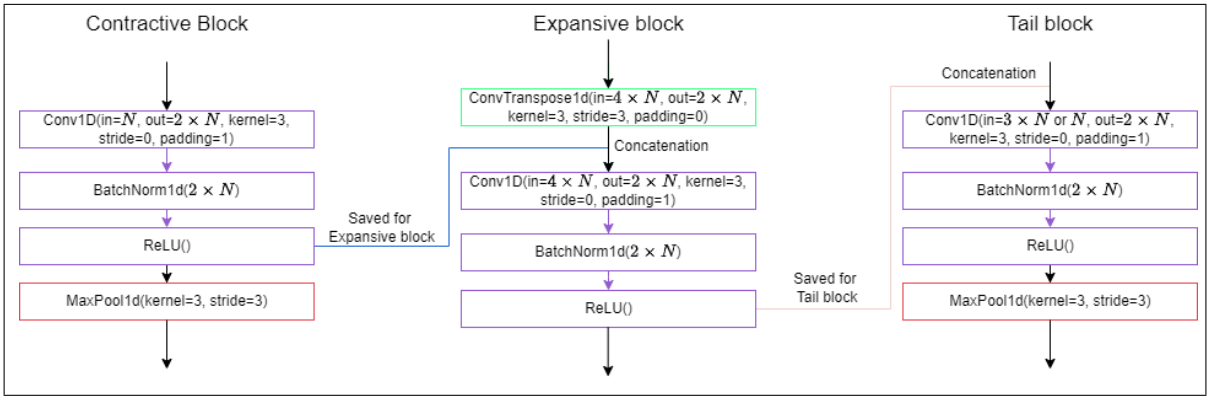

Figure S1: A schematic of what the three block types used in a TUNe architecture are made up of and how they are connected. This example schematic the dimension ratios for blocks at the same depth of the network.

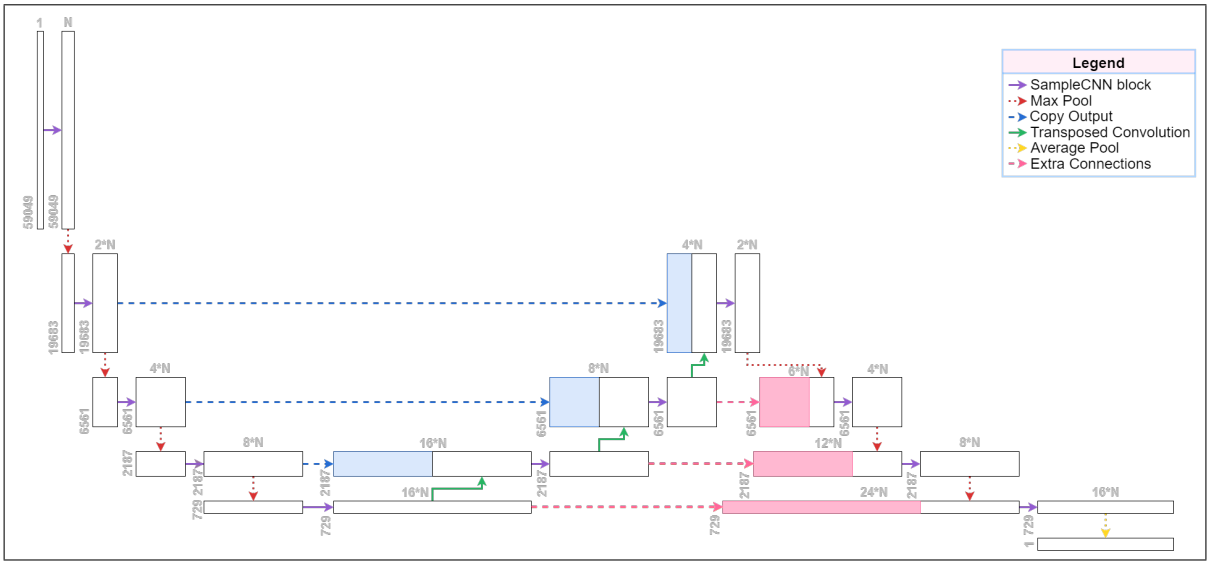

Figure S2: An example of path-length modification in TUNe. This network is an Expansive–1 network, because it is missing one layer of the expansive path (and consequently also of the tail) as compared to vanilla TUNe in Figure 1.

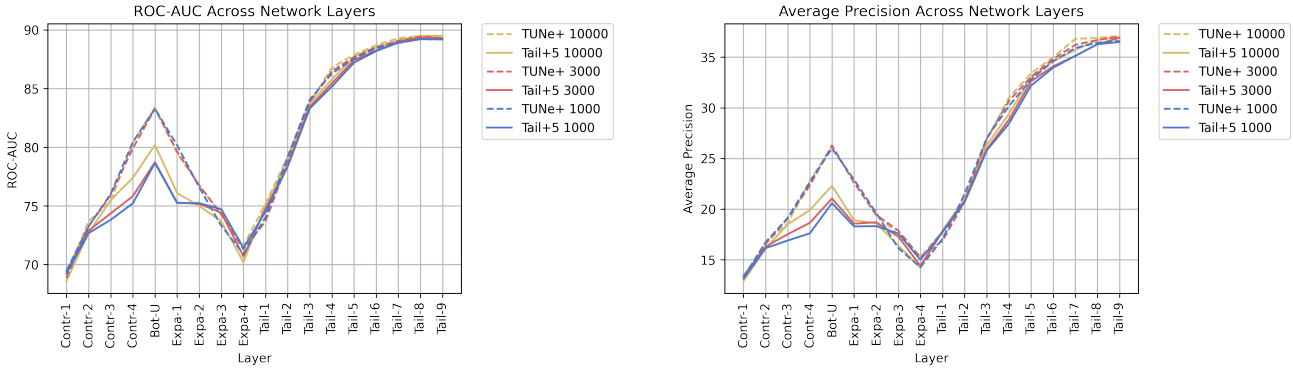

Figure S3: Two figures displaying the two evaluation metrics for probing every layer in the TUNe Tail+5 and TUNe+ models trained for 1 000, 3 000, 10 000 epochs. Along the x-axis' the name of the layers are displayed. The y-axis' display the MTT_{AUC} and MTT_{AP} performance when the output of said layer, average pooled over the time dimension, is probed on MTT.

Variant	Epochs trained	MTT_{AUC}	MTT_{AP}
CLMR [43]	1000	88.3	34.4
TUNe Tail+5	1000	89.2	36.5
TUNe+	1000	89.2	36.6
CLMR [43]	3000	88.5 (88.9)	35.1 (35.5)
TUNe Tail+5	3000	89.3 (89.6)	36.9 (36.7)
TUNe+	3000	89.3 (89.5)	37.0 (36.6)
CLMR [3]	10000	88.7 (89.3)	35.6 (36.0)
TUNe Tail+5	10000	89.5 (89.6)	37.0 (36.7)
TUNe+	10000	89.3 (89.8)	37.1 (37.1)

Table S1: CLMR, TUNe Tail+5, and TUNe+ performances at different amount of epochs trained, evaluated on the Magnatagatune tagging task.

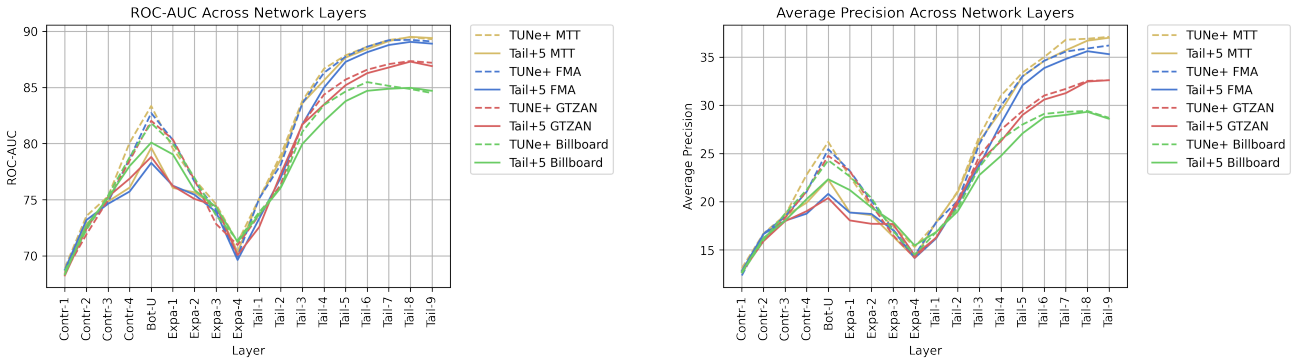

Figure S4: Two plots displaying the probing performance of every layer in the TUNe Tail+5 and TUNe+ models trained on MTT, FMA, GTZAN, and McGill Billboard. Along the x-axis' the name of the layers are displayed. The y-axis' display the MTT_{AUC} and MTT_{AP} performance when the output of said layer, average pooled over the time dimension, is probed.

Probing Approach	MTT_{AUC}	MTT_{AP}	GTZAN	GS	Emo_{A}	Emo_{V}
CLMR [3]†	89.4	36.1	68.6	14.9	67.8	45.8
Jukebox [29]†	91.5	41.4	79.7	66.7	72.1	61.7
TUNe	90.3	38.1	67.6	13.7	60.5	55.7
TUNe+	90.3	38.0	64.5	15.5	64.7	45.9
State-of-the-art [2, 12, 27, 44–46] †	92.0	38.4	82.1	79.6	70.4	55.6
Pre-trained [3, 12, 27, 45, 45, 47] †	92.0	35.9	82.1	75.8	67.1	55.6
From scratch [2, 2, 44, 44, 48, 49] †	90.7	38.4	65.8	74.3	70.4	50.0

Table S2: Probing experiment as compared to results from [14]. To see how well the best-performing model trained on their methods main dataset, in our case MTT, generalises to four different probing tasks: music tagging of the MTT dataset [34], genre classification of the fault-filtered GTZAN dataset [40, 41, 50], key detection in the Giant Steps dataset [51], and emotion recognition in the Emomusic dataset [52] (which is a subset of the FMA dataset). For each of these experiments a grid search was done for the optimal probe parameters, using the same settings as [14].

Variant	Forward/Backward pass (MB)	Weights (MB)	Batch of 96 size (GB)	Epochs/ hour	1 000 epochs (h:m)
Vanilla TUNe	466.19	9.03	44.55	17.94	56:45
TUNe Contractive+1	264.98	8.71	25.66	18.15	55:10
TUNe Contractive+2	142.46	8.18	14.12	17.95	55:45
TUNe Contractive+3	70.91	6.32	7.24	18.15	55:10
TUNe Expansive-1	343.65	8.95	33.06	17.82	56:10
TUNe Expansive-2	269.32	9.08	26.1	16.39	61:00
TUNe Expansive-3	230.56	8.94	22.45	17.64	56:45
TUNe Tail+1	264.98	8.86	37.31	17.68	56:40
TUNe Tail+2	155.62	8.62	25.9	17.75	56:25
TUNe Tail+3	155.62	8.65	20.69	17.55	57:00
TUNe Tail+4	155.62	8.83	18.07	18.06	55.25
TUNe Tail+5	155.62	8.1	15.35	17.55	57:00
TUNe CLMR-tail	253.05	9.72	24.63	17.96	55:45
TUNe+	155.62	8.57	15.39	17.48	57:15
Vanilla TUNe Small	158.53	1.39	14.99	17.71	56:30
TUNe+ Large	473.28	28.35	47.03	18.25	54:50
TUNe+ Smaller Rep	156.38	5.48	15.17	17.11	58:30
CLMR	164.47	9.13	16.27	17.47	57:20

Table S3: Each variants Forward/backward pass size (MB), parameter size, memory needed for a batch size of 96, how many epochs per hour, and how long it takes to train 1000 epochs. These statistics are gathered either by the 'torchsummary' module, see 'model_overview.py' in the repository, or the Wandb logs.