Global_idx	Year	Title	Link	GPU*	Training time	Company connection	Total Compute Info (Train)	TDP/W	Number of GPUs	Training time	Energy cost	
Unique ID for the article	Publication year	Title of the ISMIR article	Paper Link	References GPU/TPU use (citation)	Total computing time for model training (hours)	Did the authors indicate an affiliation with a company (if yes, which)	Full/Partial/None. Partial will refer to cases where number of GPU is stated.	The number of GPUs used in the model training	The number of GPUs used in the model training	Training time (hours)	The total energy cost of training (kWh)	Red columns used for energy computations in paper
81	2023	TriAD: Capturing Harmonics With 3D Convolutions	https://archives.ismir.net/ismir2023/paper/000002.pdf	All the harmonic blocks take a similar time to train, around 24h to complete in a V100 GPU.	3*2*24 hours	Yes: Huawei, Munich Research Center	Full	250	1	144	36.00	
82	2023	Impact of time and note duration tokenizations on deep learning symbolic music modeling	https://archives.ismir.net/ismir2023/paper/000009.pdf	All trainings are performed on V100 GPUs		No	Partial	N/A	N/A	N/A	N/A	
83	2023	IteraTTA: An interface for exploring both text prompts and audio priors in generating music with text-to-audio models	https://archives.ismir.net/ismir2023/paper/000014.pdf	(Inference:) The length of music audios to generate was predetermined at 10 seconds so that our GPU server harnessing an NVIDIA RTX 2080 Ti can afford the generation of 12 audios (3 audios × 4 prompts) simultaneously.		No	None	N/A	N/A	N/A	N/A	
84	2023	Efficient Notation Assembly in Optical Music Recognition	https://archives.ismir.net/ismir2023/paper/000020.pdf	The experiment was run over 8 cores of i7-7700K CPU at 4.20 GHz with 16 GB of RAM memory, with no explicit parallelization or GPU speed-up.		No	Partial	N/A	N/A	N/A	N/A	
85	2023	Transcription with Hierarchical Frequency-Time Transformer	https://archives.ismir.net/ismir2023/paper/000024.pdf	We trained our models for 50 epochs on MAPS dataset and 20 epochs for MAESTRO dataset using one NVIDIA A100 GPU. It took roughly 140 minutes and 43.5 hours to train one epoch with our model for MAPS and MAESTRO, respectively.	50*140 min + 20*43.5*60 min	Yes: Sony Group Corporation and Sony Computer Science Laboratories	Full	300	1	987	296.10	
86	2023	On the Performance of Optical Music Recognition in the Absence of Specific Training Data	https://archives.ismir.net/ismir2023/paper/000037.pdf	Finally, all experiments were run using the Python language (v. 3.8.13) with the PyTorch framework (v. 1.13.0) on a single NVIDIA GeForce RTX 4090 card with 24GB of GPU memory.	No	No	Partial	N/A	N/A	N/A	N/A	
87	2023	Zero-shot Lyrics Transcription by Whispering to ChatGPT	https://archives.ismir.net/ismir2023/paper/000040.pdf	We conducted our experiments concurrently on a server with 8xA100 80G GPUs. It takes approximately 9 hours to complete one round of inference, and each process uses up to 12G VRAM.	No	No	Partial (inference)	N/A	N/A	N/A	N/A	
88	2023	Predicting Music Hierarchies with a Graph-Based Neural Decoder	https://archives.ismir.net/ismir2023/paper/000050.pdf	The training time is roughly the same, around 1 hour on a GPU RTX 1080	Assumably 2 hours	No	Full	170	1	2	0.34	
89	2023	On the Effectiveness of Speech Self-supervised Learning for Music	https://archives.ismir.net/ismir2023/paper/000054.pdf	All the MusicHu-BERT and Music2Vec models are trained for 400k steps with 8 × NVIDIA A100-40GB GPUs. Training with 8 GPUs takes around 2~3 days.	Yes: average of 48h and 72h	No	Full	300	8	60	144.00	

90	2023	PESTO: Pitch Estimation with Self-supervised Transposition-equivariant Objective	https://archives.ismir.net/ismir2023/paper/000063.pdf	Our architecture being extremely lightweight, training requires only 545MB of GPU memory and can be performed on a single GTX 1080Ti.	No	Yes: LTCI/Télécom-Paris, Sony Computer Science Laboratories, Sony AI	Partial	N/A	N/A	N/A	N/A
91	2023	Audio Embeddings as Teachers for Music Classification	https://archives.ismir.net/ismir2023/paper/000068.pdf	(Inference:) GPU: NVIDIA 2070 Super	No	No	Partial (inference)	N/A	N/A	N/A	N/A
92	2023	Towards Improving Harmonic Sensitivity and Prediction Sta-bility for Singing Melody Extraction	https://archives.ismir.net/ismir2023/paper/000078.pdf	All models are trained and tested in NVIDIA RTX 2080Ti GPUs and implemented in PyTorch.	No	No	Partial	N/A	N/A	N/A	N/A
93	2023	Efficient Supervised Training of Audio Transformers for Music Representation Learning	https://archives.ismir.net/ismir2023/paper/000098.pdf	We trained MAEST using 4 Nvidia 2080 RTX Ti GPUs with 12GB of RAM. The training takes 31 hours for MAEST-5 and 48 hours for MAEST-30.	Yes: 31h+48h	No	Full	250	4	79	79.00
94	2023	Music Source Separation with MLP Mixing of Time, Frequency, and Channel	https://archives.ismir.net/ismir2023/paper/000100.pdf	The training was distributed across multiple GPUs, with a batch size of 4 on each GPU. + see Table 3	No	No	None	N/A	N/A	N/A	N/A
95	2023	Symbolic Music Representations for Classification Tasks: A Systematic Evaluation	https://archives.ismir.net/ismir2023/paper/000101.pdf	All our experiments are trained on a single A5000 GPU.	No	No	Partial	N/A	N/A	N/A	N/A
96	2022	Attention-Based Audio Embeddings for Query-by-Example	https://archives.ismir.net/ismir2022/paper/000005.pdf	The model was trained on a single NVIDIA Tesla V100 GPU for about 40 hours	Yes: 40h	No	Full	250	1	40	10.00
97	2022	Beat Transformer: Demixed Beat and Downbeat Tracking with Dilated Self-Attention	https://archives.ismir.net/ismir2022/paper/000019.pdf	Our model has 9.29M trainable parameters and is trained with an RTX-A5000-24GB GPU. Each training fold generally takes 20 epochs (in 11 hours) to fully converge. (AH: they have 8 folds)	Yes: 88h	No	Full	230	1	88	20.24
98	2022	Mel Spectrogram Inversion with Stable Pitch	https://archives.ismir.net/ismir2022/paper/000027.pdf	2 Volta GPUs, on which training took “approximatly 2 days”	48h	Yes: Apple	Full	250	2	48	24.00
99	2022	Latent feature augmentation for chorus detection	https://archives.ismir.net/ismir2022/paper/000028.pdf	Training “run at a Tesla-V100-SXM2-32GB GPU”, but time given in epochs only (100)	No	Yes: ByteDance	Partial	N/A	N/A	N/A	N/A
100	2022	Supervised and Unsupervised Learning of Audio Representations for Music Understanding	https://archives.ismir.net/ismir2022/paper/000030.pdf	Supervised models were trained on 8 v100 GPUs taking approximately 30 hours, while unsupervised models were trained on 16 A100 GPUs taking approximately 80 hours.	"Supervised models were trained on 8 v100 GPUs taking approximately 30 hours, while unsupervised models were trained on 16 A100 GPUs taking approximately 80 hours."	Yes: SiriusXM	Full	250;300	8;16	30;80	444.00
101	2022	Performance MIDI-to-score conversion by neural beat tracking	https://archives.ismir.net/ismir2022/paper/000047.pdf	4 GPUs used for training	No	Yes: ByteDance (China)	Partial	N/A	N/A	N/A	N/A
102	2022	Melody transcription via generative pre-training	https://archives.ismir.net/ismir2022/paper/000058.pdf	All models converge within 15k steps or about a day on a single K40 GPU.	Yes: 3 * 24h	No	Full	235	1	72	16.92
103	2022	Source Separation of Piano Concertos with Test-Time Adaptation	https://archives.ismir.net/ismir2022/paper/000059.pdf	"Train all our models on a single NVIDIA GeForce 1080 Ti GPU"	No	No	Partial	N/A	N/A	N/A	N/A
104	2022	Checklist Models for Improved Output Fluency in Piano Fingering Prediction	https://archives.ismir.net/ismir2022/paper/000063.pdf	Model "trains on an NVIDIA 2080ti GPU in roughly 12 hours".	Yes	No	Full	250	1	12	3.00
105	2022	Towards robust music source separation on loud commercial music	https://archives.ismir.net/ismir2022/paper/000069.pdf	No info on how many used	No	No	None	N/A	N/A	N/A	N/A

106	2022	EnsembleSet: a new high quality synthesised dataset for chamber ensemble separation	https://archives.ismir.net/ismir2022/paper/000075.pdf	4 x NVIDIA A100 GPUs	"Each epoch in our experiments took 40 minutes" and "We train the models for 100 epochs"	No	Full	300	4	200	240.00
107	2022	Contrastive Audio-Language Learning for Music	https://archives.ismir.net/ismir2022/paper/000077.pdf	No info: "To save GPU memory, we perform training with automatic mixed precision."	No	Yes: Universal Music Group	None	N/A	N/A	N/A	N/A
108	2022	A diffusion-inspired training strategy for singing voice extraction in the waveform domain	https://archives.ismir.net/ismir2022/paper/000082.pdf	No info: "Both models predict faster than real-time on a TITAN Xp GPU"	No	No	None	N/A	N/A	N/A	N/A
109	2022	HPPNet: Modeling the Harmonic Structure and Pitch Invariance in Piano Transcription	https://archives.ismir.net/ismir2022/paper/000085.pdf	Training time on an NVIDIA GeForce 3060 GPU with 12 GB VRAM is about 48 hours.	Yes	No	Full	170	1	48	8.16
110	2022	Improving Choral Music Separation through Expressive Synthesized Data from Sampled Instruments	https://archives.ismir.net/ismir2022/paper/000087.pdf	Type info & not single GPU: "We implemented all methods in Pytorch using NVIDIA RTX 2080Ti GPUs"	No: "All models converged within 300 epochs"	Yes: Tencent AI lab (China)	Partial	N/A	N/A	N/A	N/A
111	2022	A Transformer-Based Spellchecker for Detecting Errors in OMR Output	https://archives.ismir.net/ismir2022/paper/000095.pdf	Number: "using four NVIDIA P100-PCIE GPUs with a combined 48 GB of memory"	No	No	Partial	N/A	N/A	N/A	N/A
112	2022	Music Representation Learning Based on Editorial Metadata From Discogs	https://archives.ismir.net/ismir2022/paper/000099.pdf	Number & Type (pre-training): "three machines with two GeForce RTX 2080 Ti GPUs each"	No: "trained the models for 100 epochs"(not clear how many and how long in hours)	No	Partial	N/A	N/A	N/A	N/A
113	2022	Transfer Learning of wav2vec 2.0 for Automatic Lyric Transcription	https://archives.ismir.net/ismir2022/paper/000107.pdf	"4 RTX A5000 GPUs" (for training, also some inference resources are provided)	No	No	Partial	N/A	N/A	N/A	N/A