A APPENDIX

A.1 TOPOAUDIO PERFORMANCE

Table 1 reports classification accuracy and representational smoothness for both baseline and topographic variants of Transformer-B/32 and ResNet-50 backbones. Across datasets (ESC50, NSynth, and Speech Commands), accuracy remains nearly unchanged when introducing topographic constraints (τ), with performance differences typically within $< 1\%$ of baseline. In contrast, smoothness values increase substantially, confirming that topographic regularization induces more spatially coherent representations. These results demonstrate that TopoAudio models preserve strong classification performance while simultaneously improving internal topographic structure, supporting their utility as both effective and interpretable auditory models.

Table 1: **Topographic auditory models maintain high classification performance across evaluations.** Accuracy is reported for ESC50, NSynth, and Speech Command datasets using Transformer-B/32 and ResNet-50 backbones. While baseline models achieve slightly higher accuracy, introducing topographic constraints (τ) substantially increases representational smoothness with only modest changes in classification performance. Topographic Avg. indicates the mean performance across all non-baseline τ values.

Topography (τ)	**Accuracy**			**Smoothness**
	ESC50	NSynth	SpeechCmd	
Transformer-B/32				
Baseline	82.10	98.25	92.94	0.31
5	81.94	98.13	92.63	0.46
25	82.01	98.13	92.80	0.50
50	81.66	97.99	92.42	0.57
100	81.88	97.91	92.38	0.56
Topographic Avg.	81.87	98.04	92.56	0.52
ResNet-50				
Baseline	81.69	98.29	86.68	0.34
5	81.46	98.50	86.88	1.10
25	80.62	98.39	87.89	0.87
50	80.84	98.36	87.57	1.04
100	80.32	98.72	86.85	1.09
200	80.70	98.48	87.47	0.93
Topographic Avg.	80.79	98.49	87.33	1.01

A.2 SPATIAL LOSS

To investigate how topographic constraints shape auditory representations, we adapted the TopoLoss framework (Deb et al., 2025) to the auditory domain. As before, we define a 2D "cortical sheet" from convolutional layers in the auditory model on which to enforce topography. Each convolutional kernel in the model is mapped onto this sheet. For a convolutional layer with c_{input} input channels and c_{output} output channels, and a kernel size of $k \times k$, the weight tensor $W \in \mathbb{R}^{c_{\text{output}} \times c_{\text{input}} \times k \times k}$ is reshaped into a cortical representation $C \in \mathbb{R}^{h \times w \times d}$, where $h \times w = c_{\text{output}}$, and $d = c_{\text{input}} \cdot k \cdot k$.

To encourage smoothness in the cortical sheet $C^{h\times w\times d}$, we apply a blurring operation that removes high-frequency variations. We compute a blurred version C' of the cortical sheet using a downsampling factor $\phi_h = \phi_w = 3$ followed by upsampling:

$$\text{Blur}(X, \phi_h, \phi_w) = f_{\text{up}}\left(f_{\text{down}}\left(X, \frac{h}{\phi_h}, \frac{w}{\phi_w}\right), h, w\right) \tag{4}$$

The *TopoLoss* is then defined as the negative mean cosine similarity between the original and blurred cortical maps:

$$\mathcal{L}_{\text{topo}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{C_i \cdot C_i'}{\|C_i\|\|C_i'\|} \tag{5}$$

This encourages neurons with similar functions to be spatially clustered, enhancing topographic organization. Finally, we integrate the *TopoLoss* with the primary task loss $\mathcal{L}_{\text{training}}$ as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{training}} + \tau \cdot \mathcal{L}_{\text{topo}} \tag{6}$$

where τ is a scaling coefficient controlling the influence of topographic regularization. Higher values of τ encourage stronger topographic organization.

A.3 fMRI Datasets

A.3.1 NH2015

The fMRI data used in this study are a subset of those originally reported in (Norman-Haignere et al., 2015), with procedures summarized below.

Participants and Experimental Design. Eight right-handed, native English-speaking participants (4 female; mean age 22 years, range 19–25) with normal hearing and no formal musical training participated in the study. Each participant completed three fMRI sessions (˜2 hours each). Five additional participants were excluded due to either incomplete scanning sessions or excessive head motion and task non-compliance. All participants gave informed consent under protocols approved by the MIT Committee on the Use of Humans as Experimental Subjects (protocol 2105000382).

Stimuli. A total of 165 two-second natural sounds were selected to span a wide range of real-world auditory categories. Each sound was validated using a 10-way forced-choice classification task on Amazon Mechanical Turk and included only if recognized with at least 80% accuracy. Stimulus names and categories are available in the supplementary materials of (Tuckute et al., 2023), and the full stimulus set can be downloaded from: `http://mcdermottlab.mit.edu/downloads.html`.

fMRI Procedure. Stimuli were presented in a blocked design, with each block consisting of five repetitions of the same 2-second sound, interleaved with 200 ms of silence to minimize scanner noise. Each block lasted 17 s (TR = 3.4 s), and silence blocks of equal duration were interspersed to estimate baseline responses. To ensure attentiveness, participants performed an intensity discrimination task in each block, identifying the quietest sound (7 dB lower than the others) via button press.

Data Acquisition. Data were acquired on a 3T Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center (MIT). Each run consisted of 15 slices oriented parallel to the superior temporal plane (TR = 3.4 s, TE = 30 ms, flip angle = 90°). The in-plane resolution was 2.1 mm $\times$ 2.1 mm, with 4 mm thick slices and a 10% gap (voxel size: $2.1 \times 2.1 \times 4.4$ mm). The first 5 volumes of each run were discarded.

Preprocessing. Preprocessing was conducted using FSL, FreeSurfer, and custom MATLAB scripts. Functional data were motion- and slice-time corrected, linearly detrended, skull-stripped, and aligned to each participant's anatomical scan using FLIRT and BBRegister. Volumes were projected to the reconstructed cortical surface using FreeSurfer and smoothed with a 3-mm FWHM 2D Gaussian kernel. Percent signal change was computed relative to silence blocks, and responses were downsampled to a 2-mm isotropic grid on the FreeSurfer surface. All participants' data were registered to the `fsaverage` template.

Voxel Selection. Voxel selection followed the criteria in (Tuckute et al., 2023). We retained voxels within a superior temporal and posterior parietal mask if they met two conditions: (1) significant sound vs. silence response ($p < 0.001$, uncorrected), and (2) reliable responses to sounds across scan sessions, quantified as:

$$r = 1 - \frac{\|\mathbf{v}_{12} - \text{proj}_{\mathbf{v}_3}\mathbf{v}_{12}\|_2}{\|\mathbf{v}_{12}\|_2}, \quad \text{with} \quad \text{proj}_{\mathbf{v}_3}\mathbf{v}_{12} = \left(\frac{\mathbf{v}_3 \cdot \mathbf{v}_{12}}{\|\mathbf{v}_3\|_2^2}\right)\mathbf{v}_3$$

Here, $\mathbf{v}_{12}$ is the voxel's response vector (averaged over the first two sessions) to all 165 sounds, and $\mathbf{v}_3$ is the same voxel's response from the third session. This measure captures the fraction of variance in $\mathbf{v}_{12}$ explained by $\mathbf{v}_3$. Voxels with $r \geq 0.3$ were retained. Across participants, this yielded 7,694 voxels (mean per participant: 961.75; range: 637–1,221).

A.3.2 B2021

The B2021 fMRI dataset used in this study was originally collected and analyzed by (Boebinger et al., 2021) and reanalyzed in (Tuckute et al., 2023). We summarize the methodological details below.

Participants and Experimental Design. Twenty right-handed participants (14 female; mean age: 25 years, range: 18–34) each completed three fMRI sessions (˜2 hours per session). Half of the participants ($n = 10$) were highly trained musicians, with an average of 16.3 years (SD = 2.5) of formal training that began before age 7 and continued through the time of scanning. The other half ($n = 10$) were non-musicians with fewer than 2 years of musical training, none of which occurred before age 7 or within 5 years of scanning. All participants provided informed consent, and the study was approved by the MIT Committee on the Use of Humans as Experimental Subjects (protocol number 2105000382).

Stimuli. The stimulus set consisted of 192 natural sounds, including 165 from (Norman-Haignere et al., 2015) and 27 additional music and drumming clips representing diverse musical cultures. To ensure comparability with NH2015, all analyses in this study were restricted to the shared subset of 165 sounds.

fMRI Procedure. The scanning procedure closely followed that of (Norman-Haignere et al., 2015), with some modifications. Each stimulus block consisted of three repetitions of a 2-second sound, lasting 10.2 seconds total (TR = 3.4 s, 3 repetitions). Each participant completed 48 runs across the 3 sessions (16 runs per session), with each run containing 24 stimulus blocks and 5 randomly interleaved silent blocks. This design enabled each sound block to be repeated 6 times across the experiment. Participants performed an intensity discrimination task, pressing a button upon detecting the quietest of the three repetitions in a block (12 dB lower).

Data Acquisition. MRI data were collected using a 3T Siemens Prisma scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at MIT. Functional volumes (48 slices per volume) covered the superior temporal and parietal lobes, matching the anatomical mask used in (Norman-Haignere et al., 2015). Imaging parameters were: TR = 3.4 s (TA = 1 s), TE = 33 ms, flip angle = 90°, in-plane resolution = 2.1 mm, slice thickness = 3 mm (10% gap), and voxel size = $2.1 \times 2.1 \times 3.3$ mm. A multiband SMS factor of 4 was used to accelerate acquisition. Structural T1 images (1 mm isotropic) were also collected.

Preprocessing. Preprocessing matched the pipeline used in (Norman-Haignere et al., 2015), but with a general linear model used to estimate voxel responses due to the shorter stimulus blocks and increased overlap in BOLD responses. For each stimulus block, beta weights were computed using a boxcar function convolved with a canonical hemodynamic response function, along with 6 motion regressors and a linear trend term. Resulting beta weights were downsampled to a 2-mm isotropic grid on the FreeSurfer cortical surface. Each participant's cortical surface was registered to the `fsaverage` template.

Voxel Selection. Voxels were selected using the same reliability-based procedure described in (Tuckute et al., 2023). Reliability was computed from vectors of beta weights for the 165 shared stimuli, estimated separately from two halves of the data (v_1 = runs 1–24, v_2 = runs 25–48):

$$r = 1 - \frac{\|\mathbf{v}_{12} - \text{proj}_{\mathbf{v}_3}\mathbf{v}_{12}\|_2^2}{\|\mathbf{v}_{12}\|_2^2}, \quad \text{where} \quad \text{proj}_{\mathbf{v}_3}\mathbf{v}_{12} = \left(\frac{\mathbf{v}_3 \cdot \mathbf{v}_{12}}{\|\mathbf{v}_3\|_2^2}\right)\mathbf{v}_3$$

Voxels with $r \geq 0.3$ and significant sound-evoked responses ($p < 0.001$, uncorrected) were retained. This procedure yielded a total of 26,792 reliable voxels across 20 participants (mean: 1,340 per participant; range: 1,020–1,828).

A.4 Voxelwise Response Modeling

This procedure was repeated 10 times (once per train-test split), and the median corrected variance explained was reported for each voxel-layer pair. We evaluated all layers from each candidate model on both datasets, yielding voxelwise explained variance values for 7,694 voxels (NH2015) and 26,792 voxels (B2021).

Regularized linear regression and cross-validation. To model the relationship between model unit activations and measured brain responses, we used voxelwise linear encoding models. For each voxel, we predicted its time-averaged response to natural sounds as a linear combination of time-averaged activations from a specific model layer. We randomly split the 165 sounds into 10 unique train-test partitions of 83 training and 82 test sounds. For each split, we fit a regularized linear regression (ridge regression) model using the 83 training sounds and evaluated prediction performance on the held-out 82 sounds.

Regression formulation. Let $\mathbf{y} \in \mathbb{R}^n$ be the voxel's mean response to $n = 83$ sounds, and let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix of d regressors (i.e., time-averaged activations from a model layer). The ridge solution is:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

where λ is the regularization parameter and $\mathbf{w}$ is the vector of regression weights. Both $\mathbf{y}$ and the columns of $\mathbf{X}$ were demeaned (but not normalized) prior to regression. This allowed units with greater magnitude variance to contribute more to the prediction under a non-isotropic Gaussian prior. To avoid data leakage, all transformations were learned on the training set and applied to the test set.

We used leave-one-out cross-validation within the 83 training sounds to select λ. For each of 100 logarithmically spaced values (from 10^{-50} to 10^{49}), we computed the mean squared error of the predicted response for each left-out training sound. The λ minimizing this error was used to retrain the model on all 83 training sounds. The final model was then used to predict responses to the 82 held-out test sounds, and performance was quantified using the Pearson correlation between predicted and actual voxel responses. Negative correlations or correlations with zero variance were set to zero.

Correcting for reliability of predicted voxel responses. Because both training and test responses are affected by measurement noise, we corrected for the reliability of both the predicted and measured voxel responses. This correction was essential to fairly compare model performance across voxels and model layers. We defined the corrected variance explained using the attenuation-corrected squared correlation:

$$r^2_{\mathbf{v},\hat{\mathbf{v}}^*} = \frac{r(\mathbf{v}_{123}, \hat{\mathbf{v}}_{123})^2}{r'_v r'_{\hat{v}}}$$

where $\mathbf{v}_{123}$ is the voxel response to the 82 test sounds, $\hat{\mathbf{v}}_{123}$ is the predicted response, and r'_v, $r'_{\hat{v}}$ are the reliabilities of the measured and predicted responses, respectively. Reliability was estimated via median Spearman–Brown corrected correlations across scan pairs. For stability, we excluded voxels for which r'_v or $r'_{\hat{v}}$ was less than $k = 0.182$ and $k = 0.183$, respectively (corresponding to $p < 0.05$ thresholds for 83- and 82-dimensional Gaussian variables).