Table 2: Summary of the 19 evaluation tasks of HEAR. Includes the embedding type (timestamp (T) or scene (S)), the predictor type (multiclass (C) or multilabel (L)), the split method used during downstream evaluation (train/validation/test (TVT) or K-fold), the duration of clips in seconds, the total number of clips for each task, the primary evaluation metric, and whether or not the task is novel. Novel tasks are not comparable to the literature. For all tasks except FSD50k, clips were standardized to one length using padding or trimming, typically the 95th percentile length in the original corpus.

Task Name	Embed Type	Predictor Type	Split Method	Duration (seconds)	# clips	Evaluation Metric	Novel
Open Tasks							
DCASE 2016 Task 2	T	L	TVT	120.0	72	Onset FMS	✓
NSynth Pitch 5hr	S	C	TVT	4.0	5000	Pitch Acc.	✓
NSynth Pitch 50hr	S	C	TVT	4.0	49060	Pitch Acc.	✓
Speech Commands 5hr	S	C	TVT	1.0	22890	Accuracy	✓
Speech Commands Full	S	C	TVT	1.0	100503	Accuracy	
Secret Tasks							
Beehive States	S	C	TVT	600.0	576	AUCROC	
Beijing Opera Percussion	S	C	5-fold	4.77	236	Accuracy	✓
CREMA-D	S	C	5-fold	5.0	7438	Accuracy	
ESC-50	S	C	5-fold	5.0	2000	Accuracy	
FSD50K	S	L	TVT	0.3 - 30.0	51185	mAP	
Gunshot Triangulation	S	C	7-fold	1.5	88	Accuracy	✓
GTZAN Genre	S	C	10-fold	30.0	1000	Accuracy	
GTZAN Music Speech	S	C	10-fold	30.0	128	Accuracy	
LibriCount	S	C	5-fold	5.0	5720	Accuracy	
MAESTRO 5hr	T	L	5-fold	120.0	185	Onset FMS	✓
Mridangam Stroke	S	C	5-fold	0.81	6977	Accuracy	✓
Mridangam Tonic	S	C	5-fold	0.81	6977	Accuracy	✓
Vocal Imitations	S	C	3-fold	11.26	5601	mAP	✓
VoxLingua107 Top10	S	C	5-fold	18.64	972	Accuracy	✓

Appendix A. Evaluation Tasks

Our 19 tasks were derived from 16 datasets, as described in more detail below. Tasks described as "novel" are not comparable to the literature. A summary of task statistics is available in Table 2.

Speech Commands (version 2), 5h and full Classification of known spoken commands, with additional categories for silence and unknown commands (Warden, 2018). As per the literature, models are evaluated by prediction accuracy. We also provide a 5-hour subset of the training data. We use the predefined train and test split, and note that the test data has a different distribution of labels from the training data.

NSynth Pitch, 5h and 50h NSynth Pitch is a novel multiclass classification problem. The goal of this task is to classify instrumental sounds from the NSynth Dataset (Engel et al., 2017) into one of 88 pitches. Results for this task are measured by pitch accuracy, as well as chroma accuracy. Chroma accuracy considers only the pitch "class" i.e., pitches that are a multiple-of-octaves apart are considered equivalent. For HEAR we created two versions of this dataset: a 5 hour and 50 hour version. Unlike Carr et al. (2021), we treat this as a classification, not regression problem.

DCASE 2016 Task 2 A novel office sound event detection in synthesized scenes, adapted from DCASE 2016 Task 2 (Mesaros et al., 2018). Novel, insofar as our evaluation uses different splits. The original imbalanced splits did not work well our generic cross-validation.

Postprocessing: Predictions were postprocessed using 250 ms median filtering. At each validation step, a minimum event duration of 125 or 250 ms was chosen to maximize onset-only event-based F-measure (with 200 ms tolerance). Scores were computed using `sed_eval` (Mesaros et al., 2016).

Beehive States This is a binary classification task using audio recordings of two beehives (Nolasco et al., 2019). The beehives are in one of two states: a normal state, and one in which the queen bee is missing ("queen-less"). At 10 minutes long, this task has the longest audio clips in HEAR.

Beijing Opera Percussion This is a novel audio classification task developed using the Beijing Opera Percussion Instrument Dataset (Tian et al., 2014). The Beijing Opera uses six main percussion instruments that can be classified into four main categories: Bangu, Naobo, Daluo, and Xiaoluo.

CREMA-D CREMA-D is a dataset for emotion recognition (Cao et al., 2014). The original dataset contains audiovisual data of actors reciting sentences with one of six different emotions (anger, disgust, fear, happy, neutral and sad). For HEAR, we only use the audio recordings (which differs from much but not all of the literature).

ESC-50 This is a multiclass classification task on environmental sounds. The ESC-50 dataset is a collection of 2000 environmental sounds organized into 50 classes (Piczak, 2015). Scores are averaged over 5 folds. (The folds are predefined in the original dataset.)

FSD50K FSD50K is a multilabel task (Fonseca et al., 2020). This dataset contains over 100 hours of human-labeled sound events from Freesound (https://freesound.org/). Each of the ≈51 k audio clips is labeled using one or more of 200 classes from the AudioSet Ontology, encompassing environmental sounds, speech, and music. Unlike the other datasets, for FSD50K scene embeddings we did not alter the audio clip length. Each clip is between 0.3 and 30 seconds long. We use the predefined train/val/eval split. Evaluation is done using mean average precision (mAP).

Gunshot Triangulation Gunshot triangulation is a novel resource multiclass classification task that utilizes a unique dataset: gunshots recorded in an open field using iPod Touch devices (Cooper and Shaw, 2020). This data consist of 22 shots from 7 different firearms, for a total of 88 audio clips, the smallest dataset in HEAR. Each shot is recorded using four different iPod Touches, located at different distances from the shooter. The goal of this task is to classify audio by the iPod Touch that recorded it, i.e., to identify the location of the microphone. The dataset was split into 7 different folds, where each firearm belonged to only one fold. Results are averaged over each fold.

GTZAN Genre The GTZAN Genre Collection (Tzanetakis and Cook, 2002) is a dataset of 1000 audio tracks (each 30 seconds in duration) that are categorized into ten genres (100 tracks per genre). The task is multiclass classification. As per the literature, scores are averaged over 10 folds. However, we don't used the corrected artist-conditional splits from (Sturm, 2013).

GTZAN Music Speech GTZAN Music Speech is a binary classification task, where the goal is to distinguish between music and speech. The dataset consists of 120 tracks (each 30 seconds in duration) and each class (music/speech) has 60 examples.

LibriCount LibriCount is a multiclass speaker count identification task (Stöter et al., 2018b). The dataset contains audio of a simulated cocktail party environment with between 0 to 10 speakers. The goal of this task is to classify how many speakers are present in each of the recordings. Following Stöter et al. (2018a), we treat this as a classification, not regression, problem.

MAESTRO 5h This is a novel music transcription task adapted from MAESTRO. For HEAR, we created a subsampled version that includes 5 hours of training and validation audio, in 120 second clips. To evaluate submissions, a shallow transcription model was trained on timestamp-based embeddings provided by the participant models.

We use note onset FMS and note onset with offset FMS for evaluation, as per the original MAESTRO paper (Hawthorne et al., 2019) and the preceding Onsets and Frames paper (Hawthorne et al., 2018).

Note onset measures the ability of the model to estimate note onsets with 50 ms tolerance and ignores offsets. Note onset w/ offset includes onsets as well as requires note duration within 20% of ground truth or within 50 ms, whichever is greater.

Mridingham Stroke and Mridingham Tonic We used the Mridangam Stroke Dataset (Anantapadmanabhan et al., 2013) for two novel multiclass classification tasks: Stroke classification and Tonic classification. The Mridingam is a pitched percussion instrument used in carnatic music, which is a sub-genre of Indian classical music. This dataset comprises 10 different strokes played on Mridingams with 6 different tonics.

Vocal Imitations Vocal Imitations (Kim et al., 2018a) is a novel multiclass classification task, where the goal is to match a vocal imitation of a sound with the sound that is being imitated. The dataset contains 5601 vocal imitations of 302 reference sounds, organized by AudioSet ontology. Given a vocal sound, the classification task is to retrieve the original audio it is imitating.

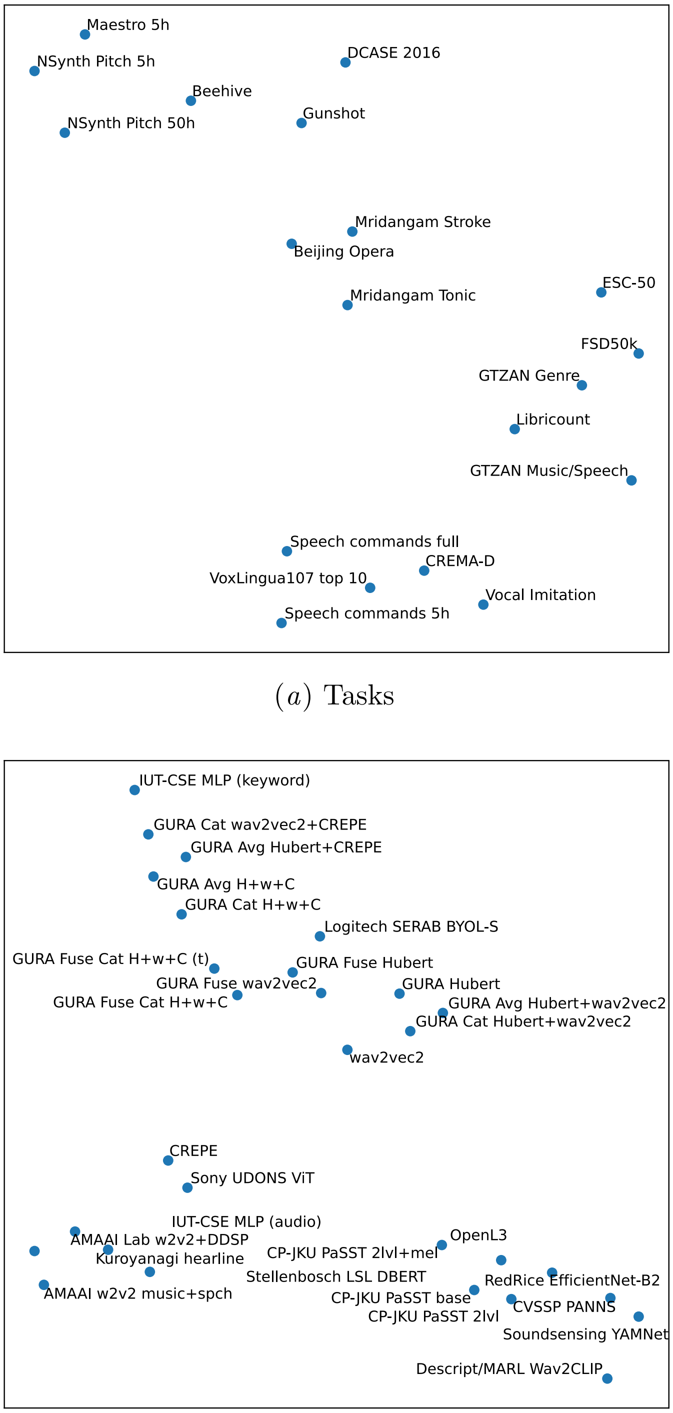
VoxLingua107 Top 10 This is a novel multiclass classification task derived from the VoxLingua107 dataset (Valk and Alumäe, 2021). The goal of the task is to identify the spoken language in an audio file. For HEAR we selected the top 10 most frequent languages from the development set, which resulted in just over 5 hours of audio over 972 audio clips.

Appendix B. Downstream training details

For each task, using a given model's frozen embeddings as input features, we train a downstream MLP classifier. For scene-based multiclass tasks, the final layer is a softmax with

Table 3: Properties of baseline and submitted models, including: whether the model processes raw audio (1-D) or spectrograms (2D); on what kind of data the model is pretrained; the number of million parameters; the size of the output embedding for scene and timestamp tasks; and the number of minutes the model spends embedding Speech Commands V2. We caution that embedding time is not the entire picture, if participants did not do simple speed optimizations. For example, the CREPE wrapper (also used by GURA) is known not to exploit GPU batch parallelism.

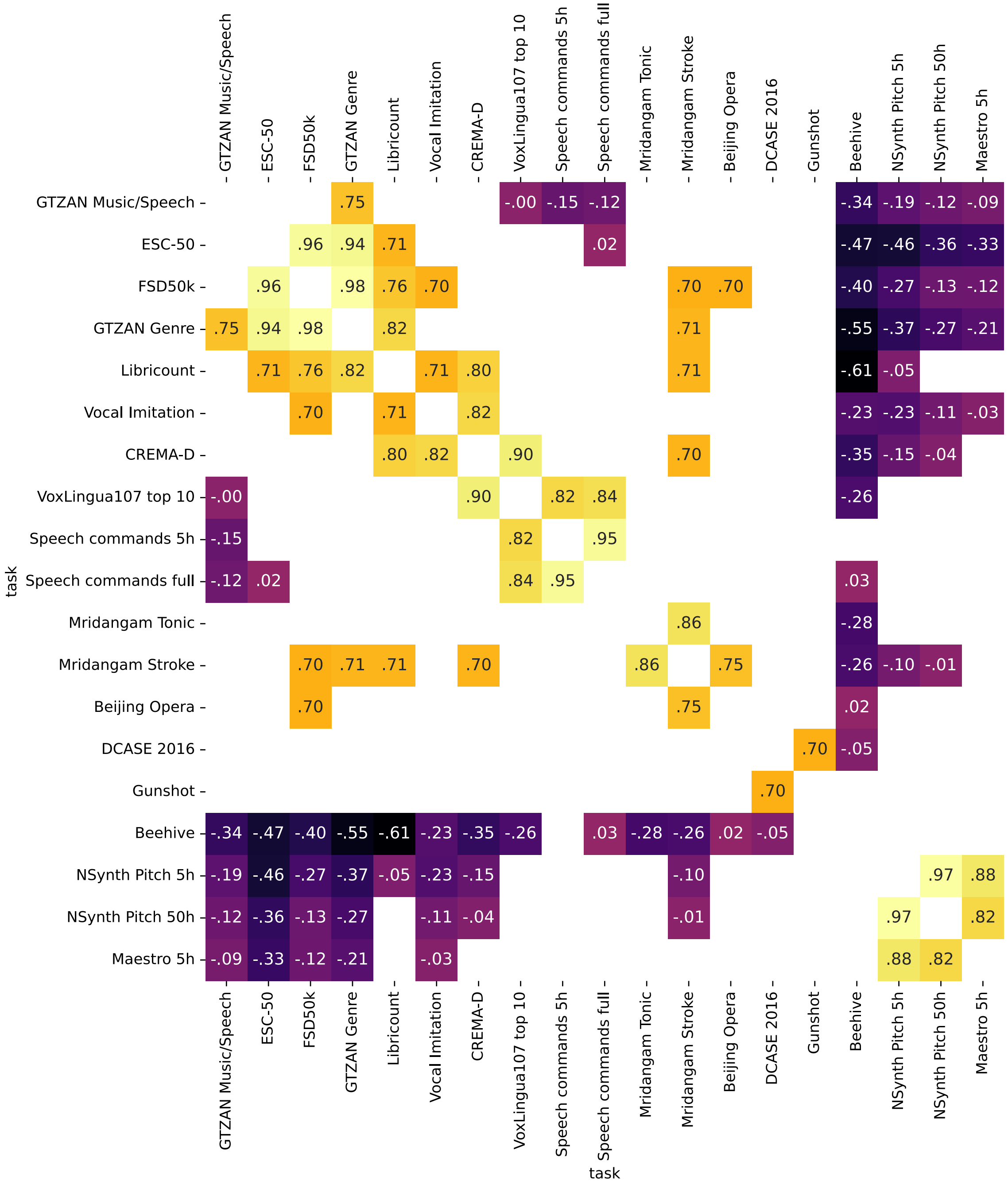
	Input		**Pretraining data**			**# M**	**Embed dim**		**Time**
Model	**1D**	**2D**	**speech**	**broad**	**music**	**params**	**scene**	**time**	**min**
OpenL3		✓		✓		4.7	512	512	94.9
wav2vec2	✓		✓			315.4	1024	1024	8.9
CREPE	✓				✓	22.2	2048	2048	38.3
AMAAI Lab wav2vec2+DDSP	✓		✓		✓	98.8	871	871	43.6
AMAAI wav2vec2 music+speech	✓		✓		✓	300.0	768	768	5.0
CP-JKU PaSST 2lvl		✓		✓		86.2	1295	2590	14.5
CP-JKU PaSST 2lvl+mel		✓		✓		86.2	1295	3358	5.8
CP-JKU PaSST base		✓		✓		86.2	1295	1295	5.8
CVSSP PANNS		✓		✓		80.8	2048	2048	3.9
Descript/MARL Wav2CLIP		✓		✓		11.7	512	512	3.1
GURA Avg H+w+C	✓		✓		✓	1339.0	1024	1024	40.0
GURA Avg Hubert+Crepe	✓		✓		✓	1022.0	1024	1024	33.9
GURA Avg Hubert+wav2vec2	✓		✓			634.0	1024	1024	14.6
GURA Cat H+w+C	✓		✓		✓	1339.0	3072	3072	40.1
GURA Cat Hubert+wav2vec2	✓		✓			634.0	2048	2048	14.4
GURA Cat wav2vec2+crepe	✓		✓		✓	339.0	2048	2048	24.7
GURA Fuse Cat H+w+C	✓		✓		✓	1339.0	3072	3072	40.1
GURA Fuse Cat H+w+C (time)	✓	✓			✓	1339.0	15360	3072	34.6
GURA Fuse Hubert	✓		✓			1000.0	1280	1280	18.1
GURA Fuse wav2vec2	✓		✓			317.0	1024	1024	8.8
GURA Hubert	✓		✓			1000.0	1280	1280	17.9
IUT-CSE MLP (audio)		✓	✓	✓	✓	0.2	1584	8	2.9
IUT-CSE MLP (keyword)		✓	✓			0.4	1024	64	3.0
Logitech AI SERAB BYOL-S		✓	✓			5.3	2048	2048	4.8
RedRice EfficientNet-B2		✓		✓		7.7	1408	1408	3.4
Sony UDONS ViT		✓	✓			11.1	768	768	3.5
Soundsensing YAMNet		✓		✓		3.8	1024	1024	15.7
Stellenbosch LSL DBERT	✓		✓			316.8	2048	2048	6.5



(*a*) Tasks

(*b*) Models

Figure 2: t-SNE visualizations of tasks and models, based upon normalized scores. Missing normalized scores were imputed using sklearn's multivariate IterativeImputer.



	GTZAN Music/Speech	ESC-50	FSD50k	GTZAN Genre	Libricount	Vocal Imitation	CREMA-D	VoxLingua107 top 10	Speech commands 5h	Speech commands full	Mridangam Tonic	Mridangam Stroke	Beijing Opera	DCASE 2016	Gunshot	Beehive	NSynth Pitch 5h	NSynth Pitch 50h	Maestro 5h
GTZAN Music/Speech				.75				-.00	-.15	-.12						-.34	-.19	-.12	-.09
ESC-50			.96	.94	.71					.02						-.47	-.46	-.36	-.33
FSD50k		.96		.98	.76	.70						.70	.70			-.40	-.27	-.13	-.12
GTZAN Genre	.75	.94	.98		.82							.71				-.55	-.37	-.27	-.21
Libricount		.71	.76	.82		.71	.80					.71				-.61	-.05		
Vocal Imitation			.70		.71		.82									-.23	-.23	-.11	-.03
CREMA-D					.80	.82		.90				.70				-.35	-.15	-.04	
VoxLingua107 top 10	-.00						.90		.82	.84						-.26			
Speech commands 5h	-.15							.82		.95									
Speech commands full	-.12	.02						.84	.95							.03			
Mridangam Tonic												.86				-.28			
Mridangam Stroke			.70	.71	.71		.70				.86		.75			-.26	-.10	-.01	
Beijing Opera			.70									.75				.02			
DCASE 2016															.70	-.05			
Gunshot														.70					
Beehive	-.34	-.47	-.40	-.55	-.61	-.23	-.35	-.26		.03	-.28	-.26	.02	-.05					
NSynth Pitch 5h	-.19	-.46	-.27	-.37	-.05	-.23	-.15					-.10						.97	.88
NSynth Pitch 50h	-.12	-.36	-.13	-.27		-.11	-.04					-.01					.97		.82
Maestro 5h	-.09	-.33	-.12	-.21		-.03											.88	.82	

Figure 3: Task versus task correlation scores, based upon normalized scores. Only the highest and lowest correlations are displayed. Cells are sorted to minimize the traveling salesperson distance, mapping correlations [-1, +1] to distances [+2, 0].

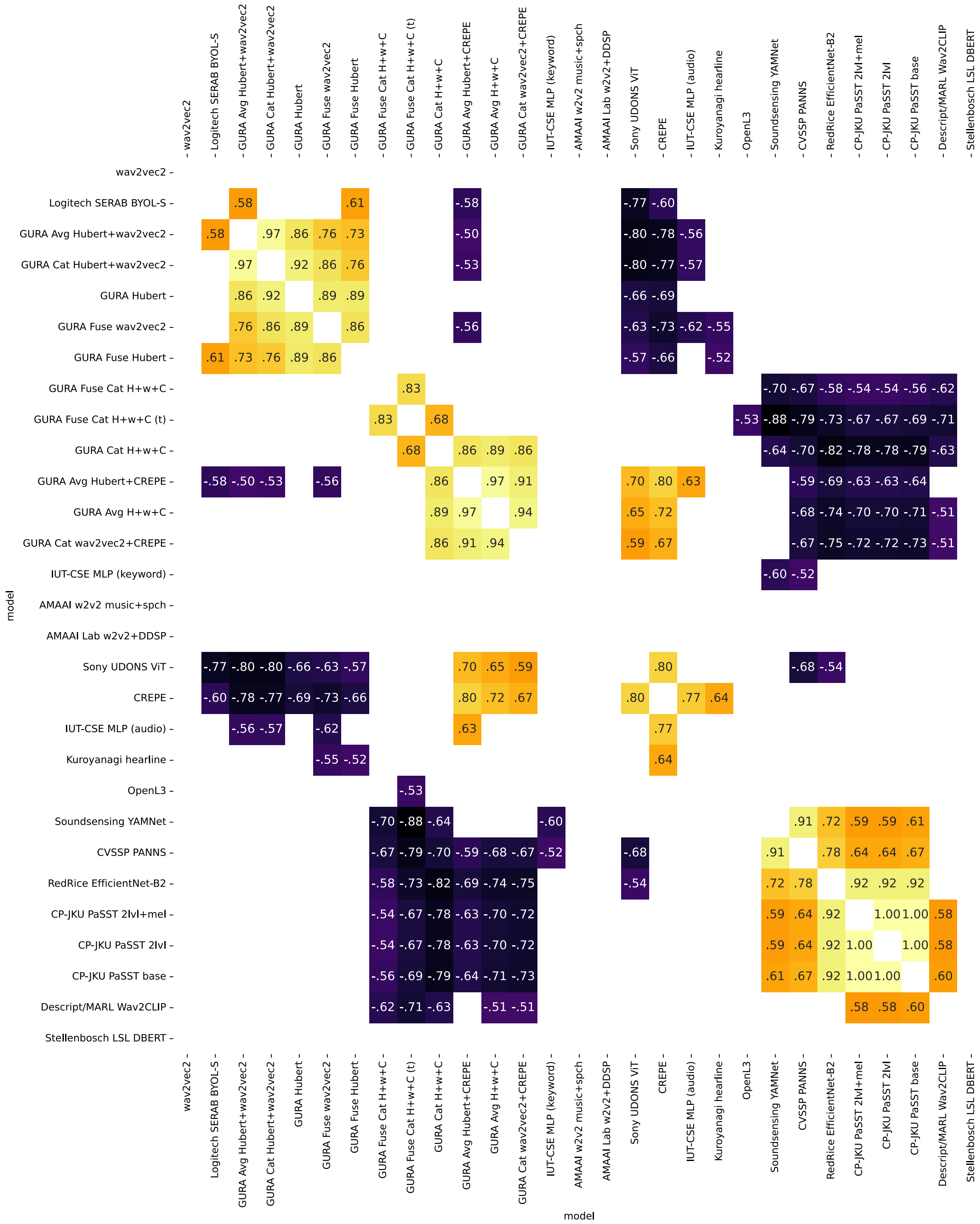


Figure 4: Model versus model correlation scores, based upon normalized scores. Only the highest and lowest correlations are displayed. Cells are sorted to minimize the traveling salesperson distance, mapping correlations [-1, +1] to distances [+2, 0].

cross-entropy loss. For scene-based multilabel tasks and multilabel frame reductions of timestamp tasks, the final layer is a sigmoid with cross-entropy loss.

We monitor the score (not loss) on the validation set. For timestamp tasks, computing the validation score involves a full CPU-based `sed_eval` (Mesaros et al., 2016) run with median filter of 250ms and minimum event duration 125 ms and 250 ms. (Both event durations are tried at each validation step and the best hyperparameter is retained for that validation step.) We train for a maximum of 500 epochs, checking the validation score every 3 epochs, early stopping if no improvement is seen after 20 validation steps. For DCASE 2015 Task 2, we check the validation score every 10 epochs.

The validation score is used for early-stopping, as well as for model selection. The same RNG seed is used for every model-task downstream training, ensuring that grid points and weight initialization is identical. Model selection is performed over 8 deterministic random grid points out of 16 possible grid points. Hyperparameters are shown in Table 4. This grid was chosen after using a much larger hyperparameter grid with the three baseline models on the open tasks. In these preliminary hyperparameter grid pruning experiments, the grid was progressively refined by discarding hyperparemeter choices that were not predictive of relatively high model performance, similarly to how Kelz et al. (2016) use tree ensemble learning to prune their hyperparameter grid.

Table 4: Hyperparameters used for training.

Hidden layers	[1, 2]
Hidden dimensions	1024
Dropout	0.1
Learning rate	[3.2e-3, 1e-3, 3.2e-4, 1e-4]
Batch size	1024
Hidden norm	Batch Norm
Initialization	[Xavier Uniform, Xavier Normal] (Glorot and Bengio, 2010)
Optimizer	Adam