positive sampling method by selecting multiple audio chunks from the same song as positive samples - with no overlap. In doing so, the model learns global representations for various parts of a song, which might not show high structural correlation when looked at separately. This rejoins a trend in exploring contrastive learning of negative sampling : How can we inform our decision of negative and positive sampling strategies to not accidentally push together dissmilar samples or push positive samples further apart? [98] explores this issue for music with different strategies - proximity sampling, random sampling, and self-augmentation, showing that adjacent chunks of the same track function best as positive anchors. A simple strategy which has proven to be quite effective in approaching state-of-the-art on automatic tagging. [99] Approaches this issue of negative sampling from another angle by training a global feature extractor model with masked modeling, and then compares local frames of a given audio to extracted global features of both augmented input samples. In doing this, it selects as positive frames those that have a high correlation to the global representation of the audio. [100] Uses a supervised - unsupervised training setup where the embeddings of a supervised model serve as guidance for positive-negative masking, another approach to the negative sampling problem. [101], a multimodal text-image approach, uses emotional annotations to group samples with similar emotional attributes.

The other contribution from CLMR that has sparked interest in the field is the incorporation of domain knowledge into the training setup. CLMR proposes musically relevant augmentations of audio (such as reverb, pitch-shifting, low and high-passing). This incorporates domain knowledge about the field of musical audio into the model without touching the architecture. [102] use source separation as a pre-processing step and consider source-separated segments of the same audio as positive samples for the mixtures. Even more musical tasks used a preprocessing or augmentation steps could be used to further adapt contrastive learning to musical audio. S3T [103] and TUNe [104] incorporate similar augmentations as CLMR in their training, but with different architectures. AST and Multi-Scale U-Net respectively. Very recently, *Akama et al.* [105] propose using the strong instance discrimination capabilities of contrastive learning as a support task to traditional automatic tagging tasks after an initial contrastive-only pretraining stage, Achieving state of the art results on the MagnaTagATune dataset - showing that including self-supervision into supervised tasks can also serve as a positive regularizer and maintaining discrimative capabilities while specializing to a given task.

2.3.3 Autoencoding for Musical Audio

The high potential of autoregressive predictive coding as a representation learning method for musical audio. Jukebox uses multi-resolution VQ-VAEs to generate long-form music. beyond generation, it was also shown to be a potent representation learner for downstream tasks. CALM learns useful representations for Music Information Retrieval [106] and shows that using a linear probe on JukeBoxes' layer embeddings significantly outperforms self-supervised state of the art and approaches or beats specialized state of the art on multiple tasks. While the sheer size of JukeBox can be attributed as a cause for such performance, [106] shows that APC is a viable learning strategy for music. Further studies have used this idea of probing Jukebox for melody transcription [107], or source separation [108]. Newer generation models [109], [110] using RVQ token generation could also follow the same trend and produce meaningful representations that provide hierarchical information by design.

Denoising Autoencoders have also been used to learn representations for music, but are still a growing field with potential for exploration. Early approaches include predicting domain-knowledge metrics and attributes through autoencoding, as PASE [111] trains multiple classification and regression heads on the outputs from an encoder to predict attributes such as tempo, chroma, waveform... with competitive results on genre and instrument classification, and automatic tagging as a first self-supervised baseline. MERT [112] proposes a multitask approach to reconstruct both acoustic information (RVQ Encodec tokens or HuBERT clustered acoustic units) and musical information (CQT frames) with masked audio modeling. One output head reconstructs the acoustic information, and the other reconstructs a CQT frame, a representation particularly adapted for harmonic tasks. MERT shows competitive or state of the art performance across a range of downstream musical tasks. Finally, MAP-Music2Vec [113] uses a data-2-vec [26] approach with a student-teacher setup, where the student

model receives the masked input and predicts the whole input embeddings of the teacher model. With 20x less parameters, MAPMusic2Vec achieves a comparable performance to jukebox on downstream tasks, training only on music data. MuLaP [114] uses cross-attention between two text and music transformer encoders and separate decoders for unmasking. This weak text supervision leads to improved performance over CLMR [95] and competitive performance with CALM [106] on downstream MIR tasks, with potential for the evaluation of cross-modal tasks in future work.

Recently, FMX [115] trains 9 masked autoencoders on different targets, data, with different architectures with the goal of providing a set of foundation models for music downstream tasks and comparing the influence of architecture, latent sequence sampling rate, data needs and tokenized targets for masked modeling for music, providing valuable insights for future work approaching such models. A few takeaways from [115] are that models trained on shorter context lengths excel on "local tasks", while longer context lengths perform better on more global tasks such as automatic tagging. Local tasks (*"token-level classification"* in this paper [115]) are more discriminative in the performance of foundation models and show that local tasks demanding a long context such as downbeat tracking and structure analysis suffer when evaluated on granular models trained on shorter context length.

In this section, we have covered approaches that have been closer in nature to applications for general audio, which have still seen success on self-supervision for musical downstream tasks, and approaches that have adapted SSL to music-specific tasks. Table 2.2 summarizes these key approaches.

Model	***SSL Method***	***Loss***	***Task***
CLMR [95]	Contrastive	NT-Xent	
CALM [106]	APC	Reconstruction, Codebook	
S3T [103]	Contrastive	NT-Xent	
TUNe [104]	Contrastive	NT-Xent	
EquiTempo [94]	Equivariance	Equivariance	
MAP-M2V [113]	Masked Modeling	MSE	
PESTO [93]	Equivariance	Equivariance	
MERT [112]	Masked Modeling	MSE	
FMX [115]	Masked Modeling	Multiple	

Table 2.2: Summary of key SSL methods for music

2.3.4 Shortcomings and outlook for SSL for music

As mentioned previously, SSL for musical audio is a developing field, thus the smaller number of studies conducted so far. Certainly, more work in the near future will emerge, but SSL for music still faces some hurdles in smooth advancement of research:

Firstly, access to large-scale datasets the scale of which matches those for text, image, or even general audio and speech, is currently very restrained. Large-scale datasets for musical audio include the Free Music Archive **fma˙dataset** and MagnaTagATune [116] dataset, as well as the MTG-Jamendo dataset [117]. The Million Song Dataset [118], which contains one million 30s samples of musical audio, is now publicly unavailable, and many of the previously cited studies use proprietary or private datasets.

Furthermore, this field is missing a set of unified evaluation benchmarks to properly compare models. Recently, the Music Audio Representation Benchmark for universaL Evaluation (MARBLE [119]), and even more recently the `mir_ref` [120] framework were released which are a first step towards reliably comparing these models, with baselines MusiCNN [96], CLMR [95], CALM [106], Music2Vec [113], MERT [112], and MULE [100]. Though MARBLE proposes a healthy assortment of tasks and baselines to evaluate, multimodal approaches were not covered, and more extensive tasks could be proposed to further enable reproducible and verifiable research for music representation learning.

However, this area of research is rife with new research directions to explore. Not only do new developments in the field of audio representation learning quasi-systematically open new alleys using these developments in the field of music representation learning, SSL for music also has directions of its own: using the hierarchical and multi-scale nature of music that results from its highly structural nature is yet under-explored and has shown promising results in representation learning. Another research direction of notable interest is including domain knowledge of musical audio directly into SSL either by baking it into the architecture or by leveraging known characteristics of music to design tasks and augmentation pipelines.

2.4 Multimodal approaches

Recent years have seen a development of interest in integrating multiple modalities into self-supervised learning tasks for audio and music with the development of newer, more voluminous multimodal datasets [38]–[44], [121], [122]. Main complementary modalities for audio include images (video, by nature, aligns corresponding audio and images well in time) and text. Indeed, additional modalities provide additional supervision signals and tasks to solve for self-supervised methods, and these joint representations hold potential for tasks leveraging an understanding of both modalities such as captioning, retrieval, conditional generation... The general framework of multimodal self-supervised representation learning leverages encoded representations of data from different modalities to solve a pretext task. This task can pertain to one or more of these modalities, thus creating a shared latent space containing robust representations for both modalities. The learned representations can then, much like unimodal approaches, be used for downstream tasks with the added possibility of serving as a support representation for a task concerning multiple modalities (See figure 2.6).

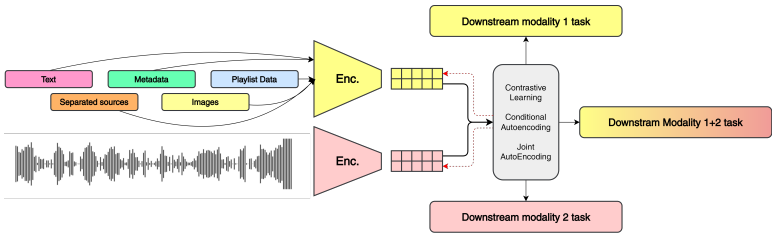


Figure 2.6: General multimodal weak supervision framework

Though not entirely unsupervised, as there is the need for correspondence between pairs of input data, it is much easier to find noisy supervision data from cross-modal pairs than to find robust, clean labeled datasets for supervised tasks, specifically for audio and even moreso for musical audio.

2.4.1 Multimodal weak supervision for audio and Music

Multimodal contrastive learning for audio and music

Multimodal approaches with contrastive learning have garnered increasing attention. By nature, multimodal datasets lend themselves well to contrastive learning as different modalities of a same sample are different views. Studies on cross-modal approaches using image and audio have largely leveraged video datasets to learn joint time-aligned representations. A first attempt at image-audio learning in a "contrastive" fashion is L3-Net [123] - which, while it isn't exactly contrastive, inspires further works down the line in co-aligning modalities [124]–[128]. A popular method has been the distillation of previous knowledge in image-language pre-training supported by much larger datasets [129]. CLIP [130] has been specifically leveraged to learn robust representations of image and audio [131], [132].

Building text-audio multimodal latent spaces is also of high interest for tasks such as text-to-speech, speech-to-text, captioning, understanding and reasoning, retrieval... One foundational approach using contrastive learning is CLAP [133], which contrasts matching pairs across modalities similarly to CLIP [130]. Beyond providing state-of-the-art zero-shot capabilities using prompting approaches, CLAP has since been widely used as a baseline text-audio encoder pair for generation [134]–[139], retrieval [140], [141], and even source separation [142] . Similar approaches have since been undertaken using more advanced encoders [143], [144] and caption-augmentation to enhance existing text-audio datasets [38], [145]. Other approaches include leveraging tags [85], [146] and speech transcriptions with ASM for speech as a supervision signal [147].

Recently, approaches attempting to unify more than two modalities using contrastive learning have begun to emerge [126], [128], [131], showing the desire of the research community to build unified representations across modalities which can be used for all manners of downstream tasks. It stands to reason that multimodal contrastive learning is a field of growing interest, as disentangled well-organized latent spaces directly correlate to out-of-domain transferability, and more modalities correlate with more means of navigating these spaces - as well as more supervision signals.

Contrastive learning has also been used for text and music, with notable examples including using metadata such as tags as a supervision signal [85], [148], [149], lyrics for lyrics alignment [150] or as a supervision signal for a captioning model [151], or descriptive noisy text labels [152]–[155]. MUSCALL

[152] specifically shows promise in zero-shot capabilities of text-audio contrastive models on music automatic tagging and genre classification.

Multimodal Autoencoding for audio and music

Some Noteworthy approaches in multimodal learning include applications to audio captioning [86], music understanding and reasoning [156], and multimodal retrieval. One interesting remark is that due to the weak supervision signal from extra modalities, Denoising Autoencoders have been less popular with multimodal learning. Instead, the most common approach in text and audio is to extract a sequence of embeddings from both audio and text and using the audio sequence as a prefix for the text sequence as input for a language model. Recent results include Listen, Think, and Understand [44], which provides state-of-the art audio reasoning and zero-shot retrieval abilities. [157] concatenates representations from pretrained tokenizers and learns to decode for both text and audio decoders. VALL-E shows that Neural Codecs are Zero-shot text to speech synthesizers by concatenating text and audio tokens and training a decoder to generate the corresponding acoustic codec tokens [158]. Alternatives include cross-attention between both sequences [86], [159]. CLSR [87] proposes that reconstruction should not only be evaluated intra-modality, but inter-modality also, and shows that learning to reconstruct intermediary latents of the other modality from a bottleneck leads to better inter-modality entanglement. Interestingly, as far as we have found in the literature, less Audio-Visual approaches exist that use denoising autoencoders : [160] experiments with concatenation, cross-attention, and separated reconstruction objectives for video, obtaining better results for mid-transformer concatenation. [21] uses a double reconstruction-contrastive objective with a ViT and AST to learn audiovisual representations. multimodal, multitask approaches include COALA [85], where both reconstruction loss and autoencoding to maximize co-alignment between tags and audio, [86] and [21], where embeddings from different modalities are contrasted as different views of the same sample as an additional task to masked modeling. CLSR [87] uses multiple contrastive losses, latent MSE reconstructive loss, and a semantic alignment loss to explicitly symmetrize the alignment between modalities.

Similar approaches to those in the audio domain for multimodal APC have seen light in the music domain, with mostly the same applications : Music captioning, Music understanding and reasoning, and Codified music generation conditioned on text. Music captioning approaches leveraging APC and thus presumably learning multimodal representations include previously mentioned LLARK [156], and other approaches using cross-attention with a music and text encoder to generate captions[151], [161]–[164]. The most notable approach for TASM music generation conditioned on text is MusicLM [165], which uses MuLan [153] and W2V-BERT [83] tokens as prefixes to autoregressively generate soundstream [69] RVQ tokens. The very recent Music-Understanding LLAMA [166] uses representations from MERT [112] fused into a frozen LLAMA model to generate answers to a newly generated Music QA dataset.

Multimodal approaches with masked modelling for music are mostly centered towards joint masked modelling using both a text encoder and audio encoder. MuLaP [114] uses cross-attention between transformer encoders and separate decoders for unmasking. This weak text supervision leads to improved performance over CLMR [95] and competitive performance with CALM [106] on downstream MIR tasks, with potential for cross-modal task evaluation in future work.

2.5 Human-in-the loop Machine Learning

2.5.1 A general overview of Human-in-the-loop

Human in the loop ML is a whole field of research on its own, and covering it exhaustively would take us outside of the scope of the valuable research for this stage 0 proposal. The general approach of human-in-the loop machine learning is to reintroduce human annotators in the training process to improve the performance of models on downstream tasks. Here, we choose to restrict our scope to active learning, which has been leveraged in the field of MIR in the past, curriculum learning, which holds promising applications for music, and Reinforcement Learning from Human Feedback (RLHF).

Active Learning

Under a first umbrella of Human-in-the-loop approaches can be found Active learning, Interactive Machine Learning, and Machine Teaching [167]. All three are similar in that the learner is always a machine learning model and the process of teaching this machine is incremental and aims to refine

unclear or difficult examples through the iterative mechanism. All three involve some form of an expert annotating data after an initial training phase and the model adapting to the new training data stemming from identified weak points. However, the manner in which this is done differs from approach to approach.

In Active Learning (AL), the 'human' is considered an all-knowing oracle used to annotate unclear data points for the learner, and the model is in control of the learning process [167]. The data is separated into unlabeled and labeled sets, and the model can **query** a label from the oracle for an unlabeled data point. The sampling process from which the model samples the queried labels can be *random* or one of two: *information based*, where the sampling criterion might be prediction uncertainty, disagreement between a set of trained models (Query by commitee), margin with regards to the decision boundary of a linear SVM classifier, or in the expected change of model or prediction. *representation-based*, where the goal is to sample the query based on what is "lacking" in the representation of real world data : the input can be selected based on density regions in the input space, rare class occurences for more robust classification, or on centroids of representative clusters within the representation space [167], [168].

With AL, an iterative process improves the performance of the models until a stopping condition is met. In this process, the model is trained on the labeled dataset. It then queries the oracle based on the unlabeled sampling strategies described above for an annotation on an unlabeled data point, which it then uses to retrain itself, either sequentially (after each data point) or in a batch fashion (after multiple annotations) [167] - See Figure 2.7.

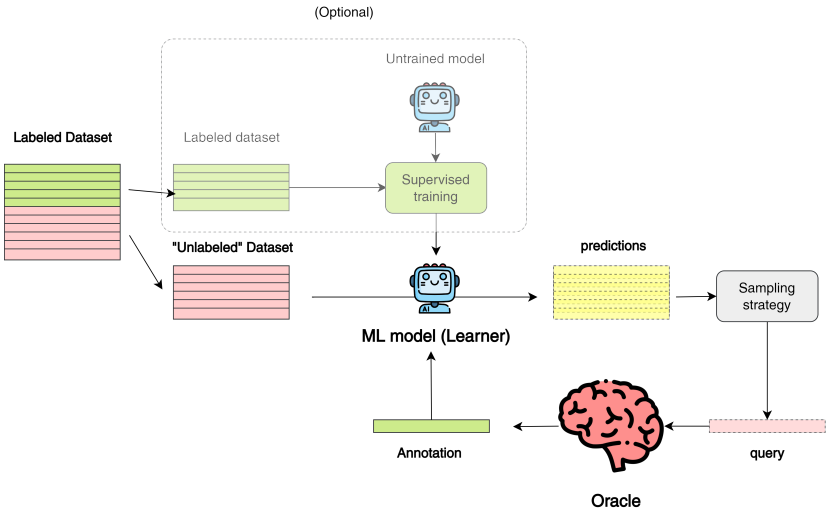


Figure 2.7: Principle of Active Learning - an iterative process allows a model to select unclear or underrepresented examples and reinforce its prediction capabilities on these unclear instances by querying an omniscient oracle.

AL has been successful in multiple domains, including Computer Vision for image classification and segmentation [169]–[173] and NLP for text classification [174], [175] Named Entity Recognition [176], and sentiment analysis [177], but has certain weaknesses: Noisy and imperfect oracles lead to degradation of the oracle annotations over time, especially in cases where the annotation is difficult or subjective, even for humans. Novel classes yet to be seen in the training or oracle datasets are difficult to incorporate once the model is interacting with real-world instances. Furthermore, when multiple tasks are involved for one same model, the issue arises of how to prevent catastrophic forgetting in other tasks due to oracle-induced improvements on other tasks.

Sister approaches to Active Learning include Interactive Machine Learning (**IML**) and Machine Teaching. **IML** is closest to **AL** and differs slightly in that the sampling strategy is interactively co-decided between the model and an expert in the loop with knowledge of what the model should be learning. In this, it includes more Human-Computer Interaction considerations and a broader consideration of *who* the human in the loop is.

Machine Teaching is different in spirit, but has recently also converged to an iterative process. In MT, the teacher is in control of the learning process and has the objective of teaching the learner

a target ML model. The teacher collects the sampling set, designs the learning process, and labels the required data for a given step of the teaching process. Because MT is highly specific in terms of relevant tasks and the need for an expert in the loop, which often does not have perfect knowledge, we elect not to cover it in depth here.

Curriculum Learning

In curriculum learning, a sequence of increasingly complex tasks or instances are given to a model, with the intuition that foundational knowledge is acquired in the earlier and simpler stages of learning, and more complex examples help the learner to refine its understanding of a given topic. Curriculum learning improves model performance and improves the efficiency of training compared to a traditional supervised learning approach by regularizing the training towards better regions of the parameter space, and denoising by focusing on high-confidence areas of the target space to gradually increase the noisiness in the data throughout the training. Such approaches have proven to be useful in.

However, the success of CL is highly dependent on the capacity to organize the "learning material" into a relevant curriculum of increasing complexity, and thus requires a measure of difficulty as well as a coherent scheduling of the increase in difficulty to be applicable. These both can be either human-determined (**predefined CL**) or learned by another ML model (**automatic CL**). With predefined CL, the specific successful combination of difficulty measure and scheduler is often hard to come across for a given task and is predefined; thus, it ignores the feedback provided by the model during training. It is also difficult to determine a difficulty measure in nontrivial cases, requiring domain experts. Automatic CL has emerged as a way to circumvent these limitations. Three main approaches exist for automatic CL: **Self-paced CL**, where the loss of the model on a given set of examples is an indicator of the perceived difficulty of these examples - and thus a difficulty measure. **Transfer teacher CL** uses a pretrained teacher model as the difficulty estimator by measuring its performance on the data points and using the set with the best results as the easier data. Finally, **Reinforcment Learning Teacher** methods dynamically adjust the provided data based on the evolution of the performance of the student model on the selected data - The teacher in this case is a RL model that selects data points for the student to learn [167].

Difficulty estimators can either estimate the **complexity** of a given data point pertaining to the intricacy of its structure, or its **diversity** [167], pertaining to the amount of different types of examples within a dataset - the noisier and more diverse the data, the more proficient in its discriminative capabilities the model has to be to be able to discern relevant information within the training examples. On the other hand, training schedulers can adjust the complexity threshold after a given number of training steps (**discrete scheduling**), or link the training iteration with a scheduling function of difficult sample proportion to include in the training (**continuous**).

Reinforcement learning from human feedback

Finish this before next week

2.5.2 Human-in-the-loop ML for Audio and Musical Audio

Human in the loop approaches are relatively rare for audio and music despite the importance of such approaches due to the perceptual and cultural complexity of these modalities. Curriculum learning and RLHF have notably few studies pertaining to audio and music. This is understandable: devising an appropriate curriculum of increasing complexity is a difficult task for music and audio, as the issue arises of how to measure the "complexity" of a musical piece, specifically as a concept that varies in relation to a specific task - a fast-paced challenging solo violin piece is complex from a transcription standpoint, but trivial from a source separation standpoint. Musical RLHF is also hard to implement as training the reward policy requires a large number of human annotations, which should ideally come from experts and are thus time-consuming and expensive to source. With the development of conversational multimodal LLMs for music [156], [166], RLHF is no doubt in development for these models, inspired by the success of the approach for conversational LLMs such as ChatGPT. Currently, however, RLHF approaches for multimodal LLMs **or** for other musical tasks are rare - even, to the best of my knowledge, non-existent beyond using RLHF-trained LLMs and fine-tuning them on multimodal music pseudo-captioning [161] without RLHF on the downstream music understanding task. So, current HIL approaches for audio and music are mostly restrained to active learning methodologies, which we briefly explore in this section.

Musical annotations are often hard to source as they require expertise, and even when experts are involved these annotations can become contradictory as experts disagree, leading to the problem of the noisy oracle as described in section 2.5.1. Human-in-the-loop is an attractive solution for MIR approaches because it allows for personalized model outputs through human annotations of subjective musical properties, as well as low-data efficiency which is a key component to dealing with the low amount of annotated musical data.

Human-in-the loop approaches for music include SERENADE, an automatic chord estimation model which leverages oracle inputs on one chord to propagate prediction modifications along the whole progression [178]. Beat tracking has been addressed as well, where the oracle modifies the position of a predicted beat, and the whole prediction shifts accordingly [179]. An AL approach of Emotion Recognition on Chinese music uses a SVM model which is updated after oracle annotations, showing that iterative annotations improve performance on classification-style emotion recognition [180]. Other approaches also use a SVM for retrieval [181] and genre/mood classification [182], also showing improvement over a baseline non-AL approach. Optical Music Recognition has also been addressed through AL, with oracle-corrected inputs imposing constraints on the generated transcriptions [183]. Finally, more recently, generation approaches have also benefited from human-in-the-loop, not necessarily through active learning but with iterative human input, either by iterating over a given first generated audio with task selection over a number of generation models [184] or with a set of commands the user can iteratively apply to any given audio [185].

So, A variety of tasks in MIR and generation have been addressed with human-in-the-loop, but only scratching the surface, as to my knowledge this is the extent of HIL approaches for music. This is due to previously mentioned hurdles but is also promising, as there is still space for human-in-the-loop to positively influence the performance and interactivity of ML models in MIR and other music-related AI tasks. Specifcially, Curriculum Learning approaches and RLHF approaches have not at all been leveraged yet, and with the new capabilities of multimodal LLMs and new architectures for music understanding, constitute a promising research direction.

Chapter 3

Problem specification

In the previous section, we have covered recent promising advances in learning music representations. While these approaches have been successful, in an ideal world, learned music representations and representation learners would combine a set of desirable characteristics that would make them **performant, explainable, and interactive**. Figure 3.1 shows a mind map / Venn diagram of the intersections of these properties and the current state of research in different key areas that should all converge towards the intersection of all three attributes to reach the "ideal" learned representations for music.

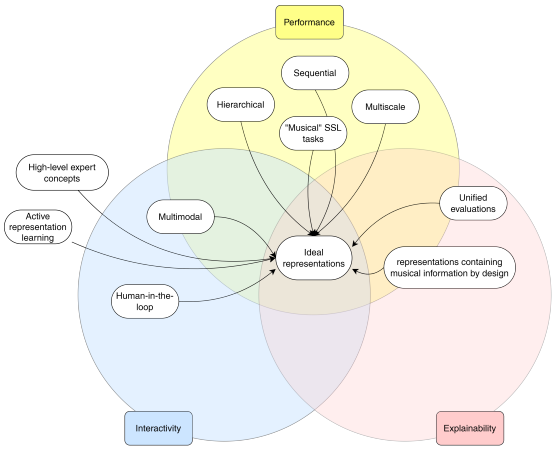


Figure 3.1: Mind map of the current state of research relating the the three aforementioned desirable attributes of learned representations. Ideally, all research directions for music representation learning should converge towards a unified "ideal" representation at the center.

Performance

We use the term **Performance** to describe the capacity of learned representations to provide state-of-the-art results on downstream tasks, with as little labeled data as possible. The current trend - like many adjacent domains in machine learning - is to leverage larger and larger datasets with larger and larger models - *in essentia*, throwing numbers at the problem. Although there have, of course, been incremental steps in performance on downstream tasks, much to the credit of previous research endeavours, the current state of open data for music warrants a more domain-appropriate approach, by infusing domain knowledge in the design process to provide representations that contain meaningful musical information *by design*.

One issue with the current state of research is that **music representation learning is not designed for music**. By designed for music, we mean representation learning tasks that emulate certain properties of music that would be desirable to also find in learned representations. This can be hierarchy, structure, sequentiality, scale... or infusing information about musical attributes into learned representation by design of the representation learning task (timbre, tempo, pitch, etc.). Some efforts are blossoming to include musical information in representation learning tasks [93]–[95], [98], [112],

showing potential for targeted downstream performance increases as well as out-of-domain transferability, but these are mostly isolated efforts, and there is no unified research front towards making deep, design-level changes to existing architectures and tasks for music representation learning.

Explainability

Explainability is another key attribute that learned representations should exhibit, specifically for music. Music is a highly personal, interpersonal, cultural, and emotional medium. Current learned representations are still mostly opaque, with the learning processes being essentially black-boxes. An ideal representation of music would be transparent, with identifiable properties of the learned representations relating to identified proprties of the audio. This attribute feeds into all the other desirable properties for these learned representations. Models would be more likely to identify key components of representations for downstream tasks, making them more performant. Interactivity is enhanced by additional navigation capacities, and musical experts would benefit from being able to dissect learned representations without resorting to *post-hoc* methods. Again, currently, efforts towards explainable learned musical representations are minimal and isolated, despite the importance of such a property.

Interactivity

Interactivity would describe the capacity of these representations to adapt to a human input, either by changing to best suit the users' needs or by providing means of navigation for the user through the learned representations. This is particularly important as music is a personal medium where many interpretations and desired outcomes for machine learning models are highly subjective. Interactivity and adaptability stems from two main approaches in the current research space : human in the loop and multimodal machine learning.

While multimodal approaches have seen a strong push in recent months, with success in improving downstream performance, providing interactive text-music modeling for somewhat conversational understanding of music, and success in retrieval and captioning, these approaches have also trended towards becoming a numbers game, and the resulting models have no music knowledge infused in them past the exposition to a large amount of musical data. Again, this is not a desirable outcome for a medium that requires an intricate understanding of structure, rules, and culture to be able to produce a cogent description for a track - especially beyond the novice level.

Due to the novelty of the field and only recent developments in text-music datasets availability, there are also no unified musically-specific benchmarks, evaluation frameworks or metrics for these text-music and music-text tasks that evaluate beyond each modality individually and into the combination of multiple modalities. This is an obstacle for unified and standardized research in the direction of attaining interactivity and performance through multimodality for music, as standardized and appropriate evaluation can lead to the identification of key shortcomings and provide fair comparisons between approaches.

Furthermore, While text *is* a coherent way of interacting with music for the layman, it requires much more high-level knowledge when interacting with an expert, and other modalities would potentially represent a better interface with machine learning models. Creating interactive representations for experts requires more than throwing data at the problem. Designing machine learning models that adapt their outcomes to the user and the task based on expert feedback are rare, and often limited to the scope of the task. An ideal outcome would be an interactive **representation** that improves upon feedback, not only reinforcing the instance capabilities of the model, but also the structure of the latent space at the same time.

The field of music representation learning is a young one. As this research area matures, it is necessary to specify the tasks and approaches to the medium, which can not simply be transferred without modification from analogous approaches in other domains. To reach the desired properties of **Performance, Interactivity, and Explainability**, it is necessary to deisolate the progress made in the respective enclaved research directions, as progress in one attribute is inevitably tied to the others, and all three must be combined to reach our ideal representation learning framework for music.

The scope of this PhD is not to aim to reach the centre by submission. However, a primary motivation for this PhD is to initiate a more generalized movement towards the intersection of all three attractive qualities by bringing attention to this intersection and incrementally moving individual research areas or groups of areas towards the centre. We turn to the following section to formulate research questions on how exactly this can be achieved.

Chapter 4

Research Questions and Hypotheses

To answer the problem space laid out in section 3, in this section we explore research questions prompted by the observed shortcomings in current approaches for moving beyond self-supervised learning for musical audio. I further formulate hypotheses these research questions will seek to validate or invalidaten, and lay out potential research directions that propose interesting avenues to explore these questions.

Recent work has shown that incorporating musical domain knowledge into the design of model architectures, into the design of self-supervised learning tasks, and into the targets for these self-supervised learning tasks themselves is beneficial to the performance of self-supervised representations on downstream musical tasks. However, these representations, while more performant, have not been shown to exhibit interpretability and explicit navigability, an important aspect for explainability - specifically for experts. Thus, my first research question:

Q1: How can we leverage domain and expert knowledge of music to learn universal, high-performance, yet transparent representations of music?

To answer this question, I will explore building architectures that leverage domain knowledge of music and its properties that would be desirable to be found in learned representations. Some of these properties could be, for instance : **sequentiality, hierarchy, structure, multiple scales, pitch, tempo, timbre...**. The intuition behind this research question is that by creating target tasks and architectures that hold this information by design, there will be no need to perform *post-hoc* analysis of learned representation to correlate desirable interpretable properties and their respective positions in the learned representations. For instance, the recent works exploring equivariance for baked-in pitch and tempo information in the latent space could be exploited to create latent representations where parts of the representation hold this information by design. Or, analogous to RVQ encoding of audio, different tokenized streams could be designed to hold different properties to reconstruct the input signal - including musically relevant ones. Following this intuition, we formulate the following hypothesis:

H1: By incorporating musical knowledge at the design stage of representation learning, we can learn representations more suitable for musical audio by design - in structure, transparency, and performance.

To test this hypothesis, I will design specific architectures and task with musical domain knowledge in mind from the design stage by designing hierarchical, multiscale, and sequential models and tasks with musical domain knowledge in mind, test the performance of the resulting representations on downstream tasks, but also analyze whether the target information designed to be held by these representations is indeed baked into the designed system through novel evaluation paradigms using retrieval and perceptual metrics.

* * *

Multimodal approaches for music have recently gained traction, with recent foundational models stemming from big stakeholders in the music industry showing state of the art performance on music understanding and reasoning for music and text. Approaches with further modalities such as video, image, and metadata, have been slower in their emergence but are no doubt on the way. Beyond these capabilities, which stem in recent studies mostly from larger musical encoders and caption augmentation using large language models, there seems to be little to no exploration of explicitly linking

musical attributes to extra modalities - be it by leveraging learned multimodal representations and extracting meaningful musical information from the representations, or baking this information into the multimodal space at the design step. As described in chapter 3, there also seems to be little work towards conversational retrieval and music understanding with large language models, and the question can be asked if text is even the most intuitive modality for musical experts to interface with machine leaning models. The broader question that can be asked is : **Are we doing multimodal learning for music right?**, leading into the more specific second research question for the proposed project:

Q2: How can weak supervision from additional modalities further disentangle latent spaces and provide intuitive navigation means for users?

As a means to answer this question, there are three main axes leveraging multimodal learning that I would like to explore: **Using the architectures and tasks of Q1** as musical encoders and evaluating whether or not the musical information contained within these representations leads to better performance on downstream multimodal tasks. **Exploring current multimodal approaches** with the goal of finding whether the semantic properties of music are well disentangled within the extra modalities, and if an additional conversational layer to the retrieval systems can specifically help identify and improve the disentanglement of musical properties within the shared latent space. **Improving current approaches for evaluating multimodal approaches for music** is another avenue that I believe is not as developed as it should be for this research community to move smoothly in this area. Evaluation methods for multimodal approaches should not only focus on the performance on one modality but rather on the joint performance of both modalities, and the successful intertwining of semantic (musical) information between both. These observations lead to the following two hypotheses for this research question:

H2: Current multimodal approaches do not result in optimized and interpretable representations for optimal performance when in comes to music, and one approach to solve this is include musical information within the training process

H3: Current evaluation metrics for multimodal approaches for music do not allow for pertinent analysis of the musical semantic knowledge of multimodal models. Improving existing metrics and building new appropriate ones will open up new perspectives and potential areas of improvement for multimodal approaches that are appropriate for music experts.

* * *

The third research question stems from the observation of the general lack of human-in-the-loop approaches for ML in the field of music and audio, despite the value it may bring. Indeed, HIL approaches for music propose a partial solution for two canonical hurdles in the field : The lack of expertly annotated data, and the frequent disagreement between these expert annotations stemming from the fact that music information retrieval tasks possess some degree of subjectivity in their evaluation. Indeed, a large number of musical attributes and properties are, to some extent, subjective in their perception. To these observations, we pose the question:

Q2: Can human feedback improve the performance of yet underexploited MIR tasks? Further, can the inclusion of human feedback improve the representations themselves, providing explainable, intuitive, high-performance, and interactive ML systems for experts?

This notion of human-in-the-loop **representation learning** is rather novel, and only few examples can be found in the literature. This is disconcerting, as a system which would improve not only its predictions, but its inner representations opens up multiple alleys for applications for ML applications personalized through feedback, and ML applications leveraging experts for improved representations. Answering this question will require exploring new paths for HIL interactions between humans and music ML systems, understanding the influence of these interactions on the latent space and the performance of these models, and then confirming the **positive** influence of these interactions, if positivity is clearly defined. Our hypothesis regarding Q3 is as follows:

H4: Involving human experts in the learning process of Machine Learning models applied to music can not only have positive repercussions on the performance of these models, but positive repercussions on the explainability, performance, and navigation of the underlying learned representations if articulated with self-supervised learning.

* * *

One should note that while these questions provide different outlines of potential research directions, they tie well into each other. Indeed, the discovery and exploration of new SSL methods for Music leads to better performance and explainability of Multimodal approaches, and to potential new HIL approaches. Multimodal approaches implemented with music in mind lead to disentangled and high-performance representations, and new means of interactions with humans, leading directly to new HIL approaches. Developing new HIL methods which demonstrably improve performance and representations has direct repercussions on both other research streams, for obvious reasons.

This positive feedback loop between the research directions identified in this section provides, in my opinion, a solid bedrock for the research to be conducted in this PhD. It not only makes incremental improvements susceptible to lead to further improvements in other research areas, but also guarantees the possibility of switching to a less challenging research direction should another become seemingly insurmountable. The wealth of areas of exploration for each of these questions further suggests that there will be no shortage of ideas to explore - a slightly ambitious yet achievable research framework with potential to be narrowed down at a later stage.

Chapter 5

Research timeline and proposed dissemination

5.1 Preliminary work

5.1.1 Dissemination and literature review

Prior to stage 0, I have written a thorough literature review paper that could possibly be targeted towards TISMIR after some polishing. After reviewing previous accepted review papers to TISMIR, this one (A more thorough version of section 2) covers a broader range of topics than most but provides a milestone review for music repressentation learning, a field in dire need of standardization and clear inventory of existing methods. In December, I presented a poster at DMRN18+, which showcases essentially the content of this stage 0, and is shown in appendex **??**.

5.1.2 Work to date : Methods and preliminary results

In these early months of my PhD, I have started working on two streams of research based on identified possible extensions and combinations of work done in literature, and directly linked to the axes of research outlined in section 3. Each of these works contains one simpler but achievable idea that can lead to publication, and one more complex and advanced idea that can lead to a separate publication, making them appropriate, separable pursuits for stage 1 and beyond if necessary. Further, both have identifiable themes and domains of interest that make them suitable publications for respective publication tracks.

MaCHUP proposes a novel masked-modeling / contrastive dual training setup for representation learning for music with further potential for music-specific setups down the line, and is detailed in Appendix A. **SemiSupCon** introduces a novel semi-supervised contrastive task without any complex machinery with high potential for both explainable learned representations and the first active *representation learning* model to my knowledge. SemiSupCon is detailed Appendix B

Overall, these two research streams provide a target achievable publication objective for stage 1 and, if successful, additional objectives suitable for additional publications beyond that milestone.

5.2 Before stage 1

Prior to stage 1 I will be focusing my efforts on the current two streams of research on which I have started working. As mentioned previously, these two projects have multiple potential publications contingent on the success of the envisioned method. I will focus on collecting publishable results for the "simpler" objectives, with the goal of publishing to **ISMIR, ICASSP, ICLR** or **NeurIPS**, conferences that all have submission deadlines within a realistic timeframe for the progress of these projects. If these first approaches are successful, I will consider other projects described in section 4, with the most interesting ones being exploration of equivariance approaches, perceptual metrics for contrastive learning, and conversational music understanding LLMs with multiple attribute-specific representations. I would also like to submit a tutorial to **ISMIR**, and I believe I have some ideas (Christos Plachouras, Yinghao Ma) as to people I could collaborate to do so - although the specific topic has yet to be decided. Figure 5.1 shows a provisional GANTT for the lead-up to stage 1.

MaCHUP will be prioritized for **ICASSP** and **ICLR**, with a possible submission to **ISMIR**. **Semi-SupCon** will be aimed at **ISMIR** in first priority, and **NeurIPS** in second *lieu*. If the human-in-the-loop extension of **SemiSupCon** is successful, it will make an impactful publication for **ISMIR** as well as **ICLR**.

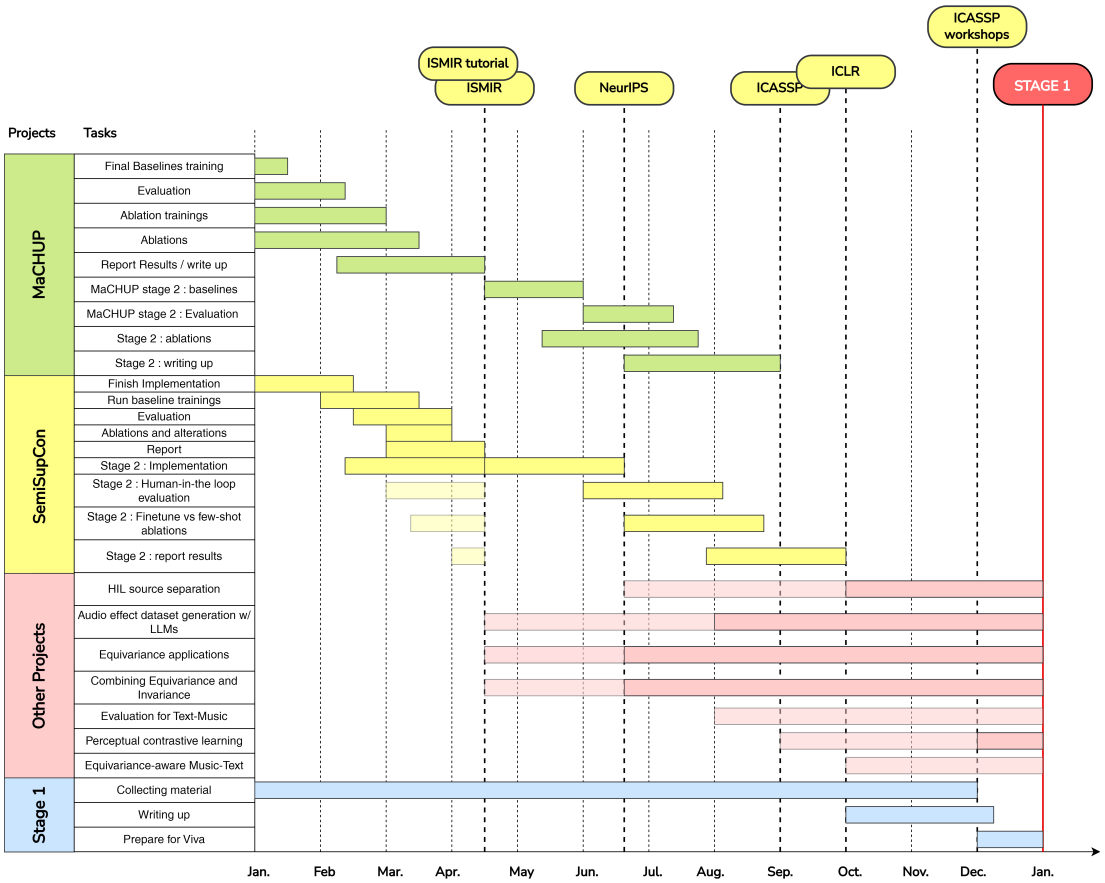


Figure 5.1: Provisional GANTT of the game plan immediately after stage 0 - low opacity bars denote optimistic, slighlty unrealistic timelines. The *Other Projects* section is coherently still vague, as I am currently more focused on MaCHUP and SemiSupCon. These other projects will extend beyond stage 1 into the rest of the PhD.

MaCHUP : So far, baselines have been established that successfully demonstrate the ability of combining masked modeling and contrastive learning tasks to beat either on their own on a limited set of tasks. If confirmed by clean evaluation with standardized training conditions, this will provide publishable results. What will then remain is conducting clean, comparable evaluations to state of the art models to compare results fairly, *i.e* on comparable data with sufficient training time and on multiple tasks to showcase the versatility of learned representations. Furthermore, this project should have multiple ablation studies due to the complexity and highly hyperparameter-dependent training to ensure that the claims of improved performance are substantiated and not just the result of stochasticity.

SemiSupCon : If we can show that including limited amounts of labeled data in the training stage can lead to improved performance on downstream tasks in-domain and out-of-domain for both architectures, this will constitute publishable results.

Further steps will include analyzing the influence of the amount of labeled vs unlabeled data in the training stage, at the batch and dataset level. Ablation studies comparing fully self-supervised to fully supervised to semi-supervised will also have to be conducted. Comparison to state-of-the-art models is a potential approach but a more interesting one would be to gauge the performance gain of different types of labeled data on different tasks (*i.e., can pitch classification labeled data help with chord identification? Can instrument classification help with source separation?*. The intuition is that by including marginal amounts of labeled data corresponding to downstream tasks, we should be able to achieve competitive performance to state of the art at lower computational and data costs. These ablation studies are numerous, but I am seeking cooperation with **Christos Plachouras**, whose research interests towards low-data regime approaches align well with this project.

Once these ablation studies have been conducted, depending on the coherence of results, this will

constitute a solid publication for the conferences mentioned above.

5.3 Long term : After stage 1 and until submission

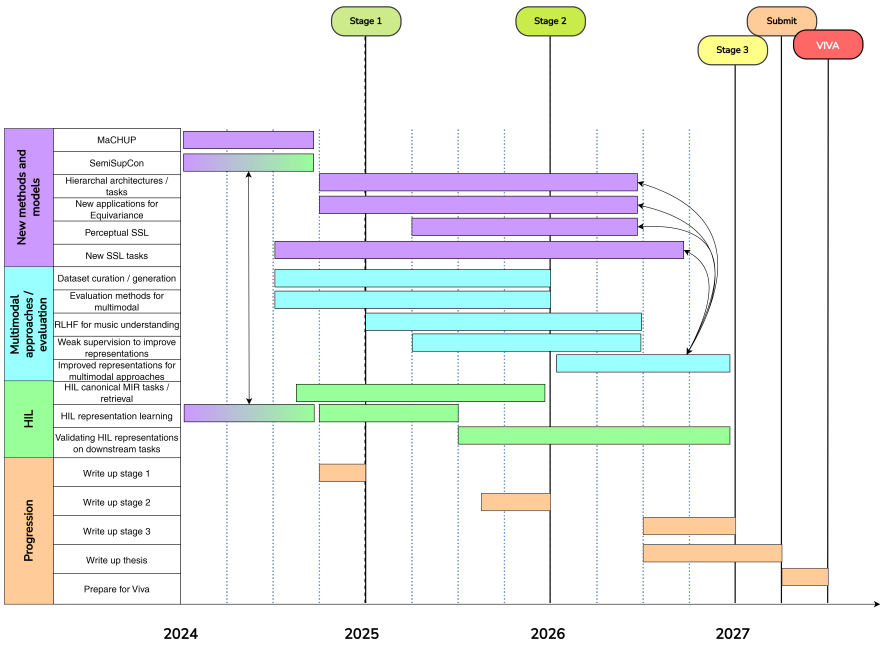


Figure 5.2: Provisional long-term research plan GANTT

The two current ongoing streams of research will, in my estimation, last approximately until shortly after stage 1. Both hold interesting research directions and the focus will be largely on the multiscale capabilities of **MaCHUP**, evaluating the usefulness of local representations for local tasks, and attempting to leverage masking and efficient attention mechanisms to fit a whole song of context into learned representation. On the **SemiSupCon** front, the main second research direction will be the human-in-the-loop approach, which if successful will be the first human-in-the-loop representation learning system and will certainly open further avenues for exploration.

In the lead up to stage 1, defining the priority of research projects and ideas past the two current ongoing ones will be a crucial point in determining a specific narrative through the choice of specific research projects to tackle, which will determine the specific game plan following stage 1. Figure 5.2 shows some broad priorities and timelines.

I expect that the focus up to stage 2 will be mostly on research questions **Q1** (Multiscale, sequential, hierarchical architectures, Equivariance, Perceptual Metrics and other new objectives for SSL...) and **Q3** (Developing human-in-the-loop approaches for existing MIR tasks, and developing human-in-the-loop representation learning methods), mostly because a large part of the methods to be applied in answering **Q2** will be contingent on the success of developing new architectures, tasks, and interactive systems for improving representations. However, there are multiple steps related to **Q2** that can be explored prior. Improving evaluation methods for multimodal approaches, leveraging existing multimodal approaches to improve learned representations of music, data collection and building new datasets for multimodal approaches, and human-in-the-loop music understanding with text, are all avenues that can be explored prior to the global objective of **Q2**, *i.e* improving multimodal approaches with our improved representations.

With regard to the final stages of the PhD, I expect all three streams of research will be loosely intertwined up about three to six months after stage 2, with some contributions mixing research questions and themes, and some standalone. At this point, I hope to have a body of work with enough progress in all research directions to warrant new projects tying the research together, to confirm that the positive feedback envisioned in Section 4 has indeed been achieved, and to demonstrate the positive synergy of the methods from each research direction.

Bibliography

[1] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, Jun. 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[2] 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.

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

[4] J. Zbontar, L. Jing, I. Misra, *et al.*, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, PMLR, 2021, pp. 12 310–12 320.

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

[6] T. Chen, S. Kornblith, K. Swersky, *et al.*, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[7] X. Chen, H. Fan, R. Girshick, *et al.*, *Improved Baselines with Momentum Contrastive Learning*, Mar. 2020. DOI: 10.48550/arXiv.2003.04297.

[8] K. He, H. Fan, Y. Wu, *et al.*, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[9] X. Chen, S. Xie, and K. He, *An Empirical Study of Training Self-Supervised Vision Transformers*, Aug. 2021. DOI: 10.48550/arXiv.2104.02057.

[10] R. Qian, T. Meng, B. Gong, *et al.*, "Spatiotemporal contrastive video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6964–6974.

[11] H. Fang, S. Wang, M. Zhou, *et al.*, "Cert: Contrastive self-supervised learning for language understanding," *arXiv preprint arXiv:2005.12766*, 2020.

[12] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. [Online]. Available: https://aclanthology.org/2020.acl-main.703.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, https://arxiv.org/abs/1312.6114v10, Dec. 2013.

[15] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[16] P. Dhariwal, H. Jun, C. Payne, *et al.*, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[17] A. Baevski, S. Schneider, and M. Auli, *Vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations*, Feb. 2020. DOI: 10.48550/arXiv.1910.05453.

[18] H. Bao, L. Dong, S. Piao, *et al.*, *BEiT: BERT Pre-Training of Image Transformers*, https://arxiv.org/abs/2106.08254v2, Jun. 2021.

[19] C. Meng, J. Ao, T. Ko, *et al.*, "CoBERT: Self-Supervised Speech Representation Learning Through Code Representation Learning," in *Proc. INTERSPEECH 2023*, 2023, pp. 2978–2982. DOI: 10.21437/Interspeech.2023-1390.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, *et al.*, "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.

[21] Y. Gong, A. Rouditchenko, A. H. Liu, *et al.*, "Contrastive audio-visual masked autoencoder," in *The Eleventh International Conference on Learning Representations*, arXiv, Apr. 2023. DOI: 10.48550/arXiv.2210.07839. [Online]. Available: https://openreview.net/forum?id=QPtMRyk5rb.

[22] K. He, X. Chen, S. Xie, *et al.*, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[23] Z. Xie, Z. Zhang, Y. Cao, *et al.*, "SimMIM: A Simple Framework for Masked Image Modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[24] S. Woo, S. Debnath, R. Hu, *et al.*, "ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[26] A. Baevski, W.-N. Hsu, Q. Xu, *et al.*, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, PMLR, 2022, pp. 1298–1312.

[27] A. Baevski, Y. Zhou, A. Mohamed, *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, no. arXiv:2006.11477, pp. 12 449–12 460, Oct. 2020.

[28] T. Mikolov, K. Chen, G. Corrado, *et al.*, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[31] C. Wei, K. Mangalam, P.-Y. Huang, *et al.*, "Diffusion models as masked autoencoder," in *ICCV*, 2023.

[32] A. Makhzani, J. Shlens, N. Jaitly, *et al.*, "Adversarial autoencoders," in *International Conference on Learning Representations*, 2016. [Online]. Available: http://arxiv.org/abs/1511.05644.

[33] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto, "Adversarial latent autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 104–14 113.

[34] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[36] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-Shot Text-to-Image Generation," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8821–8831.

[37] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[38] Y. Wu, K. Chen, T. Zhang, *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[39] J. Huh, J. Chalk, E. Kazakos, *et al.*, "EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound," in *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, arXiv, Feb. 2023. DOI: 10.48550/arXiv.2302.00646.

[40] L. Sun, X. Xu, M. Wu, *et al.*, *A Large-scale Dataset for Audio-Language Representation Learning*, Sep. 2023. DOI: 10.48550/arXiv.2309.11500.

[41] H. Chen, W. Xie, A. Vedaldi, *et al.*, *VGGSound: A Large-scale Audio-Visual Dataset*, IEEE, Sep. 2020. DOI: 10.48550/arXiv.2004.14368.

[42] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[43] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, arXiv, Oct. 2020, pp. 736–740. DOI: 10.1109/ICASSP40776.2020.9052990.

[44] Y. Gong, H. Luo, A. H. Liu, *et al.*, *Listen, Think, and Understand*, May 2023. DOI: 10.48550/arXiv.2305.10790.

[45] J. Turian, J. Shier, H. R. Khan, *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*, PMLR, 2022, pp. 125–145.

[46] 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. DOI: 10.1109/ICASSP43922.2022.9746790.

[47] P. Manocha, R. Badlani, A. Kumar, *et al.*, "Content-Based Representations of Audio Using Siamese Neural Networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 3136–3140. DOI: 10.1109/ICASSP.2018.8461524.

[48] D. Niizumi, D. Takeuchi, Y. Ohishi, *et al.*, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2021. DOI: 10.1109/ijcnn52387.2021.9534474. [Online]. Available: http://dx.doi.org/10.1109/IJCNN52387.2021.9534474.

[49] J. Anton, H. Coppock, P. Shukla, *et al.*, "Audio barlow twins: Self-supervised audio representation learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[50] G. Elbanna, N. Scheidwasser-Clow, M. Kegler, *et al.*, "Byol-s: Learning self-supervised speech representations by bootstrapping," in *HEAR: Holistic Evaluation of Audio Representations*, PMLR, 2022, pp. 25–47.

[51] P.-J. Last, H. A. Engelbrecht, and H. Kamper, "Unsupervised Feature Learning for Speech Using Correspondence and Siamese Networks," *IEEE Signal Processing Letters*, vol. 27, pp. 421–425, 2020, ISSN: 1558-2361. DOI: 10.1109/LSP.2020.2973798.

[52] A. Mehrotra, A. G. C. P. Ramos, N. D. Lane, *et al.*, "Resource Efficient Self-Supervised Learning for Speech Recognition," Sep. 2022.

[53] U. Khan and F. J. Hernando Pericás, "Unsupervised training of siamese networks for speaker verification," in *Interspeech 2020: The 20th Annual Conference of the International Speech Communication Association: 25-29 October 2020: Shanghai, China*, International Speech Communication Association (ISCA), 2020, pp. 3002–3006. DOI: 10.21437/Interspeech.2020-1882.

[54] M. Mohammadamini, D. Matrouf, J.-F. A. Bonastre, *et al.*, "Barlow Twins self-supervised learning for robust speaker recognition," in *Interspeech 2022 - Human and Humanizing Speech Technology*, Incheon, South Korea, Sep. 2022. DOI: 10.21437/Interspeech.2022-11301.

[55] M. Sang, H. Li, F. Liu, *et al.*, "Self-Supervised Speaker Verification with Simple Siamese Network and Self-Supervised Regularization," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6127–6131. DOI: 10.1109/ICASSP43922.2022.9747526.

[56] A. Saeed, D. Grangier, and N. Zeghidour, *Contrastive Learning of General-Purpose Audio Representations*, IEEE, Oct. 2020. DOI: 10.48550/arXiv.2010.10915.

[57] 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.

[58] 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.

[59] L. Wang and A. van den Oord, *Multi-Format Contrastive Learning of Audio Representations*, Mar. 2021. DOI: 10.48550/arXiv.2103.06508.

[60] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575. DOI: 10.21437/Interspeech.2021-698.

[61] 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. DOI: 10.1109/ICASSP39728.2021.9413711.

[62] T. Namgyal, A. Hepburn, R. Santos-Rodriguez, *et al.*, "Data is Overrated: Perceptual Metrics Can Lead Learning in the Absence of Training Data," in *37th Conference on Neural Information Processing Systems (NeurIPS), Machine Learning for Audio Workshop*.

[63] A. van den Oord, Y. Li, and O. Vinyals, *Representation Learning with Contrastive Predictive Coding*, Jan. 2019. DOI: 10.48550/arXiv.1807.03748.

[64] E. Kharitonov, M. Rivière, G. Synnaeve, *et al.*, *Data Augmenting Contrastive Learning of Speech Representations in the Time Domain*, Jul. 2020. DOI: 10.48550/arXiv.2007.00991.

[65] S. Schneider, A. Baevski, R. Collobert, *et al.*, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469. DOI: 10.21437/Interspeech.2019-1873.

[66] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3497–3501.

[67] Y.-A. Chung, W.-N. Hsu, H. Tang, *et al.*, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. Interspeech 2019*, 2019, pp. 146–150. DOI: 10.21437/Interspeech.2019-1473.

[68] Y.-A. Chung and J. Glass, "Improved speech representations with multi-target autoregressive predictive coding," Jan. 2020, pp. 2353–2358. DOI: 10.18653/v1/2020.acl-main.213.

[69] N. Zeghidour, A. Luebs, A. Omran, *et al.*, *Soundstream: An end-to-end neural audio codec*, 2021.

[70] A. Défossez, J. Copet, G. Synnaeve, *et al.*, *High Fidelity Neural Audio Compression*, Oct. 2022. DOI: 10.48550/arXiv.2210.13438.

[71] L. Pepino, P. Riera, and L. Ferrer, *EnCodecMAE: Leveraging neural codecs for universal audio representation learning*, Sep. 2023. DOI: 10.48550/arXiv.2309.07391.

[72] Z. Li, H. Wang, and X. Jiang, *AudioFormer: Audio Transformer Learns Audio Feature Representations from Discrete Acoustic Codes*. Aug. 2023.

[73] 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, 2022, pp. 10 699–10 709.

[74] D. Chong, H. Wang, P. Zhou, *et al.*, "Masked spectrogram prediction for self-supervised audio pre-training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[75] P.-Y. Huang, H. Xu, J. Li, *et al.*, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 708–28 720, 2022.

[76] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked Autoencoding Audio Spectrogram Transformer," in *Proc. INTERSPEECH 2023*, arXiv, Mar. 2022. DOI: 10.48550/arXiv.2203.16691.

[77] S. Yadav, S. Theodoridis, L. K. Hansen, *et al.*, *Masked Autoencoders with Multi-Window Attention Are Better Audio Learners*, Jun. 2023. DOI: 10.48550/arXiv.2306.00561.

[78] S. Ghosh, A. Seth, S. Umesh, *et al.*, "Mast: Multiscale audio spectrogram transformers," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

[79] S. Atito, M. Awais, W. Wang, *et al.*, *ASiT: Audio Spectrogram vIsion Transformer for General Audio Representation*, Nov. 2022. DOI: 10.48550/arXiv.2211.13189.

[80] S. Chen, Y. Wu, C. Wang, *et al.*, *BEATs: Audio Pre-Training with Acoustic Tokenizers*, Dec. 2022. DOI: 10.48550/arXiv.2212.09058.

[81] Z. Fei, M. Fan, and J. Huang, *A-JEPA: Joint-Embedding Predictive Architecture Can Listen*, Nov. 2023. DOI: 10.48550/arXiv.2311.15830.

[82] A. T. Liu, S.-w. Yang, P.-H. Chi, *et al.*, "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6419–6423. DOI: 10.1109/ICASSP40776.2020.9054458.

[83] 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.

[84] A. Baevski, M. Auli, and A. Mohamed, *Effectiveness of self-supervised pre-training for speech recognition*, May 2020. DOI: 10.48550/arXiv.1911.03912.

[85] X. Favory, K. Drossos, T. Virtanen, *et al.*, *COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations*, Jul. 2020. DOI: 10.48550/arXiv.2006.08386.

[86] S.-L. Wu, X. Chang, G. Wichern, *et al.*, *Improving Audio Captioning Models with Fine-grained Audio Features, Text Embedding Supervision, and LLM Mix-up Augmentation*, Sep. 2023. DOI: 10.48550/arXiv.2309.17352.

[87] K. Luo, X. Zhang, J. Wang, *et al.*, "Contrastive latent space reconstruction learning for audio-text retrieval," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2023, pp. 913–917.

[88] Y. Ma, R. Yuan, Y. Li, *et al.*, *On the Effectiveness of Speech Self-supervised Learning for Music*, Jul. 2023. DOI: 10.48550/arXiv.2307.05161.

[89] M. Pulis and J. Bajada, "Siamese neural networks for content-based cold-start music recommendation.," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 719–723.

[90] K. Park, S. Baek, J. Jeon, *et al.*, "Music plagiarism detection based on siamese cnn," *Hum.-Cent. Comput. Inf. Sci*, vol. 12, pp. 12–38, 2022.

[91] R. Agrawal and S. Dixon, "Learning frame similarity using siamese networks for audio-to-score alignment," in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 141–145.

[92] B. Gfeller, C. Frank, D. Roblek, *et al.*, "SPICE: Self-supervised Pitch Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020, ISSN: 2329-9290, 2329-9304. DOI: 10.1109/TASLP.2020.2982285.

[93] A. Riou, S. Lattner, G. Hadjeres, *et al.*, "Pesto: Pitch estimation with self-supervised transposition-equivariant objective," in *International Society for Music Information Retrieval Conference (ISMIR 2023)*, 2023.

[94] E. Quinton, "Equivariant Self-Supervision for Musical Tempo Estimation," in *International Society for Music Information Retrieval Conference (ISMIR)*, arXiv, Sep. 2022. DOI: 10.48550/arXiv.2209.01478.

[95] J. Spijkervet and J. A. Burgoyne, "Contrastive Learning of Musical Representations," in *International Society for Music Information Retrieval (ISMIR)*, arXiv, Sep. 2021. DOI: 10.48550/arXiv.2103.09410.

[96] J. Pons and X. Serra, *Musicnn: Pre-trained convolutional neural networks for music audio tagging*, Sep. 2019. DOI: 10.48550/arXiv.1909.06654.

[97] J. Lee, J. Park, K. L. Kim, *et al.*, *Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms*, May 2017. DOI: 10.48550/arXiv.1703.01789.

[98] 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.

[99] 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. DOI: 10.1145/3485447.3512011.

[100] M. C. McCallum, F. Korzeniowski, S. Oramas, *et al.*, "Supervised and Unsupervised Learning of Audio Representations for Music Understanding," in *International Society for Music Information Retrieval (ISMIR)*, arXiv, Oct. 2022. DOI: 10.48550/arXiv.2210.03799.

[101] S. Stewart, K. Avramidis, T. Feng, *et al.*, *Emotion-Aligned Contrastive Learning Between Images and Music*, Sep. 2023. DOI: 10.48550/arXiv.2308.12610.

[102] C. Garoufis, A. Zlatintsi, and P. Maragos, "Multi-Source Contrastive Learning from Musical Audio," arXiv:2302.07077, arXiv, May 2023. DOI: 10.48550/arXiv.2302.07077.

[103] 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.

[104] M. A. V. Vásquez and J. A. Burgoyne, "Tailed u-net: Multi-scale music representation learning.," in *ISMIR*, 2022, pp. 67–75.

[105] 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, e0294643, Nov. 2023, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0294643.

[106] R. Castellon, C. Donahue, and P. Liang, *Codified audio language modeling learns useful representations for music information retrieval*, Jul. 2021. DOI: 10.48550/arXiv.2107.05677.

[107] C. Donahue, J. Thickstun, and P. Liang, *Melody transcription via generative pre-training*, Dec. 2022. DOI: 10.48550/arXiv.2212.01884.

[108] W. Z. E. Amri, O. Tautz, H. Ritter, *et al.*, "Transfer Learning with Jukebox for Music Source Separation," in vol. 647, 2022, pp. 426–433. DOI: 10.1007/978-3-031-08337-2_35.

[109] J. Copet, F. Kreuk, I. Gat, *et al.*, *Simple and Controllable Music Generation*, Jun. 2023. DOI: 10.48550/arXiv.2306.05284.

[110] G. L. Lan, V. Nagaraja, E. Chang, *et al.*, *Stack-and-Delay: A new codebook pattern for music generation*, Sep. 2023. DOI: 10.48550/arXiv.2309.08804.

[111] H.-H. Wu, C.-C. Kao, Q. Tang, *et al.*, *Multi-Task Self-Supervised Pre-Training for Music Classification*, Feb. 2021. DOI: 10.48550/arXiv.2102.03229.

[112] Y. Li, R. Yuan, G. Zhang, *et al.*, *MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training*, Jun. 2023.

[113] Y. Li, R. Yuan, G. Zhang, *et al.*, *MAP-Music2Vec: A Simple and Effective Baseline for Self-Supervised Music Audio Representation Learning*, Dec. 2022. DOI: 10.48550/arXiv.2212.02508.

[114] I. Manco, E. Benetos, E. Quinton, *et al.*, "Learning music audio representations via weak language supervision," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 456–460.

[115] M. Won, Y.-N. Hung, and D. Le, *A Foundation Model for Music Informatics*, Nov. 2023. DOI: 10.48550/arXiv.2311.03318.

[116] E. Law, K. West, M. I. Mandel, *et al.*, "Evaluation of algorithms using games: The case of music tagging.," in *ISMIR*, Citeseer, 2009, pp. 387–392.

[117] 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)*, Long Beach, CA, United States, 2019. [Online]. Available: http://hdl.handle.net/10230/42015.

[118] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, *et al.*, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[119] R. Yuan, Y. Ma, Y. Li, *et al.*, *MARBLE: Music Audio Representation Benchmark for Universal Evaluation*, Jul. 2023. DOI: 10.48550/arXiv.2306.10548.

[120] C. Plachouras, P. Alonso-Jiménez, and D. Bogdanov, "Mir$_{ref}$: A representation evaluation framework for music information retrieval tasks," in *37th Conference on Neural Information Processing Systems (NeurIPS), Machine Learning for Audio Workshop*, New Orleans, LA, USA, 2023.

[121] M. Anwar, B. Shi, V. Goswami, *et al.*, *MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation*, Mar. 2023. DOI: 10.48550/arXiv.2303.00628.

[122] C. D. Kim, B. Kim, H. Lee, *et al.*, "AudioCaps: Generating Captions for Audios in The Wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132. DOI: 10.18653/v1/N19-1011.

[123] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 609–617.

[124] J. Wang, J. Jiao, Y. Song, *et al.*, *Speed Co-Augmentation for Unsupervised Audio-Visual Pre-training*, Sep. 2023. DOI: 10.48550/arXiv.2309.13942.

[125] S. Jenni, A. Black, and J. Collomosse, *Audio-Visual Contrastive Learning with Temporal Self-Supervision*, Feb. 2023. DOI: 10.48550/arXiv.2302.07702.

[126] J. Wilkins, J. Salamon, M. Fuentes, *et al.*, *Bridging High-Quality Audio and Video via Language for Sound Effects Retrieval from Visual Queries*, Aug. 2023. DOI: 10.48550/arXiv.2308.09089.

[127] H.-W. Dong, N. Takahashi, Y. Mitsufuji, *et al.*, *CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos*, Mar. 2023. DOI: 10.48550/arXiv.2212.07065.

[128] L. Wang, P. Luc, A. Recasens, *et al.*, *Multimodal Self-Supervised Learning of General Audio Representations*, Apr. 2021. DOI: 10.48550/arXiv.2104.12807.

[129] C. Schuhmann, R. Beaumont, R. Vencu, *et al.*, *LAION-5B: An open large-scale dataset for training next generation image-text models*, 2022. DOI: 10.48550/ARXIV.2210.08402.

[130] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[131] A. Guzhov, F. Raue, J. Hees, *et al.*, "Audioclip: Extending Clip to Image, Text and Audio," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 976–980. DOI: 10.1109/ICASSP43922.2022.9747631.

[132] H.-H. Wu, P. Seetharaman, K. Kumar, *et al.*, "Wav2CLIP: Learning Robust Audio Representations from Clip," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 4563–4567. DOI: 10.1109/ICASSP43922.2022.9747669.

[133] B. Elizalde, S. Deshmukh, M. Al Ismail, *et al.*, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[134] Y. Yuan, H. Liu, X. Liu, *et al.*, *Retrieval-augmented text-to-audio generation*, Sep. 2023. DOI: 10.48550/arXiv.2309.08051.

[135] Y. Yuan, H. Liu, X. Liu, *et al.*, *Text-Driven Foley Sound Generation With Latent Diffusion Model*, Sep. 2023. DOI: 10.48550/arXiv.2306.10359.

[136] H. Liu, Q. Tian, Y. Yuan, *et al.*, *AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining*, Sep. 2023. DOI: 10.48550/arXiv.2308.05734.

[137] H. Liu, Z. Chen, Y. Yuan, *et al.*, *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models*, Sep. 2023. DOI: 10.48550/arXiv.2301.12503.

[138] J. Huang, Y. Ren, R. Huang, *et al.*, *Make-An-Audio 2: Temporal-Enhanced Text-to-Audio Generation*, May 2023. DOI: 10.48550/arXiv.2305.18474.

[139] D. Ghosal, N. Majumder, A. Mehrish, *et al.*, *Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model*, May 2023. DOI: 10.48550/arXiv.2304.13731.

[140] S. S. Kushwaha and M. Fuentes, *A Multimodal Prototypical Approach for Unsupervised Sound Classification*, Aug. 2023. DOI: 10.48550/arXiv.2306.12300.

[141] S. Deshmukh, B. Elizalde, and H. Wang, *Audio Retrieval with WavText5K and CLAP Training*, Sep. 2022. DOI: 10.48550/arXiv.2209.14275.

[142] X. Liu, Q. Kong, Y. Zhao, *et al.*, *Separate Anything You Describe*, Aug. 2023. DOI: 10.48550/arXiv.2308.05037.

[143] K. Koutini, J. Schlüter, H. Eghbal-zadeh, *et al.*, "Efficient Training of Audio Transformers with Patchout," in *Interspeech 2022*, Sep. 2022, pp. 2753–2757. DOI: 10.21437/Interspeech.2022-227.

[144] B. Elizalde, S. Deshmukh, and H. Wang, *Natural Language Supervision for General-Purpose Audio Representations*, Sep. 2023. DOI: 10.48550/arXiv.2309.05767.

[145] P. Primus, K. Koutini, and G. Widmer, *Advancing Natural-Language Based Audio Retrieval with PaSST and Large Audio-Caption Data Sets*, Aug. 2023. DOI: 10.48550/arXiv.2308.04258.

[146] X. Favory, K. Drossos, T. Virtanen, *et al.*, *Learning Contextual Tag Embeddings for Cross-Modal Alignment of Audio and Tags*, Oct. 2020.

[147] C. Qiang, H. Li, Y. Tian, *et al.*, *Learning Speech Representation From Contrastive Token-Acoustic Pretraining*, Sep. 2023. DOI: 10.48550/arXiv.2309.00424.

[148] 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, ISSN: 1070-9908, 1558-2361. DOI: 10.1109/LSP.2021.3071082.

[149] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from discogs," in *Rao P, Murthy H, Srinivasamurthy A, Bittner R, Caro Repetto R, Goto M, Serra X, Miron M, editors. Proceedings of the 23nd International Society for Music Information Retrieval Conference (ISMIR 2022); 2022 Dec 4-8; Bengaluru, India.[Canada]: International Society for Music Information Retrieval; 2022. p. 825-33.*, International Society for Music Information Retrieval (ISMIR), 2022.

[150] S. Durand, D. Stoller, and S. Ewert, "Contrastive Learning-Based Audio to Lyrics Alignment for Multiple Languages," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096725.

[151] Z. He, W. Hao, W.-T. Lu, *et al.*, *ALCAP: Alignment-Augmented Music Captioner*, Dec. 2022. DOI: 10.48550/arXiv.2212.10901.

[152] I. Manco, E. Benetos, E. Quinton, *et al.*, *Contrastive Audio-Language Learning for Music*, Aug. 2022. DOI: 10.48550/arXiv.2208.12208.

[153] Q. Huang, A. Jansen, J. Lee, *et al.*, *MuLan: A Joint Embedding of Music Audio and Natural Language*, Aug. 2022. DOI: 10.48550/arXiv.2208.12415.

[154] S. Doh, M. Won, K. Choi, *et al.*, *Toward universal text-to-music retrieval*, 2022.

[155] T. Chen, Y. Xie, S. Zhang, *et al.*, "Learning Music Sequence Representation From Text Supervision," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 4583–4587. DOI: 10.1109/ICASSP43922.2022.9746131.

[156] J. Gardner, S. Durand, D. Stoller, *et al.*, *LLark: A Multimodal Foundation Model for Music*, Oct. 2023.

[157] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, *et al.*, *AudioPaLM: A Large Language Model That Can Speak and Listen*, Jun. 2023. DOI: 10.48550/arXiv.2306.12925.

[158] C. Wang, S. Chen, Y. Wu, *et al.*, *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*, Jan. 2023. DOI: 10.48550/arXiv.2301.02111.

[159] X. Mei, X. Liu, Q. Huang, *et al.*, *Audio Captioning Transformer*, Jul. 2021. DOI: 10.48550/arXiv.2107.09817.

[160] M.-I. Georgescu, E. Fonseca, R. T. Ionescu, *et al.*, *Audiovisual Masked Autoencoders*, Jul. 2023. DOI: 10.48550/arXiv.2212.05922.

[161] S. Doh, K. Choi, J. Lee, *et al.*, *LP-MusicCaps: LLM-Based Pseudo Music Captioning*, Jul. 2023. DOI: 10.48550/arXiv.2307.16372.

[162] Y. Zhang, J. Jiang, G. Xia, *et al.*, *Interpreting Song Lyrics with an Audio-Informed Pre-trained Language Model*, Aug. 2022. DOI: 10.48550/arXiv.2208.11671.

[163] T. Cai, M. I. Mandel, and D. He, "Music autotagging as captioning," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, Online: Association for Computational Linguistics, 2020, pp. 67–72.

[164] I. Manco, E. Benetos, E. Quinton, *et al.*, "MusCaps: Generating Captions for Music Audio," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533461.

[165] A. Agostinelli, T. I. Denk, Z. Borsos, *et al.*, *MusicLM: Generating Music From Text*, Jan. 2023. DOI: 10.48550/arXiv.2301.11325.

[166] S. Liu, A. S. Hussain, C. Sun, *et al.*, *Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning*, Aug. 2023. DOI: 10.48550/arXiv.2308.11276.

[167] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, Apr. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-022-10246-w. (visited on 12/17/2023).

[168] A. Tharwat and W. Schenck, "A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions," *Mathematics*, vol. 11, no. 4, p. 820, Jan. 2023, ISSN: 2227-7390. DOI: 10.3390/math11040820. (visited on 10/13/2023).

[169] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 417–424.

[170] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[171] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[172] A. Casanova, P. O. Pinheiro, N. Rostamzadeh, and C. J. Pal, "Reinforced active learning for image segmentation," *arXiv preprint arXiv:2002.06583*, 2020.

[173] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 580–588.

[174] D. Angluin, "Queries and concept learning," *Machine learning*, vol. 2, pp. 319–342, 1988.

[175] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," in *Acm Sigir Forum*, ACM New York, NY, USA, vol. 29, 1995, pp. 13–19.

[176] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active learning from stream data using optimal weight classifier ensemble," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 6, pp. 1607–1621, 2010.

[177] D. Aldoğan and Y. Yaslan, "A comparison study on active learning integrated ensemble approaches in sentiment analysis," *Computers & Electrical Engineering*, vol. 57, pp. 311–323, 2017.

[178] H. V. Koops, G. Micchi, I. Manco, and E. Quinton, *Serenade: A Model for Human-in-the-loop Automatic Chord Estimation*, Oct. 2023. arXiv: 2310.11165 [cs, eess]. (visited on 11/28/2023).

[179] K. Yamamoto, "HUMAN-IN-THE-LOOP ADAPTATION FOR INTERACTIVE MUSICAL BEAT TRACKING," 2021.

[180] D. Su and P. Fung, "Personalized music emotion classification via active learning," in *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, ser. MIRUM '12, New York, NY, USA: Association for Computing Machinery, Nov. 2012, pp. 57–62, ISBN: 978-1-4503-1591-3. DOI: 10.1145/2390848.2390864. (visited on 12/17/2023).

[181] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Support vector machine active learning for music retrieval," *Multimedia Systems*, vol. 12, no. 1, pp. 3–13, Aug. 2006, ISSN: 1432-1882. DOI: 10.1007/s00530-006-0032-2. (visited on 12/17/2023).

[182] G. Chen, T.-J. Wang, L.-Y. Gong, and P. Herrera Boyer, "Multi-class support vector machine active learning for music annotation," 2010, ISSN: 1751-648X. (visited on 12/17/2023).

[183] L. Chen and C. Raphael, "Optical Music Recognition and Human-in-the-loop Computation," Jun. 2018. (visited on 12/17/2023).

[184] Y. Zhang, A. Maezawa, G. Xia, K. Yamamoto, and S. Dixon, *Loop Copilot: Conducting AI Ensembles for Music Generation and Iterative Editing*, Oct. 2023. DOI: 10.48550/arXiv.2310.12404. arXiv: 2310.12404 [cs, eess]. (visited on 11/29/2023).

[185] B. Han, J. Dai, X. Song, *et al.*, *InstructME: An Instruction Guided Music Edit And Remix Framework with Latent Diffusion Models*, Sep. 2023. DOI: 10.48550/arXiv.2308.14360. arXiv: 2308.14360 [cs, eess]. (visited on 11/29/2023).

[186] K. He, X. Chen, S. Xie, *et al.*, "Masked autoencoders are scalable vision learners," *CoRR*, vol. abs/2111.06377, 2021. [Online]. Available: https://arxiv.org/abs/2111.06377.

[187] Z. Huang, X. Jin, C. Lu, *et al.*, *Contrastive Masked Autoencoders are Stronger Vision Learners*, Nov. 2022.

[188] Y. Yao, N. Desai, and M. Palaniswami, *Masked Contrastive Representation Learning*, Nov. 2022.

[189] S. Mishra, J. Robinson, H. Chang, *et al.*, *A simple, efficient and scalable contrastive masked autoencoder for learning visual representations*, Oct. 2022. DOI: 10.48550/arXiv.2210.16870.

[190] P. Khosla, P. Teterwak, C. Wang, *et al.*, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[191] C. Tao, X. Zhu, W. Su, *et al.*, *Siamese Image Modeling for Self-Supervised Vision Representation Learning*, Nov. 2022. arXiv: 2206.01204 [cs]. (visited on 12/14/2023).

[192] X. Yuan, Z. Li, and G. Wang, "ActiveMatch: End-To-End Semi-Supervised Active Representation Learning," in *2022 IEEE International Conference on Image Processing (ICIP)*, Oct. 2022, pp. 1136–1140. DOI: 10.1109/ICIP46576.2022.9898008. (visited on 12/15/2023).

Appendix A

MaCHUP : Masked Contrastive Hierarchical Unsupervised Pretraining for music classification

A.1 Background and Motivation

Learning general music representations suitable for multiple downstream tasks is a field of growing interest, garnering several recent publications on foundation models and unified benchmarks for better comparison of learned representations. Specifically, two studies - MERT [112], the to-date state of the art on multiple downstream tasks [119], and EncodecMAE [71], a general-purpose audio representation learning system evaluated on musical genre classification and pitch classification - leverage Encodec [70] to learn these representations via masked modeling tasks.

Encodec [70] is a RVQ Hierarchical Neural audio codec, which compresses continuous bottleneck embeddings into a sequence of hierarchical discrete tokens. EncodecMAE and MERT have leveraged the information contained in these sequences by using them as reconstruction targets for a masked modeling task using transformer architectures. Although some differences exist, namely EncodecMAE using frozen continuous embeddings from Encodec and a MAE-like [186] method in dropping out masked tokens, and MERT using its own Convolutional waveform encoder and keeping masked timesteps to be fed to the encoder - The principle is mostly the same, and has shown state-of-the-art results on Multiple tasks.

However, these approaches, while successful, are not specifically designed for music beyond MERT using a second CQT-reconstruction task. I argue that by "only" reconstructing RVQ embeddings (and CQT frames), the information contained in the learned representations is mostly perceptual, and not necessarily semantically meaningful. Additionally, MAE-like setups that save a lot of memory during training show high potential for the encoding of longer sequences, a desired trait for music representation learning systems, which so far have only been able to fit 30 seconds of music into their encoder [119], where a typical music piece is about six times as long.

Infusing additional semantic information into learned representations has been successful through combining masked modeling and contrastive learning. Multiple studies in the computer vision field have shown that combining both tasks can lead to increased performance and better semantic separation in the latent space [187], [188] - a desirable property for a perceptual-oriented system such as the previously described music representation learners. Until recently, these systems were complex, often requiring a second momentum encoder which sees the whole image, not only the unmasked token - which negates the potential of MAE for longer sequences. Recently, CAN [189] proposes a simple end-to-end approach which combines autoencoding and masked modeling with a single model, and improves upon the baselines of either method on their own - this opens the way for similar approaches in music representation learning with potential similar improvements over existing state of the art.

The aims of this work will be threefold : Adapt the CAN architecture with RVQ tokens as a reconstruction target to showcase the potential of a simple, combined autoencoding and contrastive approach for general representation learning, propose a first attempt at learning representations over song-level sequence lengths, and propose a novel hierarchal task combining autoencoding and contrastive learning which includes musical knowledge in the representation learning task design.

A.2 Methods

Dual masked-contrastive learning task

As mentioned previously, combinations of the contrastive and masked modeling objectives are commonplace, either end-to-end or in two stages. This serves the purpose of disentangling representations by proposing two self-supervised tasks of different kinds : the masked modeling task instills the learned representations with structural knowledge about the input, and the contrastive task disentangles and co-aligns representations within the latent space to provide pseudo-clusters.

Approximately following the design of EncodecMAE [71], we propose the following dual-task setup for end-to-end training of Encodec representations: The encodec encoder is frozen and used to generate training targets and inputs. A proportion p of the sequence is masked and fed into a transformer encoder, which outputs a sequence of embeddings of the same shape as the discarded sequence.

This encoded sequence is then used for two tasks : First, it is leveraged for either simpler traditional contrastive learning task, or our novel hierarchical contrastive learning task, which will be covered more extensively in the following section. It is also unmasked to be the same size as the input sequence to the encoder and fed to a comparatively smaller **de**coder for the masked reconstruction task. At fine-tuning and probing time, only the outputs of the encoder are used. (See figure A.1)

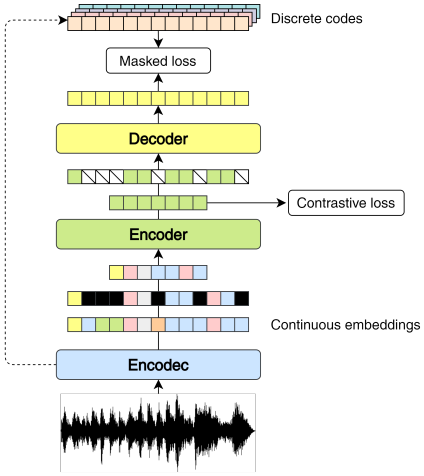


Figure A.1: Dual masked-contrastive learning task pipeline. output embeddings from the transformer encoder are used for the contrastive task, while the decoder reconstructs the masked tokens.

When combining these two training objectives, the spirit of one is prevalent over the other. In a contrastive-first setup, multiple views of the same sample are augmented and contrasted. In a masked-modeling first setup, the sample is not augmented into multiple samples and the contrastive information must come from within the structure of the sample itself. the A previous study, Audioformer [72], uses the contrastive task as a secondary objective, considering codebooks of the same timestep to be positive samples, and of different time-steps to be negative samples. We find this to be counter-intuitive for music. Firstly, there is no reason that different hierarchical representations of the signal at a given time-step should be similar in semantic content - if they were, there would be no need for hierarchical levels. Further, music is highly locally correlated, so considering other timesteps to be negative samples is not a musically sound approach.

Using this and given the success of the contrastive-first approach in previous studies leveraging contrastive learning for music, we decide to use this approach and consider the masked reconstruction task as our secondary objective.

First Iteration : MAE and Contrastive learning with global representations

In a first iteration, we consider a combination of masked modeling and contrastive learning with the contrastive approach based on global representations of the input sequence. consider a batch B of un-augmented data points $\{x_i\}, i \in \{1, 2...N\}$. We apply M augmentations by chunking the audio into non-overlapping chunks and applying a stochastic augmentation pipeline described in [95] such that

$\{x_i\}$ becomes $\{\tilde{x}_j\}, j \in \{1, 2.....M * N\}$. for simplicity, we maintain the notation x_i.

After processing by the frozen Encodec [71] - either by directly using the continuous output embeddings or passing each codebook through a trainable embedding layer , each of our samples within a batch B x_i becomes a sequence of embeddings $x_i = (x_i^1....x_i^T)$ where T is the length of the sequence output - dependent on the selected length of the signal T_s in the time domain and the hop size of the encodec model. Encodec also produces Q sequences of codes streams $C_i^k, k \in [1...Q] = (C_i^{k,t}), t \in [1....T]$ which act as our target sequence for the masked modeling task. For simplicity we use x, C for further reference to the various i indexed variables.

We mask a proportion p of tokens in the input sequence, corresponding to a number of masked tokens $|M|$ where M is the set of masked indices. These masked tokens are discarded from the original embedding sequence E such that $x_M = (x^t), t \notin M$. a shared learned class token x_{CLS} is appended to the beginning of the sequence, and the input sequence is processed by a transformer encoder into masked encoded embeddings e_M:

$$e_M = E(x_M)$$

These embeddings are then "unmasked", with the original masked indices being replaced by a shared learned mask token e_M and unmasked indices being kept as the encoded sequence. This unmasked encoded sequence is fed into a transformer **de**coder and the output is a sequence of logits d. In our original design, the encoded masked sequence e_M is used for the contrastive task by being fed into a projection head g: $z = g(e)$.

We use the following contrastive loss between global representation projections z_i of samples within the minibatch, an instance of supervised contrastive loss :

$$\mathcal{L}_{sup}^{out} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(sim(z_i, z_p)/\tau)}{\sum_{a \in A(i)} \exp(sim(z_i, z_a)/\tau)} \tag{A.1}$$

$P(i)$ is the set of positive samples for index i, in this case the set of indices corresponding to augmentations of the original sample, $A(i)$ is the set of indices in the multiview batch, sim is a similarity metric (in our case, cosine similarity).

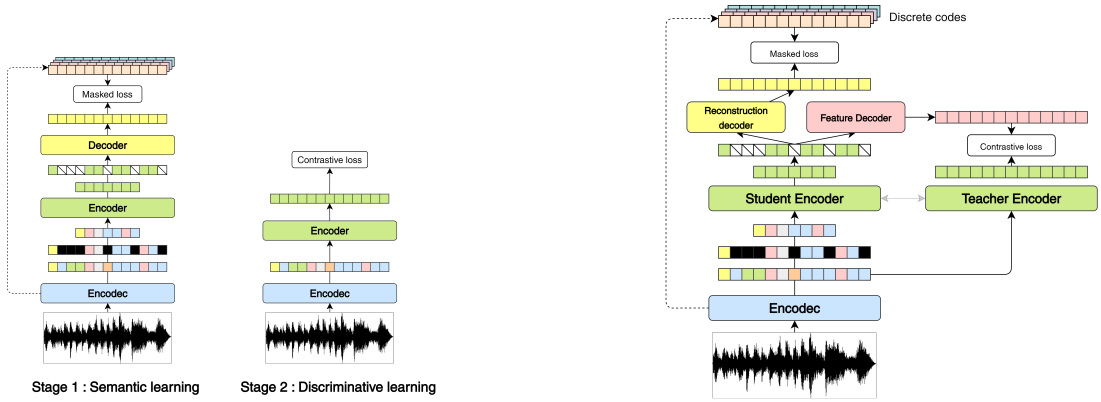
The global representations used for the contrastive task can be chosen to be the z^{CLS} token appended to the beginning of each sequence before processing by the encoder - never masked - or the mean average pooling along the time axis for each sample. This is a design choice which will be empirically selected according to results. Consider it is also possible to use a combination of both global representations for the contrastive objective by computing the contrastive loss for both and linearly combining them into a single loss term.

The decoded logits are used for the reconstruction task by projecting them through one linear classification head per codebook, then evaluating them against the targets C, as seen figure A.1 using the usual cross-entropy loss, called reconstruction loss $\mathcal{L}_r$:

$$\mathcal{L}_r(C, d) = -\frac{1}{QT} \sum_{t,Q} C^{Q,t} \log d^{Q,t}$$

End-to-end vs two-stage approach vs two-model approach

In our original approach, the whole model is trained end-to-end, with bottleneck masked representations from the encoder used as both inputs and targets for the contrastive task (see figure A.1). However, approaches in literature generally use different approaches. One is to use two stage approaches, notably by pretraining on the masked task first and then the contrastive task (presumably to infuse encoded representations with semantic information before, as mentioned in **empty citation**), and as shown to work in contrastive representation learning for natural language using pretrained bert models (see figure A.2a). This approach poses the issue of being clunkier, taking longer to train, and requiring additional exploration of loss balancing and hyperparameter tuning during the different stages for optimal performance, as well as which global representation to use for the second contrastive stage. Further, by retraining over learned representations, there is the risk of forgetting the advantageous representations learned previously without keeping the reconstruction task enforced during the secondary stage.



(a) Two-stage approach for MaCHUP

(b) Two-model approach for MaCHUP

Figure A.2: Comparison of approaches for MaCHUP

Other previous works use a dual encoder setup, as in **empty citation** where the target representations for the contrastive task are taken from a teacher encoder which sees the whole input. The masked encoded representationsfrom the "online" - or student - encoder are fed into a "reconstruction decoder", the output of which is used as input for the reconstruction loss, and a "feature encoder", the output of which is used for the contrastive task against the teacher representations (see figure A.2b).

The issue with this approach is the memory and compute requirements for doubling already voluminous encoders and decoders. It remains to be seen if the teacher and student encoders can be the same model with shared weights, although **empty citation** uses two different models and the outputs of the student encoder as representations for downstream probing tasks. However, it makes sense that the feature decoder and the representation decoder be separate, as they aim to minimize loss objective for semantically different tasks, as shown in **empty citation**

Considering these two optional approaches, and their respective advantages and disadvantages, we attempt to prioritize our end-to-end approach, and choose to explore the other two with lower priority, though we keep them in mind should the default approach fail.

Masked Autoencoding design choices

We aim to leverage masked autoencoding as a secondary task to ensure the learned representations are acoustically meaningful. we design the following task : given a sequence of encodec inputs, a proportion p of these inputs are masked (discarded) and the resulting $(1-p) * T$ representations are fed into a bidirectional transformer encoder (potentially with linear attention to accommodate the memory requirements for longer sequences). the discarded masked tokens are then re-introduced into the encoded representation embedding sequence and the reformed T-sized sequence is fed into a bidirectional transformer decoder, with the output of this decoder used for the reconstruction task (See figure A.1).

It has been shown in many studies that discarding masked tokens is beneficial for model training, as it discards unnecessary information in the decoder, acts as a regularizer for the model, and saves large amounts of memory during training. We elect to use this technique regardless of other design choices. We elect to weigh the loss of the masked task by masked and unmasked tokens as has proven to be benefecial for the performance of the model and crucial in de-trivializing the reconstruction task. Regardless of the loss function $\mathcal{L}_r$ used for the reconstruction task, we formalize the total loss $\mathcal{L}_r^T$

$$\mathcal{L}_r^T = \frac{\delta}{|M|}\mathcal{L}_r(C_i, d_i * 1_{i \in M}) + \frac{1-\delta}{T-|M|}\mathcal{L}_r(C_i, d_i * 1_{i \notin M}) \tag{A.2}$$

Further, we also aim to balance the contrastive and reconstruction loss with a hyperparameter λ:

$$\mathcal{L} = \lambda \mathcal{L}_r^T + (1-\lambda)\mathcal{L}_c \tag{A.3}$$

where $\mathcal{L}_c$ is the loss for the contrastive task.

Second iteration : Local-window hierarchical contrastive learning

The intuition of our second contrastive learning task, specifically designed for music, is as follows : music is not a sum of its parts. Furthermore, music has local correlations that describe "instantaneous" views of a music piece, but the aggregate representation of the piece is not straightforward. It is clear that the global representation of a piece of music is linked to its local components, but in no staightforward way.

We design a novel contrastive learning task, in which local representations of the music (given by time-step embeddings) are pushed together in the latent space, and pushed away from the global representation, considered here to be the class token embedding in the BERT-setup.

As in a classic contrastive learning setup, we want the global representations of augmented views of the same sample to be considered as similar. We also want the global representation of each sample to **push away** the local representations of the same sample. Because music is highly correlated locally, we introduce a **positive sample window**, in which local features from the same sample at the same timestep are considered as positive samples, and local features outside of this window are considered as "neutral" - not pushed away, not pushed together, simply ignored. This is to alleviate the potential impact of a hard switch from positive to negative samples at the border of an arbitrarily-decided window. This re-joins developments beyond CLMR that show that the best positive sampling strategy is simply adjacent sections of audio clips However, local features from one sample are pushed away from the same timestep in other samples, given that the samples are non-overlapping and/or of different audio clips altogether. Figure A.3 shows the whole process, with a batch size of 2, each sample augmented into 2 samples, and a sequence length of 3. The window for local contrastive learning is of size 2.

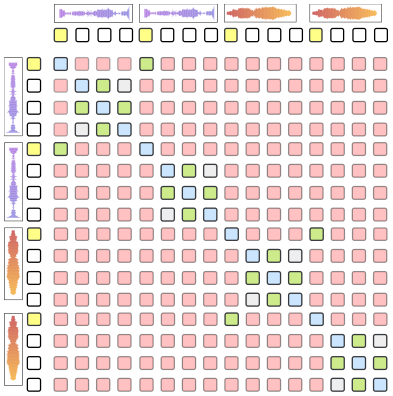


Figure A.3: *Overview of the pseudo-supervised contrastive learning objective. light red cells show negatives, blue cells show ignored self-anchors, green cells show positives, and greyed-out cells show ignored neutral correspondences. Yellow cells in the original sequence show class tokens*

In summary, global views of audio samples are pushed together if they are from the same sample, away if not, and are always pushed away from local features. Local features are pushed together if within the same window of the same sample, apart if they are from a different sample, including augmented, and ignored if outside the "correlation window" within a same sample. using the same supervised contrastive loss function defined in A.1. In this case, $P(i)$ is determined by whether or not i is the index of a global representation token, and if not, the size of the window w. Further, we slightly modify the implementation so that $A(i)$ is the set of indices in the multiview batch that are either positive or negative, not neutral.

The intuition is that this contrastive task, while complex, will provide a hierarchical latent space with global representations at the center of a "cloud" of local representations. It also opens alleys such as hard vs. soft negatives, e.g. global representations could be soft negatives of local representations from the same sample and hard negatives for local representations of other samples, likewise for local representations between themselves. Fine-grained temporal features can be uses as-is for time-sensitive self-similarity downstream tasks, and can be aggregated to whatever target length for a meso-representation. Global representations can be used on their own for downstream evaluation tasks and probing. Further, we surmise that the masked and discarded tokens actually provide a regularization technique for the contrastive task by removing directly adjacent tokens stochastically.

First results

For this first study, we evaluate the performance of five models, all trained End-to-End, as in Figure A.1: **MaCHUP-R**, where only he reconstruction loss is used, **MaCHUP-C**, where only the contrastive task is used, and three variants of **MaCHUP-2L**, trained end to end with two losses : with and without augmentations and $\lambda = 0.1$, and with augmentations and $\lambda = 0.5$. Without augmentations, adjacent chunks of audio clips are used as positive samples but without applying augmentations to the individual clips. We use the Encodec-32kHz pretrained for MusicGen by FAIR [109] for our input representations and targets. Note that MaCHUP-R is simply EncodecMAE-B with less data and slightly different hyperparameters ($\delta = 0.8$ vs 0.9, per-codebook weighting which "slightly improves validation performance" [71], and a different Encodec model.)

We train for 100 epochs with a learning rate reducer on validation loss plateaus for fair comparison between models. We will later conduct studies on the performance for longer training. We use the Adam optimizer from pytorch with default values. We train baseline runs with two augmentations per sample, batch size of 32*2=64, contrastive temperature of $\tau = 0.1$, and masked vs unmasked reconstruction loss ratio of $\delta = 0.8$. The masking ratio when masking is applied is set to $p = 0.5$. For our preliminary comparison study we train with $6s$ of audio on MTG-Jamendo [117] as our pretraining dataset.

For evaluation of learned representations, in this first approach, we propose finetuning and evaluating on two tasks : automatic tagging on the MTG-Jamendo top 50 tags task and the MagnaTagATune [116] top 50 tags task, and genre classification on the GTZAN dataset. We use a 2-layer MLP similar to the setup adopted in [119] for fine-tuning, and also report evaluation with a simple linear classifier on top of the frozen model. Further down the line we will add additional evaluation tasks, matching those presented in **MARBLE** [119]. All finetuning experiments are trained for 1000 epochs on GTZAN, 100 epochs for automatic tagging, and evaluation is performed by averaging these learned representations from the best validation model over all the $6s$ chunks of the song. Results are reported in table for our models, as well as other model performances reported in [119].

			Loss Values		Evaluation Task		
					GTZAN	MTAT top 50 tags	
Model	λ	Aug	R	C	Acc	AUROC	AP
Ours							
Machup-R	1	✗	4.27	-	0.802	0.8829	0.3952
MaCHUP-C	0	✓	-	1.45	0.849	0.8836	0.3998
	0.1	✓	4.0	1.0	0.811	0.8856	0.4041
MaCHUP-2L	0.1	✗	4.2	0.75	0.773	0.8749	0.375
	0.5	✓	1.0	4.3		0.8969	**0.4246**
Literature							
CLMR [95]	-				0.652 †	0.895 † / 0.893 *	0.36 †*
JukeBox 5B [106]	-				0.728 †	**0.914** †	0.406 †
MERT-95M-public [112]	-				0.728 †	0.907 †	0.384 †
MERT-330M [112]	-				0.776 †	0.911 †	0.395 †
EncodecMAE [71]	-				**0.862** *	-	-

Table A.1: Preliminary results for training and evaluation using a pure contrastive and pure masked modeling baseline. Results denoted † by are reported in MARBLE [119]. Results denoted by $*$ are reported in their original paper.

Despite being trained for only 100 epochs on MTG-Jamendo (compared to MERT which is trained on substantially more data for longer, and CLMR and EncodecMAE which are trained both longer and on more data **or** on the downstream dataset, we achieve competitive results. Training for longer (closer to 1000 epochs or until convergence) should substantially improve results and land MaCHUP close to state of the art.

Furthermore, combining both losses consistently outperforms either single baseline on its own. A lower λ parameter already provides small performance boosts, and $\lambda = 0.5$, an arbitrary value so far, provides close to **1.5%** performance increase on the evaluation tasks - suggesting that further investigating informed loss balancing can lead to an even higher performance model.

Future work

Despite these promising results, there are still some steps to be taken before a publication is ready for this work. Firstly, models must be trained on additional data, to match the scale of recent models **MERT**, [71] and make the comparison fair - I also plan to train the best model on only public data to be able to provide pretrained weights and open data.

pretraining and downstream evaluation must also be standardized for fair comparison between our own models. The two-stage approach is to be tested, and dependent on the available time before conference deadlines, the two-model approach as well. Once the best model has been isolated, ablation studies, hyperparameter searches and final training runs to convergence will be undertaken to report the best possible results.

Hyperparameter searches will be undertaken over contrastive temperature τ, contrastive-to-masked ratio λ, masked-to-unmasked-ratio δ, number of augmentations M, batch size B, masking ratio . If it is shown that the model is robust to high masking ratios (as is [186]), I will be able to run training and evaluation on sequence lengths of up to three minutes (a 50Hz embedding framerate with an high masking ratio of 0.8 would mean an entire song (180 seconds) could fit within 1800 tokens during training - not uncommon for transformer models.

Ablation studies will include:

- Removing masking altogether and during the second stage of the two-stage approach
- Using class tokens versus average pooled representations for both pretraining with the contrastive task and downstream finetuning

And of course, future work will be focused on the second iteration and the implementation of the novel hierarchal contrastive learning task, contingent on the success of the first iteration.

Planned dissemination

Because of the broader interest of fitting long sequences into audio representation learning models and combining contrastive learning and masked autoencoding with encodec as a target, the target conference for this publication will be **ICASSP**, with a submission deadline around september. This will give the project ample time to develop and for the second phase to produce results before publication. **ICLR**, with a submission deadline in october, will be another target conference. Should phase 1 produce results before mid-april, the usual deadline for **ISMIR** submission, then it will be submitted to ISMIR and phase 2 will be prioritized for **ICASSP** and **ICLR**.

Appendix B

Semi-Supervised Contrastive learning for active representation learning

B.1 Background and aim

Contrastive learning has emerged as a successful paradigm for representation in many domains, including music information retrieval [95], [98], [104]. However one weak point of contrastive learning is the negative sampling strategy, which is a limiter for the performance of contrastive learning approaches. Because pure self-supervised contrastive learning relies only on augmented views of a same sample as positive samples, this can cause "semantic blurring", where samples that would, from whichever metrics, be considered positive are in effect considered as negatives in the latent space (e.g two samples of the same class), causing less separation of relevant samples and thus lesser downstream performance on relevant evaluation tasks. One approach to solve this issue has been supervised contrastive learning [190], which reverts to a fully-supervised approach using the labels of the data points to extract positive and negative information for the contrastive task. This method boasts higher resistance to data corruption, more robustness to batch size, and an inherent performance improvement with the number of positives in the batch when compared to contrastive learning [5].

However, this approach also has its flaws. it is highly specific to the supervised dataset it is trained on, and learned representations are still highly dependent on the augmentation chain selected - because the network is still trained to be invariant to these augmentations. In response to this, a few methods have emerged attempting to approach semi-supervised contrastive learning with relative success, but all with complex machinery, leveraging pseudo-label classification heads and confidence or two-stage approaches for their architecture [105], [191], [192].

Our approach aims to propose a simple framework for semi-supervised contrastive learning that can :

- Improve upon fully self-supervised learning contrastive approaches by leveraging labeled datasets, even with a small amount of labeled data
- Infuse the learned representations with desired characteristics on top of a general self-supervised learned representations by selecting the supervised dataset and augmentations to augment the self-supervised training with
- Provide a **by design** representation learning system that can integrate a human in the loop without any modifications to improve not only downstream performance but representation quality itself.

This approach will be considered in the context of music classification, but can easily be extended to other areas. The second and third contributions are specifically relevant for music information retrieval, as many MIR tasks are highly subjective, and music is teeming with desirable properties to infuse learned representations with.

B.2 Methods

In this work, we propose a novel approach to representation learning with human in the loop potential : Semi-supervised contrastive learning. Esentially, we propose leveraging labeled data when available as support for a contrastive learning task to learn general representations of music.

Self-Supervised contrastive learning

In the SSL setting for traditional approaches such as SimCLR and CLMR, each batch of data points is augmented twice. let B be a batch of augmented data points x_i. Indices $i \in I = \{1, 2...2N\}$ represent the position of the data point in the batch. $j(i)$ represents the index of the other augmented data point from the same original sample in the batch ($i+1$ or $i-1$ in the canonical approach to contrastive learning). Let z_i be the projection of the encoded data point by an encoder E and a projection head g into the contrastive latent space : $z_i = g(E(x_i))$ In the SSL setting, the objective function for the contrastive method is defined as follows:

$$\mathcal{L}_{ssl} = -\frac{1}{2N}\sum_{i\in I}\log\frac{\exp(sim(z_i, z_{j(i)})/\tau)}{\sum_{a\in A(i)}\exp(sim(z_i, z_a)/\tau)} \tag{B.1}$$

Where τ is a temperature hyperparameter, sim is a similarity function - usually, cosine similarity:

$$sim(x, y) = \frac{x \cdot y}{||x||_2||y||_2}$$

$A(i)$ is the set of all other data points in the batch excluding the anchor i: $A(i) = I \setminus i$. In this scenario, index i is called the anchor, $j(i)$ is called the positive sample, and the other $2(N-1)$ indices are called the negatives.

Now, consider the general case of M augmentations applied to the input samples, such that $i \in \{1, 2....M*N\}$. In this case, for each anchor sample there are M positives, $M(N-1)$ negatives. We now define $P_u(i)$ the set indices containing positive samples for index i - naming it the set of unsupervised positives. We can rewrite the loss as in **khoslaSupervisedContrastiveLearning2021**:

$$\mathcal{L}_{ssl} = -\frac{1}{MN}\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\left(\frac{\exp(sim(z_i, z_p)/\tau)}{\sum_{a\in A(i)}\exp(sim(z_i, z_a)/\tau)}\right) \tag{B.2}$$

Supervised contrastive learning

In the supervised setting, consider now a set of labeled data points $(x_i, y_i)_{i\in\{1...N\}}$. we re-use equation B.2 with the notable exception that the set of **supervised** positives $P_s(i)$ and negatives $A_s(i)$ are now defined by the label information in the set of labels y_i:

$$P_s(i) = \{p \in I | y_p = y_i\} \setminus i$$

and $A_s(i) = \bar{P_s(i)} \setminus i$. **khoslaSupervisedContrastiveLearning2021** shows that this formulation is superior to another one that considers the sum of positives within the log on downstream classification performance.

Semi-supervised

let $\mathcal{U}$ be a set of unlabeled data points (musical audio clips), and $\mathcal{S}$ a set of labeled data points belonging to a labeled dataset (pertaining to any task). let $\mathcal{A} = \mathcal{U} \cup \mathcal{B}$ the set of all data points. we can now leverage both labeled and unlabeled data points for semi-supervised contrastive learning. Consider again the general case of M augmentations. we consider a batch B sampled from $\mathcal{A}$ such that B might contain unlabeled and labeled data points. we redefine $P_s(i)$ to allow for the empty set $P(i) = \varnothing$ if i is the index of an unlabeled data point. We keep the defintion of $P_s(i)$ the same with the additional information that given a sample x with class y, then the class of the augmented sample $\tilde{x}$ remains the same.

Briefly, if in the supervised setting a sample had 2 positives in a given batch because of shared class, in the semi-supervised setting it now has $2M$ positives because the label is invariant to augmentations. We now define our semi-supervised contrastive loss, notating $P_A(i) = P_s(i) \cup P_u(i)$:

$$\mathcal{L}_{semi} = -\frac{1}{MN}\sum_{i\in I}\frac{1}{|P_A(i)|}\sum_{p\in P_A(i)}\log\left(\frac{\exp(sim(z_i, z_p)/\tau)}{\sum_{a\in A(i)}\exp(sim(z_i, z_a)/\tau)}\right) \tag{B.3}$$

The modification is minimal, but with the inclusion of both sets of positives, we can leverage both labeled and unlabeled data in our representation learning task. Note that in the case $\mathcal{U} = \varnothing$ or $\mathcal{S} = \varnothing$, the semi-supervised contrastive loss formula degenerates back to repectively the fully-supervised loss (as $P_u(i) = \varnothing$) or the fully self-supervised loss (as $P_s(i) = \varnothing$). This approach contrasts to simply adding the supervised and self-supervised contrastive losses together, as the self-supervised contrastive

objective will consider samples from the same class as negatives while the supervised objective will consider them as positives. Conversely, the samples not having a class would be considered as having no positive samples in the fully supervised case while they would in the self-supervised case). Thus, when separated, these objectives actually work against each other, leading us to consider them together.

Optional batching proportion parameters.

We introduce two hyperparameters p_s the proportion of labeled data vs unlabeled data in $\mathcal{U} \cup \mathcal{S}$:

$$p_s = \frac{|\mathcal{S}|}{|\mathcal{S} \cup \mathcal{U}|}$$

and b_s the proportion of supervised data in a given batch B. If within a given batch B, B_u represents the set of unlabeled data points and B_s the set of labeled data points,

$$b_s = \frac{|B_s|}{|B_s \cup B_u|}$$

Potential of changing the labeled dataset

One advantage of this method is that the labeled dataset used for the supervised portion of the learning task is - quite intuitively - highly influential in the learned representations, as well as the augmentations used on this labeled dataset. In the case of music information retrieval, if we use NSYNTH, a pitch or instrument classification dataset as our labeled dataset, then the learned representations will not only be generally separated based on the semantic information learned in the self-supervised portion, but by virtue of the selected labeled task, instruments or pitches will be separated by design in the latent space.

This holds potential for music representation learning, as leveraging a given dataset can infuse the learned representations with discriminative powers regarding a desired attribute, and can even lead to explainable representations based on which parts of the projected embeddings are used for each part of the task. One of the later studies in this work will be to experiment with the influence of canonical labeled MIR datasets on downstream performance on another task when added to a constant self-supervised dataset.

Human-in-the-loop potential

Furthermore, this method has high potential for human-in-the-loop by design. Consider the following use-case, when such a model is trained on a set of labeled and unlabeled data for a downstream task and now a human user (or an oracle) is assessing the performance of the model on another target classification task. Because contrastive learning operates by similarity, let us imagine a set of prototypes for the downstream task classes, computed by the pretrained model. Inference is run for a given sample by selecting the class prototype with maximum similarity to the input sample. The model gives the estimated class output, and there are three potential cases:

- **The oracle is happy with the prediction.** The sample is now labeled and can go back into the pretraining set for another pretraining iteration when needed - further separating learned representations based on positive feedback.
- **The oracle is unhappy with the prediction, but does not provide a class as part of its feedback**. The data point can now join the set of unlabeled data points, which still provide additional separation. As we know that the data point was closest to the wrong class, further separating it from the prototype will be beneficial in downstream performance.
- **The oracle is unhappy and provides a ground truth**. If the ground truth class already exists in the dataset, then the now labeled data point can be leveraged as part of the supervised dataset to correct the learned representations. If the class is new and unknown, the labeled data point can still be used as part of the unsupervised dataset and at the end of the pretraining iteration, has its own prototype.

In essence, this method can potentially learn for human interactions, but not only positive ones, which is a capacity that not many human-in-the-loop approaches can leverage today. One essential part of experimenting with this capacity will be determining a ROI formula for the oracle and determining whether or not this ROI can evolve positively with a small amount of oracle annotations. Another

will be determining the rectification strategy once the oracle provides annotations for maximum ROI - iterating over the new data points until convergence, only performing one pre-training iteration, etc.

Finally, another strength of this approach is that the human-in-the-loop doesn't only serve to improve the performance on the downstream task. Because of the design of the semi-supervised contrastive task, the representations themselves benefit from the human feedback, which, again, close to no human-in-the-loop approaches are able to boast as a feature today.

Experiments and early results

We evaluate our new semi-supervised contrastive approach on music classification tasks with two architectures, SampleCNN (as is used in Contrastive Learning of Music Representations) [95] and a Tailed UNet (TUNe), as proposed in [104]. We compare our results to CLMR as a baseline on music automatic tagging as reported in the original paper. Further, we explore the usefulness of learned musical representations using only the self-supervised, only the supervised, and the combination of both.

For our first round of experiments, we train a fully-self-supervised baseline (identical to CLMR), a fully supervised contrastive baseline with and without augmentations, and five variants of our semi-supervised approach with different proportions of supervised data used in the global dataset ($p \in [0.01, 0.05, 0.1, 0.2, 0.5]$). For pretraining, 2 augmentations of all samples are produced with the same augmentation chain as [95]. Labels from MagnaTagATune further augment the contrastive matrix $(i, P(i))$ with positives in the case of supervised or semi-supervised pretraining. All our models are pretrained on the train split of the MagnaTagATune dataset for 1000 epochs. For now, we only vary the global proportion of supervised data used, not the in-batch proportion. All models are fine-tuned on the top-50 tags for MagnaTagATune task using an early-stopping mechanism on validation loss, with a learning rate of 0.0003, as in [95]. We use a 1-hidden layer MLP as our classifier on top of frozen representations and report comparable results from literature in table B.1.

				Evaluation - MTAT50	
Ours	**Architecture**	p	**Augmentations**	**AUROC**	**AP**
SSL Contrastive	SampleCNN	0	✓	0.901	0.438
SL Contrastive	SampleCNN	1	✓	0.918	0.465
			✗	**0.925**	**0.499**
Semi-SL Contrastive	SampleCNN	0.01	✓	TBD	TBD
		0.05	✓	TBD	TBD
		0.1	✓	TBD	TBD
		0.2	✓	TBD	TBD
		0.5	✓	0.916	0.4502
Literature					
CLMR [95]	SampleCNN	0	✓	0.893	0.36
TUNe	TUNe [104]	0	✓	0.898	0.371
SOTA	MuLan [153]	-	-	0.920	0.414

Table B.1: Preliminary results for SemiSupCon

Despite the lower number of parameters, when trained for 1000 epochs, the fully supervised contrastive approach beats out even the current state of the art MuLan [153], as well as much larger self-supervised approaches such as MERT [112] and FMX [115] and greatly outperforms the original CLMR implementation. Future experiments will allow for more detailed analysis but a first look at semi-supervised learning with only 50% of the labeled data used to mine positives and negatives is enough to improve upon fully self-supervised training and approach supervised contrastive training with augmentations. *Interestingly, supervised training without augmentations beats out supervised training with augmentation, intuitively this makes sense if the fine-tuning dataset is the same as the pretraining dataset. This will be verified with domain transferability experiments.*

B.3 Planned dissemination

Because of the more desirable features of Semi-Supervised contrastive learning and the already existing approaches using supervised contrastive learning for computer vision, publications for this project make more sense for music or audio oriented conferences and journals. **ISMIR** will be the first priority of SemiSupCon. If the human-in-the-loop experiments are successful, this becomes a novel approach for representation learning using human feedback and would warrant a submission at higher-profile conferences such as **NeurIPS** or **ICLR.**

Appendix C

DMRN18+ poster

RETHINKING MUSIC REPRESENTATION LEARNING FOR MUSIC AND MUSICIANS

Disentangled Human-in-the-loop self-supervised learning for musical audio : towards navigable and interpretable musical representations.

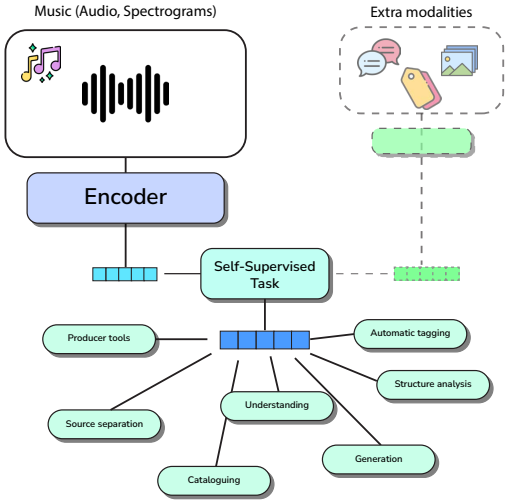
Julien Guinot [1,2] György Fazekas[1] Elio Quinton[2]

1 : Queen Mary University of London 2: Universal Music Group

AI + MUSIC

j.guinot@qmul.ac.uk
Juj_Guinot
julienguinot.com

Problem : Representation learning for music is **not** designed for Music.



A classic representation learning pipeline - the same as general audio.

Some examples of work specialized to music but otherwise :

No taking into account rules / nature of music (hierarchical, sequential, harmonic, structured)

Extra modalities not adapted for high-level musical abstraction (text not from experts)

Representations are not navigable / explainable for humans

No help from expert humans to refine model outputs and representations.

Why?

Inaccessibility of musical experts / expert annotations

Copyright laws

Ontological and structural complexity

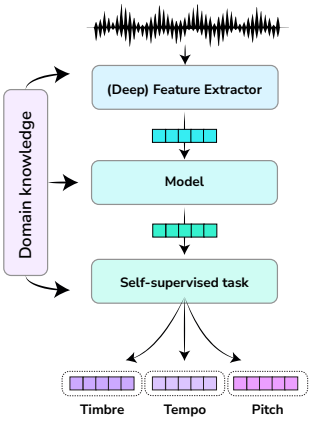
(past) lack of interest

Let's refine self-supervised learning techniques for musical audio, to create interpretable, navigable and explainable learned representations for musicians and musical experts.

How do we do that?

Domain knowledge

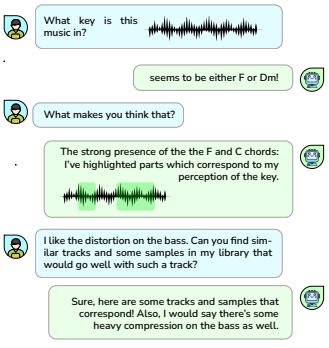
Learn interpretable representations by infusing model and training design with domain knowledge



Extra modalities

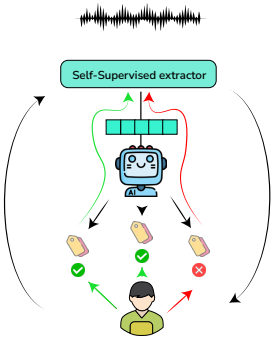
Build musically-relevant latent spaces by baking musical information into outside modalities - for better disentanglement and interactivity.

Improve evaluation for multimoal approaches to identify weaknesses



Human-in-the-loop

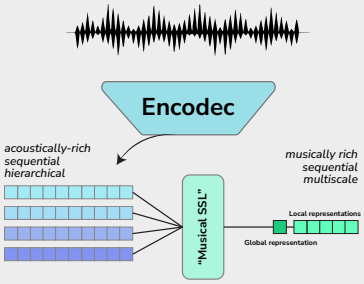
Build models that adapt their outputs and representations with human feedback - to build adaptable, personnalized, and interactive representations.



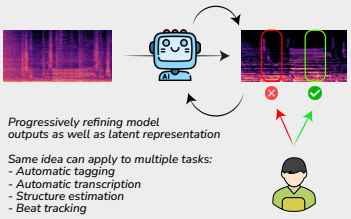
My main question is : can HIL approaches improve the outputs while also improving the representations - in the case of music specifically?

Some ideas

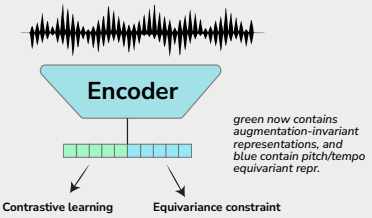
RVQ for repr. learning



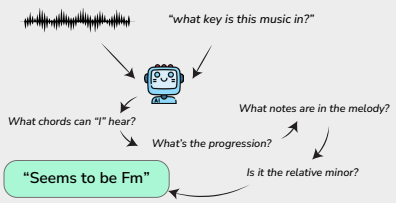
Human-in-the-loop source separation



Disentangled equivariance and invariance



Chain-of-thought for Audio-LLMs



This work was supported by the UKRI and EPSRC under grant EP/S022694/1

UK Research and Innovation

AI + MUSIC

Queen Mary University of London