- In this paper, we investigate the task of musical stem affinity estimation, i.e. from a set of music recordings finding the ones that would match from a musical perspective if mixed together.
- We propose to tackle this problem by training a deep neural network on a multi-track dataset, relying on the assumption that different stems from the same music track do match.
- More precisely, we introduce Stem-JEPA, a novel Joint-Embedding Predictive Architecture (JEPA), in which two neural networks (an encoder and a predictor) are jointly trained to learn semantic representations from an input mix and predict the representations of another stem from the same track but not present in the mix.
- This training strategy enables our model to capture both global and local semantic information from music excerpts and to predict at inference time a musically coherent representation of a given instrument.

- As there is no metric for musical stem affinity estimation, we evaluate our model on a retrieval task and demonstrate its ability to recover the missing stem from a given mix.

We also investigate the semantic content of the learned representations by evaluating them on various MIR tasks, such as key detection, tempo estimation, genre classification and tagging, and reveal that our model captures musical features that would (implicitly or explicitly) be used by a human to determine whether two musical excepts do fit together, such as tempo and key. \TODO{make a good abstract}

In this paper, we explore the task of estimating musical stem accordance, which involves identifying music recordings that would fit each other musically when mixed together.
To do so, we introduce Stem-JEPA, a novel Joint-Embedding Predictive Architecture (JEPA) trained on a multi-track dataset in a self-supervised way.
Our model is composed of two networks (an encoder and a predictor) that are jointly trained to learn semantic representations from an input mix and predict the representations of another stem from the same track but not present in the original mix.
This training strategy enables it to capture both global and local semantic information from music excerpts and to predict at inference time a musically coherent representation of a given instrument.

We validate the capabilities of our model by measuring its performances on a retrieval task using the MUSDB18 dataset, and demonstrate its ability to recover the missing stem from a given mix. We also show that the embeddings capture beat information and present some failure cases of our model. Finally, we evaluate the representations learned by of our model on several downstream tasks, highlighting the fact that they effectively capture meaningful musical features.