PNAS RESEARCH ARTICLE | NEUROSCIENCE

A hierarchy of processing complexity and timescales for natural sounds in the human auditory cortex

Kyle M. Rupp[a,1], Jasmine L. Hect[a], Emily E. Harford[a], Lori L. Holt[b], Avniel Singh Ghuman[a], and Taylor J. Abel[a,c,1]

Affiliations are included on p. 10.

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received June 18, 2024; accepted March 21, 2025

Efficient behavior is supported by humans' ability to rapidly recognize acoustically distinct sounds as members of a common category. Within the auditory cortex, critical unanswered questions remain regarding the organization and dynamics of sound categorization. We performed intracerebral recordings during epilepsy surgery evaluation as 20 patient-participants listened to natural sounds. We then built encoding models to predict neural responses using sound representations extracted from different layers within a deep neural network (DNN) pretrained to categorize sounds from acoustics. This approach yielded accurate models of neural responses throughout the auditory cortex. The complexity of a cortical site's representation (measured by the depth of the DNN layer that produced the best model) was closely related to its anatomical location, with shallow, middle, and deep layers associated with core (primary auditory cortex), lateral belt, and parabelt regions, respectively. Smoothly varying gradients of representational complexity existed within these regions, with complexity increasing along a posteromedial-to-anterolateral direction in core and lateral belt and along posterior-to-anterior and dorsal-to-ventral dimensions in parabelt. We then characterized the time (relative to sound onset) when feature representations emerged; this measure of temporal dynamics increased across the auditory hierarchy. Finally, we found separable effects of region and temporal dynamics on representational complexity: sites that took longer to begin encoding stimulus features had higher representational complexity independent of region, and downstream regions encoded more complex features independent of temporal dynamics. These findings suggest that hierarchies of timescales and complexity represent a functional organizational principle of the auditory stream underlying our ability to rapidly categorize sounds.

auditory cortex | intracerebral recordings | deep neural networks | encoding | natural sounds

Significance

We compared human brain recordings during natural sound listening to representations within an artificial (deep) neural network (DNN) designed to categorize sounds. Despite not being exposed to real neural data during training, the DNN's representations closely resembled auditory cortex responses. We found that the complexity of a cortical site's representation (i.e., its best-matching DNN layer) increased across the auditory hierarchy and varied smoothly within regions, even those traditionally thought to represent uniform information types (e.g., primary auditory cortex). Finally, cortical sites with more complex representations required longer from sound onset to begin representing those features, even within regions of the auditory hierarchy, suggesting that both region and temporal dynamics are independently related to the complexity of a site's representation.

Humans encounter diverse sounds in daily life that require rapid categorization. Is a cell phone vibrating in the other room or did a bee get inside? Was that a whisper or just the wind? The brain recognizes sounds with vastly different acoustic signatures as functionally equivalent (e.g., a laugh and scream are human vocalizations) and also differentiates acoustically similar sounds across incongruent classes (e.g., a man humming and a flute playing the same note). The cortical mechanisms responsible for our ability to differentiate and identify sounds according to the spectrotemporal acoustic patterns that characterize classes of sound have been studied extensively as auditory categorization (1–4). There is widespread consensus for a hierarchical organization of the auditory cortex, with primary areas representing acoustic features and downstream regions progressively encoding representations better aligned with categories. Yet, the representational complexity of auditory cortical encoding has been difficult to quantify objectively, limiting our ability to characterize the progression of information throughout this hierarchy.

Prior work suggests early auditory responses can be explained by their responsivity to sounds' spectrotemporal features and modulations thereof (5–8), with successful characterization of speech encoding via a spectrotemporal modulation (STM) tuning framework applied to both primary and downstream auditory cortex (9, 10). Furthermore, STM models can be applied to reconstruct speech (11) and natural sounds from neural responses (12). However, the majority of acoustic sensory input we receive falls within a fraction of STM space (13, 14), which only partially explains why this approach is unable to produce a fully generalizable model of neural responses outside of the primary auditory cortex (2, 5, 15–20). More importantly, STM features measure acoustic properties and are thus not well-suited to describe responses in the higher-order auditory cortex, which responds to sound categories such as vocalizations (21) and music (16), exemplars of which may have highly distinct spectrotemporal features.

Author contributions: K.M.R. and T.J.A. designed research; K.M.R., J.L.H., and E.E.H. performed research; K.M.R. analyzed data; and K.M.R., L.L.H., A.S.G., and T.J.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: kyle.m.rupp@pitt.edu or abeltj@upmc.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2412243122/-/DCSupplemental.

Published April 28, 2025.

Recently, task-optimized deep neural network (DNN) models have been used to explore representations throughout the auditory cortex, generating compelling evidence for a nonlinear hierarchically organized auditory network (15, 17, 18, 22–25) that transforms incoming acoustic input into salient, behaviorally relevant representations (e.g. speech, music, environmental sounds). Critically, commonly used feedforward DNN architectures that incorporate nonlinear transformations at each layer result in stimulus representations that become increasingly complex with DNN layer depth (26). In the case of a DNN trained to classify sounds into semantic categories, this transformation approximates a continuum from acoustic feature encoding in shallow layers to abstract, semantic category representations in deep layers (15, 27, 28). This framework can be used to quantitatively estimate representational complexity, i.e., the complexity of feature representations, across the auditory cortex.

Here, in the context of epilepsy surgery evaluation, we acquired intracerebral electrophysiology from 20 patient-participants as they listened to a rich set of natural sounds that spanned multiple categories, including speech and nonspeech vocalizations, music, animal vocalizations, and environmental sounds (16). We extracted representational features from different layers within a sound-categorization DNN, which we will refer to as DNN features, and built encoding models to predict neural responses throughout the auditory cortex. This task-optimized DNN was trained to evaluate a sound's spectrogram and classify it within a broad hierarchical taxonomy (e.g., human vocalizations such as speech and laughter, music genres and instruments, types of mechanical sounds, and so on) and thus was well suited to explore categorization of natural sounds in the human auditory cortex. Across recording sites, we used these encoding models to assess the relationship between neural representations and the representations within different layers of the DNN. Representational complexity was inferred by the layers of the DNN that best explained the neural responses, with deeper layers corresponding to more complex representations; we then characterized how complexity varied across and within core, belt, and parabelt regions. Finally, we used sliding encoding models to estimate the time relative to sound onset when a recording site begins encoding DNN features, termed the encoding onset time, and then explored how representational complexity was related to both this property of temporal dynamics and anatomical position within the auditory cortical hierarchy. By characterizing how complexity and timescales vary (and covary) across the auditory cortex, we elucidate functional organizational principles underlying the processing of natural sounds in the auditory cortex.

Results

A total of 20 patient-participants performed a one-back auditory task using the freely available Natural Sounds stimulus set (16), which consists of 165 two-second sounds (Fig. 1*A*). One-back repeats occurred on 17% of trials, and patients recorded their responses using a button box. The full stimulus set was presented in random order three separate times, and broadband high gamma activity (HGA, 70 to 150 Hz) aligned to stimulus onset was extracted and used in all analyses (Fig. 1*B*).

Across these 20 patients, there were a total of 3145 intracerebral recording sites, of which 811 were auditory-responsive, defined as a statistically significant difference between HGA responses to stimuli compared to baseline (two-sample *t* test across 594 trials, df: 1186, $P < 0.01$, false discovery rate corrected). For this test, HGA values for each trial were averaged across baseline and response windows, which were both 800 ms long and began at −900 and 0 ms relative to stimulus onset, respectively; henceforth, the t-statistics from these tests will be referred to as auditory responsiveness. Of those 811 channels, 755 exhibited increased responses, while 56 channels showed decreased responses. There

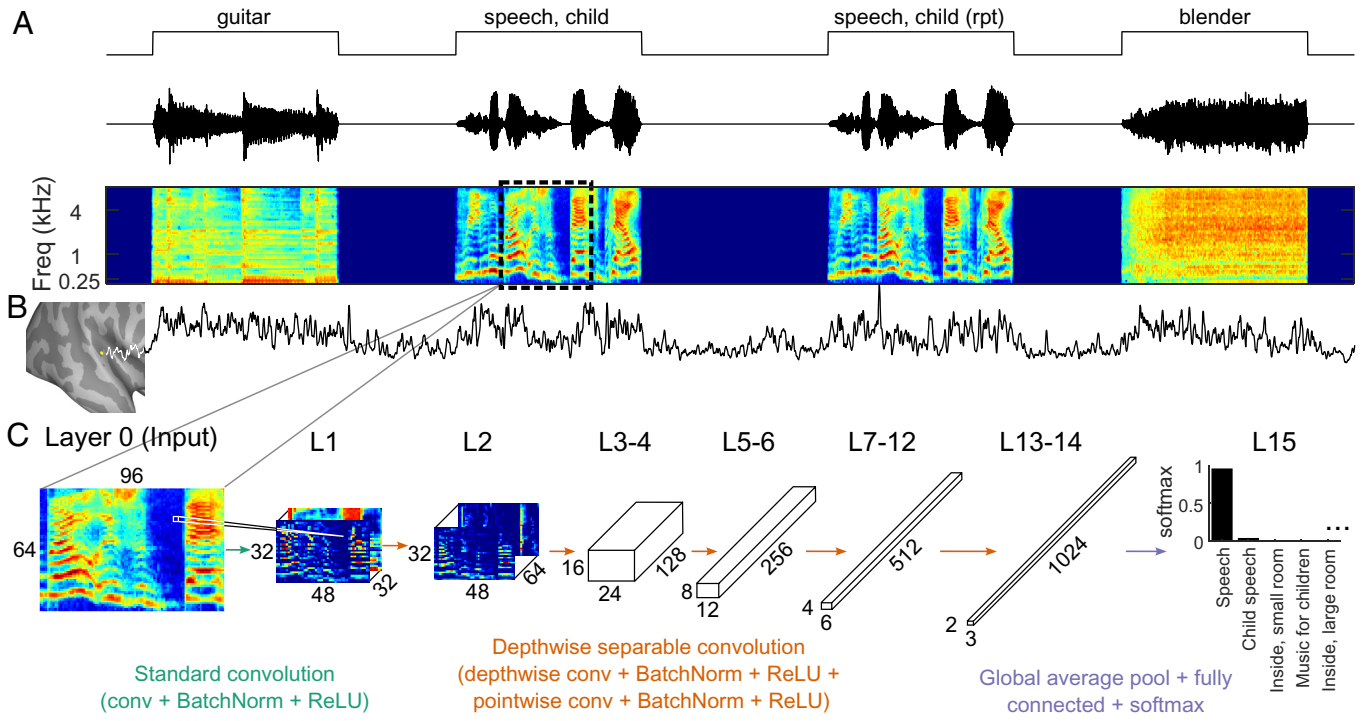

Fig. 1. Methods. (*A*) Patients performed an auditory 1-back task using Natural Sounds stimuli. The dashed black box in the auditory spectrogram represents the 975 ms input window for the DNN (see panel *C*). (*B*) Broadband high gamma activity (HGA) from an example channel. (*C*) YAMNet DNN model architecture. Arrow colors represent different blocks of DNN layer operations. Depthwise separable convolutions were also used between the grouped layers in the figure (L3-4, L5-6, L7-12, and L13-14). Using this pretrained DNN, layer activations for each stimulus were extracted and used to build encoding models to predict HGA.

were 388 channels localized to the auditory cortex, with 303 of these exhibiting auditory responses (core: 93 auditory-responsive/95 total channels, medial belt: 10/10, lateral belt: 105/106, and parabelt: 95/177). All auditory-responsive channels in the auditory cortex exhibited increased responses. Due to the sparse coverage within medial belt, this region was excluded from all region-of-interest (ROI) analyses. Fig. 2*A* shows mean HGA responses across all auditory-responsive channels within each region (core, lateral belt, and parabelt).

First, we sought to establish whether a DNN trained to categorize sounds possessed similar representations to HGA recorded throughout the human auditory cortex. To this end, we built lasso-regularized linear regression encoding models to predict instantaneous HGA responses for all auditory-responsive channels, sampled every 50 ms from 50 to 2,000 ms poststimulus onset, resulting in 6,600 observations (40 timepoints × 165 stimuli). For each of the 165 sounds, we used the DNN YAMNet (27, 28) to extract features by calculating the node activations at each layer (Fig. 1*C*), with deeper layers containing features of increasing representational complexity. The DNN features were calculated using a window spanning from 975 to 0 ms before a given HGA sample (e.g., for the sample at 50 ms, the stimulus waveform was extracted from −925 to 50 ms with appropriate zero-padding; the acoustic spectrogram was then calculated and input to YAMNet). Due to computational constraints, we applied dimensionality reduction to these features using sparse random projection; we then used the resultant feature matrices as inputs to our encoding models, with separate models built for each layer. We refer to these as full models to differentiate them from sliding models (built at individual timepoints), which are introduced later. Model performance was assessed using the coefficient of determination (R^2), which corresponds to one minus the ratio of the sum of squared residuals over a null model's sum of squared errors. The null model used here was one that predicts the time-dependent mean neural response (i.e., the mean across stimuli but not across time) rather than the global mean (i.e., the mean across stimuli and time) to ensure that models were not simply capturing the generic stimulus response dynamics.

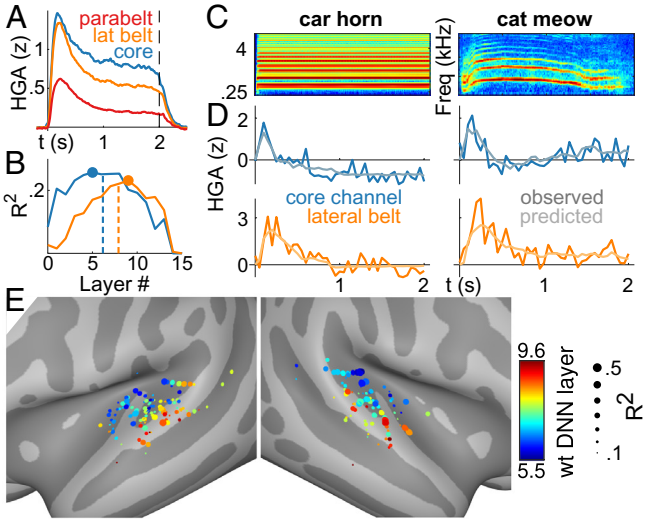

Fig. 2. Full encoding model results. (*A*) Mean HGA responses across all trials and channels within different regions of the auditory cortex. (*B*) Encoding model accuracies (R^2) across DNN layers for two representative channels (blue: core, lateral belt: orange). Points show the layer that maximizes R^2, and dashed lines show the weighted DNN layer, which is the weighted mean of each curve. (*C*) Acoustic spectrograms for two example stimuli. (*D*) Example responses to the stimuli in *C*, averaged across all presentations of a given stimulus, from the same channels as *B* (core: *Top*; lateral belt: *Bottom*). Dark and light lines correspond to observed responses and encoding model predictions respectively. Predictions were generated by the peak models (i.e., the points) in *B*. (*E*) Encoding model results across patients and channels. Neural prediction accuracy for the best model is shown by marker size. Color represents the weighted DNN layer (the dashed lines from panel *B*), with red indicating greatest representational complexity.

Fig. 2 *B–D* shows example modeling results for two channels. Fig. 2*B* shows encoding accuracy across all layers of YAMNet for these two channels, with the blue (core) and orange (lateral belt) channels best predicted by shallower and middle layers respectively. Fig. 2*D* shows predicted versus observed responses for two example stimuli, whose auditory spectrograms are shown in Fig. 2*C*; predictions were generated using the peak models of each channel (i.e., the dots in Fig. 2*B*).

Neural responses throughout the auditory cortex could be predicted well, achieving R^2 values up to 0.57; conversely, most auditory-responsive channels outside of the auditory cortex were poorly modeled using this encoding approach. Of the 303 auditory-responsive channels in the auditory cortex, 204 channels (67%) achieved a peak R^2 (i.e., maximum R^2 across DNN layers) greater than 0.1. In contrast, only 4% (20/508) of auditory-responsive channels outside of the auditory cortex achieved a peak R^2 over 0.1, suggesting that most responses outside of the auditory cortex do not encode the stimulus features captured by the DNN. Finally, channels with decreased stimulus responses relative to baseline (all of which were outside of the auditory cortex) could not be predicted well with these encoding models; all 56 such channels achieved a peak R^2 less than 0.011.

We then used linear mixed-effects models with patient as a random effect to compare encoding performance across hemisphere and ROIs, controlling for auditory responsiveness by including it as a fixed-effects variable. Within auditory-responsive channels in the auditory cortex, peak R^2 values were lower in parabelt compared to core [$t(288) = 4.27$, $P = 2.7 \times 10^{-5}$]and lateral belt [$t(288) = 4.84$, $P < 10^{-5}$] regions, even when controlling for auditory responsiveness. This is partly due to the fact that relatively fewer parabelt channels could be predicted well ($R^2 > 0.1$) compared to core and lateral belt (core: 80/93, 86%; lateral belt: 83/105, 79%; parabelt: 36/95, 38%). However, these differences in proportions do not fully explain differences in model performance between ROIs; an LME model using only channels with $R^2 > 0.1$ still found that parabelt performed worse than core [$t(194) = 2.36$, $P = 0.019$] and lateral belt [$t(194) = 2.65$, $P = 0.0087$]. We also explored a threshold of $R^2 > 0.05$; while the proportion of channels exceeding this threshold increased substantially (*SI Appendix*, Table S1), the LME modeling results were not impacted significantly (*SI Appendix*, Table S2). In other words, only a subset of parabelt channels appear to encode the DNN features embedded within YAMNet, potentially due to the functional heterogeneity within this region: for example, YAMNet was trained to categorize speech sounds at the superordinate level of "Speech" and thus may not extract phonetic features that are known to be encoded within superior temporal gyrus [STG; (29)]. Furthermore, the long window over which the DNN operates (a single category-level prediction for a 975 ms window) may result in slow-varying high-order DNN representations that are poorly reflected in instantaneous parabelt HGA responses. While a DNN that operates over shorter timescales might capture features with a higher correspondence to parabelt responses, building and training such a model was outside the scope of this work. A final possibility is that a one-back auditory task does not explicitly require sound categorization but can instead be completed using lower-level acoustic feature matching. Given these task demands, it is possible that parabelt regions were not activated with sufficient strength to demonstrate widespread encoding success (30).

Nevertheless, the parabelt channels that were modeled well displayed interesting properties that we elucidate in subsequent analyses.

Next, we estimated representational complexity, which we define as the complexity of features encoded by each channel. This was quantified as the DNN layer that produced the most accurate predictions of neural responses, with deeper layers having undergone more nonlinear transformations and thus corresponding to more complex features. While this could be done by selecting the DNN layer that maximizes R^2 (24), this approach is susceptible to noise. Instead, we calculated the weighted DNN layer, defined as the weighted mean of R^2 across DNN layers (i.e., the center of mass of the curves in Fig. 2*B*, shown as dashed lines). Note that we used this weighted DNN layer metric, calculated from full model results, for all remaining analyses that include representational complexity. Fig. 2*E* shows weighted DNN layers and peak R^2 values plotted across patients and channels, revealing a relationship of increasing representational complexity in the higher-order compared to the lower-order auditory cortex. This effect was statistically validated using an LME model with patient as a random effect and channels with peak $R^2 > 0.1$. Controlling for auditory responsiveness, core mapped to shallower DNN layers relative to lateral belt [$t(194) = -5.57$, $P < 10^{-5}$] and parabelt mapped to deeper DNN layers [$t(194) = 3.20$, $P = 0.0016$; also see Fig. 5 *B*, rightmost panel]. There was also a hemispheric effect, with left hemisphere mapping to deeper DNN layers than right hemisphere when controlling for ROI [$t(194) = 2.14$, $P = 0.034$]. Results were not qualitatively different when using a threshold of $R^2 > 0.05$ (*SI Appendix*, Table S3).

In addition to complexity differences between ROIs, we hypothesized that complexity gradients existed within each ROI along specific axes (see Methods for descriptions on how each axis was calculated). In core, the primary axis under consideration was oriented from posteromedial to anterolateral (PM–AL, Fig. 3*A*), which is roughly parallel to the presumed axis from human A1 to R. An orthogonal secondary axis that was parallel to the supratemporal plane (STP), as well as a tertiary axis perpendicular to STP, were also investigated (axes not shown in Fig. 3*A*). In lateral belt, axes with the same orientation as core were used, with the origin shifted to the lateral belt centroid. In parabelt, the three axes pointed from posterior to anterior (P-A) along the length of the STG, from ventral to dorsal (V-D), and from medial to lateral (Fig. 3*B*, M-L axis not shown). Channels belonging to a given ROI were projected to each of that ROI's axes, and correlations were calculated between these positions and representational complexity (i.e., weighted DNN layer). In both left and right core as well as right lateral belt, representational complexity increased as a function of position from posteromedial to anterolateral (Fig. 3*A*). Complexity also increased in right core from ventral to dorsal channels. In right hemisphere parabelt channels, complexity increased moving anteriorly as well ventrally (Fig. 3*B*). Position along the M-L axis was not correlated with complexity in parabelt. The correlations shown in Fig. 3 were also calculated for channels with $R^2 > 0.05$; these results were similar to those using the $R^2 > 0.1$ threshold and can be found in *SI Appendix*, Table S4.

We then tested whether these within-ROI complexity gradients were best explained by a linear or sigmoidal relationship. A linear relationship would demonstrate a smoothly varying degree of complexity along the axis, while a sigmoidal relationship might indicate a more step-like transition, which would be consistent with homogenous-complexity subregions within an ROI (for example, transitioning from A1 to R within core). This test was performed by fitting both linear and sigmoid models to the channel position versus weighted DNN layer relationship (i.e., the scatter plots in Fig. 3), and then comparing the two models using the Akaike information criterion. The sigmoid model contained four parameters that described the lower and upper asymptotes, the x-position of the midpoint between asymptotes, and the slope at that midpoint (i.e., how step-like the transition is). We required the x-position to fall between 0 and 1 but placed no other constraints when fitting the sigmoid model. We found that a linear model better explained these relationships than a sigmoidal one in left core (PM–AL, relative likelihood of sigmoidal compared to linear = 0.241), right lateral belt (PM–AL, relative likelihood = 0.387), and parabelt (P-A relative likelihood = 0.135; V-D relative likelihood = 0.203). In right core, complexity gradients were better explained by a linear model along the ventral–dorsal axis (relative likelihood = 0.135); along the PM–AL axis, sigmoidal and linear models were approximately equivalent (sigmoid > linear, relative likelihood = 0.854). In summary, representational complexity appeared to vary linearly along anatomical axes in nearly all ROIs with a significant correlation, with the exception of the PM–AL axis in right core (which had only slightly more support for a sigmoidal model).

To investigate whether there were any patterns in the encoding results, we calculated the stimulus-specific R^2 values from the best performing full model for each channel (Fig. 4*A*). We then applied hierarchical clustering across both channels and stimuli and identified the most prominent cluster. The stimulus cluster consisted almost entirely of speech and music containing singing (Fig. 4*B*). The cluster of channels, which showed stronger encoding results

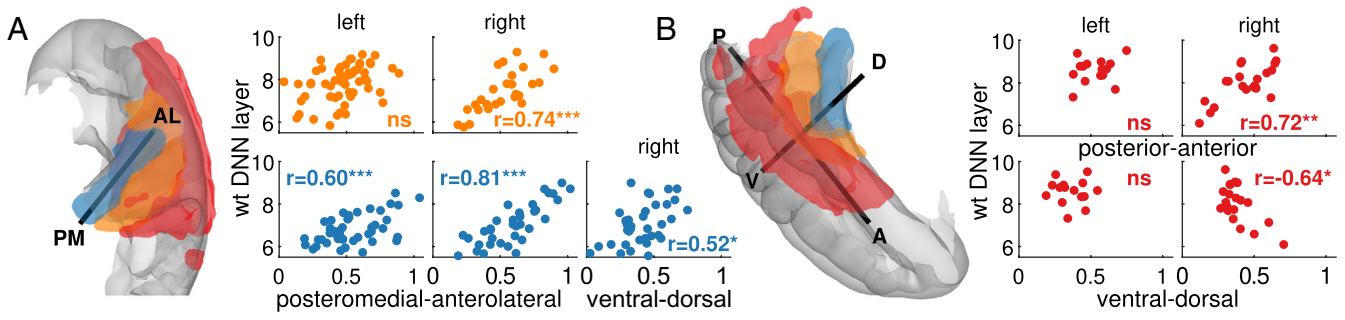

Fig. 3. Complexity gradients within ROIs. (*A*) A gradient of increasing representational complexity (indexed by weighted DNN layer) was found along a posteromedial–anterolateral (PM–AL) axis in both *Left* and *Right* core (*Bottom Left* two plots); additionally, a gradient was found from ventral–dorsal position in core (*Right* hemisphere only). In lateral belt, a PM–AL gradient was found in *Right* hemisphere only (*Top* plots). (*B*) Representational complexity gradients were also found in parabelt (*Right* hemisphere only) along the posterior–anterior and ventral–dorsal axes, with complexity increasing in the anterior and ventral directions. Results are across all patients and channels with full model peak $R^2 > 0.1$. Correlation significance was assessed using permutation testing with 100,000 permutations (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, Bonferroni corrected).

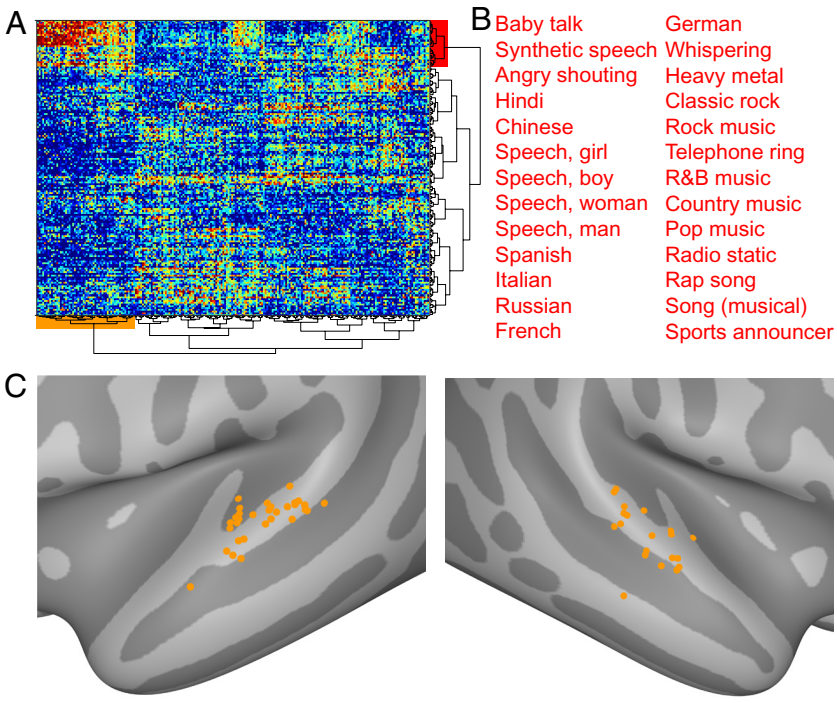

Fig. 4. Stimulus-specific encoding results. (*A*) R^2 values across channels (columns) and stimuli (rows), with hierarchical clustering applied. The most prominent cluster is highlighted in red (stimuli) and orange (channels). (*B*) The stimuli belonging to this cluster were primarily speech and music that included singing. (*C*) Channels in this cluster were localized to the bilateral higher-order auditory cortex and showed more accurate predictions for speech/singing responses compared to other stimuli.

for speech/singing compared to other stimuli, was primarily localized to the higher-order auditory cortex in both hemispheres.

Next, we sought to characterize the temporal dynamics of encoding by focusing on a property we call the encoding onset time, which describes the timepoint following sound onset when a channel begins encoding DNN features. We were interested in exploring how this property varied across the auditory cortex and how it related to representational complexity. We hypothesized that encoding onset times would be shortest in core and longest in parabelt and that this property would increase with representational complexity (i.e., channels with more complex feature encoding would also have later encoding onset times). For each channel with full model peak $R^2 > 0.1$, we identified the DNN layer that produced the best full model. This layer was then used to build sliding encoding models, where separate models were built for each timepoint (from 0 to 1,000 ms, sliding by 20 ms). As in the full model analysis, sliding models were trained to predict instantaneous HGA values, and input features consisted of layer activations from a window spanning −975 to 0 ms relative to each HGA sample. These input features underwent dimensionality reduction via sparse random projection before model fitting; see Methods for more details. We identified the encoding onset time as the timepoint where model performance reached 50% of its maximum value for a given channel. In this analysis, two channels were discarded because they had no timepoints where the sliding model R^2 exceeded 0.1.

Anatomical maps of encoding onset times are shown in Fig. 5*A*, with the earliest onsets appearing in core and the longest onsets in parabelt (see also the marginal distributions in Fig. 5 *B*, *Top* panel). A linear mixed-effects model (fixed effects: hemisphere and ROI, random effects: patient) confirmed that these visual trends were statistically significant [parabelt > core: $t(193) = 4.08$, $P = 6.7 \times 10^{-5}$]. Additionally, lateral belt was found to exhibit onset times that were later than core [$t(193) = 2.31$, $P = 0.022$] and earlier than parabelt [$t(193) = 2.34$, $P = 0.020$]. No hemispheric effect was observed. LME results for channels with $R^2 > 0.05$ can be found in *SI Appendix*, Table S5, as well as the effects of varying the onset threshold from 50% to 70%. The only qualitative difference when using these different thresholds was that for $R^2 > 0.1$ and an onset threshold of 70%, encoding onset times no longer differed significantly between parabelt and lateral belt. A highly significant relationship was also observed between encoding onset time and response latency (i.e., the first timepoint where the HGA response differed from baseline). The response latency for each channel was estimated by calculating its time-varying 95% CI across trials, finding the first window that exceeded the baseline mean for at least three consecutive samples (equivalent to 30 ms), and taking the first timepoint in this window (31). This analysis was restricted to timepoints following stimulus onset (i.e., $t \geq 0$ ms). Response latencies were highly correlated with encoding onset times (Spearman's rho = 0.72; $P < 10^{-5}$, permutation testing with 100,000 permutations).

Finally, we investigated the relationship between encoding onset times and representational complexity, shown in Fig. 5*B*. Features of higher complexity often span longer timescales, and neural populations that represent these features may require a longer time from sound onset to synthesize and code this information. However, as was shown in our previous analyses, both weighted DNN layers and encoding onset times vary systematically across the auditory cortical hierarchy. To investigate the relationship between encoding onset times and representational complexity stratified by ROI, we built an LME model with the weighted DNN layer as the response variable, patient as a random effect, and hemisphere, log(encoding onset time), and ROI as fixed effects. Finally, given its relationship with encoding onset time, log(response latency) was included as a fixed effect; its inclusion in the model helps to isolate the primary relationship of interest by accounting for this potential confound. (See *SI Appendix*, Table S7, for results

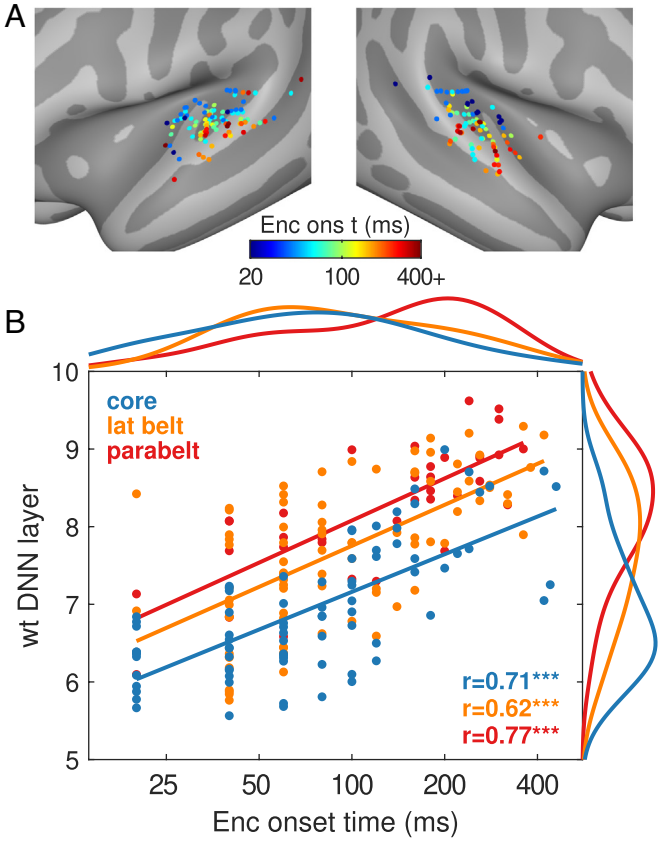

Fig. 5. Encoding onset times. (*A*) The first timepoint where sliding model encoding accuracy exceeds 50% of its maximum, termed the encoding onset time, is shown across all patients and channels (with full model peak $R^2 > 0.1$). Smaller onset times were observed in core and lateral belt while longer times appear in parabelt regions. (*B*) In all three ROIs (core, lateral belt, and parabelt), channels with later encoding onset times also encoded more complex representations (weighted DNN layer), with strong positive correlations observed between these two metrics. Marginal distributions show differences in representational complexity between all three regions (*Right*); encoding onset times (marginal distribution, *Top*) differed strongly between parabelt and both core and lateral belt; while differences were smaller between core and lateral belt, they were statistically significant. See main text for details. Correlation statistical significance assessed via permutation testing with 100,000 permutations (***$P < 0.001$, Bonferroni corrected). Correlations did not differ qualitatively when varying the R^2 and onset thresholds (*SI Appendix*, Table S6).

when varying the R^2 and onset thresholds; the only difference was the emergence of a hemispheric effect for $R^2 > 0.05$ and onset threshold = 50%.) While no hemispheric effect was observed for $R^2 > 0.1$ and onset threshold = 50%, we found highly significant effects of both encoding onset time and ROI, even when accounting for response latency (which also had a significant effect; see *SI Appendix*, Table S7). First, the strong positive relationship between encoding onsets and representational complexity (independent of ROI) indicates that channels that encode more complex features take longer for those features to be represented in the neural response, even within region [$t(192) = 7.75$, $P < 10^{-5}$; also see the correlations in Fig. 5*B*]. Furthermore, the significant independent effect of ROI (also apparent in Fig. 5*B*) shows that representational complexity increases across the auditory hierarchy, even for channels (across regions) with similar encoding onsets [$t(192) = 8.39$, $P < 10^{-5}$]. In other words, while parabelt channels with rapid onsets encoded less complex features than parabelt channels with slower onsets, their representations were still more complex than rapid onset core channels. Conversely, feature encoding in slow-onset core channels was more complex than rapid-onset core channels but still less complex than parabelt channels. This suggests that throughout the auditory cortex, there is a strong link between the temporal dynamics of encoding and underlying representational complexity, alongside an independent trend of increasing feature richness through the auditory processing hierarchy.

Discussion

In this work, we demonstrate that neural responses to natural sounds share similarities to representations embedded in a DNN model trained to categorize sounds. Notably, this DNN accepts acoustic spectrograms as inputs and outputs a semantic category label, was not exposed to any neural data during its training, and was trained on a separate stimulus set from the one eliciting the neural responses we measured. Nevertheless, the DNN representations could be used to predict neural responses with high accuracy using an encoding model approach, suggesting an overlap in the hierarchical representation of sounds across the DNN and human cortex. Furthermore, sites in increasingly higher-order cortical areas were best predicted by increasingly deeper layers of the DNN, suggesting areas further downstream in the auditory cortical hierarchy represent auditory information with increasing representational complexity. This relationship even existed within specific regions, with smoothly varying gradients of representational complexity observed along relevant anatomical axes. Sliding encoding models revealed that early auditory areas begin encoding DNN features more rapidly (relative to sound onset) than higher-order areas, which is consistent with existing models of auditory processing. Finally, we found that even when controlling for position in the auditory hierarchy, channels with increasing complexity were slower to begin encoding DNN features, suggesting a relationship between representational complexity and the temporal dynamics of encoding that is independent of position within the auditory hierarchy.

We showed that auditory cortical responses to natural sounds could be modeled using DNN layer activations and that higher-order areas in the auditory hierarchy are best modeled by increasingly deeper layers; both of these findings are aligned with prior studies (15, 24). This relationship between position within the auditory hierarchy and DNN layer depth is consistent with a broadly accepted model in which auditory core, which comprises the posteromedial two-thirds of Heschl's gyrus and is considered to be the primary auditory cortex (32, 33), processes low-level acoustic features (2, 34–36). In this model, lateral belt comprises planum temporale along with the anterolateral aspect of Heschl's gyrus (37, 38) and processes more complex or compositional acoustic features (35, 39–41). Finally, parabelt is situated in STG and the upper bank of the superior temporal sulcus (STS). While representations in this region are far more heterogeneous, they are best described as high level, abstract features (2, 16, 21, 29, 34).

We also found gradients of complexity within each of core, lateral belt, and parabelt, with these effects lateralizing to the right hemisphere outside of core. Within parabelt, complexity increased linearly along a posterior-to-anterior as well as a dorsal-to-ventral dimension. These findings are consistent with other studies that have proposed a flow of information along these axes (33, 42–47). For example, multiple voice processing nodes have been identified in STG and extending into STS (21), with the most posterior node found to process physical speaker characteristics such as vocal tract length (43, 44) and the most anterior one shown to encode individual voice identities (42, 45–47). Notably, these models describe discrete nodes within STG and STS, which would suggest a more step-like transition of complexity along the posterior–anterior axis.

However, we found that a linear model was better able to describe the gradient along this axis (as well as the gradient from dorsal to ventral), suggesting a more continuous and smoothly varying degree of complexity. Multiple studies have demonstrated a lateralization of voice processing toward the right hemisphere (21, 45–47), though this finding is extensively disputed (48). Our complexity measure was correlated with anatomic position only in the right hemisphere, which may be because conspecific vocalizations represent a particularly salient auditory category, and individual species may have evolved dedicated networks to process it (49). If the shared representations between parabelt sites and DNN layers are largely biased toward voice processing as suggested by our clustering analysis in Fig. 4, then this may account for these specific hemispheric differences. Notably, left parabelt/STG is often associated with speech processing (9, 34, 50–52), exhibiting a posterior-to-anterior flow of information (e.g., phonetic features to syllables/words) along the ventral stream (50, 52). While we could still predict left parabelt channels with high accuracy using DNN features, there was no relationship between DNN layer depth and anatomic position within this ROI. This may be because the DNN's training goals did not require learning progressively compositional speech features since the speech-related categories it was trained to predict were quite coarse (e.g., speech, babbling, speech synthesizer, etc.).

Notably, we found complexity gradients within the core region in both hemispheres along a posteromedial–anterolateral axis. Traditionally, core is believed to represent low-level acoustic features such as relatively simple spectrotemporal representations, typically showing a preference for single frequency bands (36, 53) that are relatively narrow in bandwidth (54, 55). This region is often described as having two subregions, termed hA1 and hR (human homologs to monkey A1 and the rostral area R), which are positioned sequentially along the PM–AL axis. In this model, hA1 and hR exhibit mirrored tonotopic gradients, with hA1 transitioning from high to low frequencies and hR from low to high along the PM–AL dimension (53). Interestingly, our results appear to contradict this model, which would predict a similar representational complexity throughout the core region. One possibility is that this gradient is driven by an increase in integration window length, which would require deeper layers to model within the DNN, given the narrow 3 × 3 extent of the DNN's convolutional filters. This possibility follows from the observation in primates that the core subregion R integrates over longer windows than A1 (56). However, this model would predict a stepwise or sigmoidal relationship across the hA1-hR boundary, which is inconsistent with our findings in left core of a linear relationship; the results in right core were ambiguous since the linear and sigmoidal models were similarly supported by the data.

Finally, lateral belt areas have been shown to respond to more complex sounds such as band-passed noise (35, 40). This area is also parcellated into multiple subregions, including hML and hAL (human homologs to middle lateral and anterolateral belt regions), which are positioned sequentially along the PM–AL axis. Studies have found that hML contains a tonotopic gradient, suggesting relatively lower-level acoustic processing, while hAL overlaps with voice-sensitive regions and exhibits a strong preference for low frequencies (a prominent feature of human vocalizations) (37, 55, 57). This latter finding suggests that hAL may engage in higher-order processing to support complex or abstract representations of voice (55). The complexity gradient we observed along the PM–AL axis of lateral belt was partially consistent with these findings, though we observed a linear gradient as opposed to a sigmoidal (i.e., step-like) one. It should be noted that for all analyses comparing linear and sigmoidal relationships for complexity gradients within ROIs, it is possible that these findings may be impacted by spatial blurring due to factors such as individual differences in anatomy and volume conduction; nevertheless, we believe they represent intriguing findings that warrant further investigation in future studies.

Finally, we found that channels with increased representational complexity had later encoding onset times, defined as the amount of time from sound onset for a channel to begin representing those features. This relationship persisted even when accounting for gross anatomical location (ROI position within the auditory hierarchy); this control is necessary, since our results showed that these onset times increased from core to lateral belt to parabelt; furthermore, other studies have shown that temporal properties such as integration windows also increase across the auditory hierarchy (58–60). Further exploration revealed that this relationship was observable in all three of these regions. These significant correlations throughout the auditory cortex support the general principle that more complex neural representations require longer time windows to integrate and combine simpler features into higher-order compositional representations, likely due (at least in part) to the longer temporal extent that these representations span.

In considering the findings presented here, several caveats should be taken into account. First, as with nearly all studies involving intracranial recordings, our electrodes are implanted based on clinical necessity, which may cause a selection bias for recording sites. While the patients involved are not neurotypical, their perceptual performance often is, and the spatial and temporal resolution of these direct cortical recordings provide substantial value both in testing existing theories and offering findings that can then be tested using other experimental modalities and approaches. Furthermore, the large number of patients included in this study (N = 20) and the heterogeneity of their pathologies (Table 1) provide some assurance that any individual deviations would have minimal impact on the group results, assuming that those deviations are distributed randomly. Our sample included several pediatric patients, which raises the question of whether responses recorded from these participants are comparable to those of adults, as explored in a recent review of the developmental trajectory of voice perception (61). While the primary auditory cortex reaches maturity in childhood (62), the posterior segment of the STG is one of the latest to reach maturity and continues to develop into adolescence (63, 64). Though it is possible that certain regions, especially in the parabelt, were not structurally mature in some of our participants, results from several studies support maturity of perception and representation of auditory categories in childhood. For example, adult-like voice-selective responses have been identified in fMRI studies of children as young as 5 to 8 y old (65–67). Another caveat related to the nature of the recordings is the sparsity of sEEG recordings, both within individual regions and across hemispheres (only a subset of patients had bilateral implantations). This limitation would have the effect of reducing statistical power, and thus positive statistical results can be considered with this in mind. Finally, throughout the paper we have interpreted the DNN layer depth as a measure of representational complexity. While layer depth may seem like a rough proxy for this concept, it is a direct measure of the number of nonlinear transformations the input has undergone (26). Thus, we believe it is a straightforward and readily interpretable index of complexity, especially considering the way it has neatly mapped on to existing models of the auditory cortex.

Table 1. Participant demographic and clinical information

ID	Age (yrs)	Sex	Handedness	Language-dominant hemisphere	Epilepsy onset (years)	Seizure locus	Epilepsy etiology	sEEG laterality	Electrode # (auditory responsive/all)
P1	19	Female	Right	Unknown	14	Right SMA	Lesional	Bilateral	55/106
P2	13	Female	Right	Unknown	3	Bilateral MTL	Unknown	Bilateral	27/104
P3	14	Male	Right	Unknown	13	Tumor	Tumoral	Left	47/54
P4	18	Male	Right	Left	13	Left multifocal	Autoimmune	Bilateral	39/127
P5	14	Female	Left	Unknown	14	Right MTL	Unknown	Right	24/125
P6	15	Male	Right	Unknown	14	Right posterior MTL	FCD	Bilateral	55/117
P7	15	Female	Right	Left	13	Left MTL	MTS	Left	63/119
P8	10	Male	Right	Unknown	8	Bilateral multifocal	Autoimmune	Bilateral	24/120
P9	16	Male	Right	Unknown	0	Right frontal	Unknown	Right	51/226
P10	19	Male	Ambidextrous	Left	17	Left MTG/STG	Tumoral	Left	23/56
P11	16	Male	Left	Right	1	Right parietal + cingulate	FCD	Bilateral	71/256
P12	17	Male	Right	Unknown	14	Right posterior MTL	FCD	Right	34/164
P13	19	Male	Right	Unknown	2	Right STG	Unknown	Right	37/205
P14	16	Male	Left	Unknown	12	Left frontal	Unknown	Bilateral	25/256
P15	22	Male	Right	Right	12	Right multifocal	Unknown	Bilateral	51/243
P16	17	Female	Right	Left	15	Left multifocal	Unknown	Bilateral	44/200
P17	19	Female	Right	Unknown	12	Left MTL	MTS	Left	41/151
P18	16	Male	Ambidextrous	Unknown	12	Left frontal	Unknown	Bilateral	8/218
P19	25	Male	Left	Left	2	Left parietal	Lesional	Left	72/126
P20	13	Female	Right	Unknown	11	Left MTL	Tumoral	Left	20/172

Materials and Methods

Patient-Participants. Data were collected from 20 adult and pediatric patients (age 10 to 25 y old, 7 females) with epilepsy undergoing in-patient stereo-electroencephalography (sEEG) monitoring at UPMC Children's Hospital of Pittsburgh. Each patient had between 6 to 20 electrode trajectories, with each trajectory containing between 8 to 18 electrode contacts, which we refer to as channels. sEEG electrodes were implanted with a robot-assisted frameless stereotactic technique, as previously described (68). All electrode locations were selected based purely on clinical considerations. Further demographic information can be found in Table 1.

The research protocol was approved by the University of Pittsburgh Institutional Review Board (STUDY20030060). All eligible patient-participants were identified from the clinical practice of coauthor TJA. The informed consent process for research took place prior to surgical implantation of electrodes and was separate from the clinical informed consent process for surgery. Patient-participants 18 y of age or older provided written consent for participation in research activities. For patient-participants younger than 18 y, written consent was provided by a parent and written assent was obtained for patient-participants who were 14 to 17 y old.

Data Collection. Patients performed an auditory one-back task while listening to a well-characterized stimulus set of natural sounds (16) consisting of 165 2 s clips of speech (English and non-English), human and animal vocalizations, music, and other environmental sounds (Fig. 1*A*). Environmental sounds consisted of a broad category that included sounds from manmade objects (engine revving, typing, alarm clock, dishes clanking) and nature (wind, thunder, stream). The 165 stimuli were presented in random order, with a random 20% chosen to be followed by an immediate repeat (i.e., a one-back target), and an interstimulus interval that varied randomly from a uniform distribution ranging from 1 to 2 s. This resulted in 198 trials, which were split into two blocks. The process that generated this stimulus order was repeated three times, resulting in six blocks and 594 trials. Patients indicated the presence of a one-back target using a button box (RT Box, model v6). The experiment was run using Psychtoolbox-3 and custom MATLAB code.

Neural data were sampled at 1,000 Hz using a Ripple Grapevine Nomad neural interface processor (model R02000), with line noise notch filters at 60, 120, and 180 Hz. Audio was split using a distribution amplifier (Rolls model DA134), with one stream presented to the patient at a volumed deemed loud but comfortable via Etymotic ER-3C insert earphones, and a second stream recorded synchronously with neural data using a Ripple Digital/Analog IO box (model R02010-0017) and sampled at 30 kHz. Digital stimulus triggers marking the onsets and offsets of each stimulus were sent using a Measurement Computing Data Acquisition device (model USB-1208FS) and were recorded via the IO box.

Preprocessing. Data were common average reference filtered and epoched to include 1,000 ms prestimulus onset to 3,000 ms poststimulus onset. Stimulus onsets were determined via cross-correlation between the original stimulus files and the audio recorded synchronously with neural data. Data in each channel were then z-scored relative to the baseline signal calculated across all trials.

High gamma activity (HGA, 70 to 150 Hz) was then extracted in the following way. For each channel, data were forward- and reverse-filtered (Butterworth, sixth order) in eight different bands, with center frequencies and bandwidths logarithmically spaced between 70 to 150 Hz and 16 to 64 Hz respectively. The analytic signal amplitude for each band was extracted using the Hilbert transform, and the resultant signal was z-scored across all trials relative to a baseline period of −900 to −100 ms relative to stimulus onset. The first and last 100 ms of the 1,000 ms baseline period were excluded to avoid contamination from edge effects and any rapid onset neural responses, respectively. After z-scoring, the signal was averaged across frequency bands and down-sampled to 100 Hz, and then z-scored across trials once more, again using a baseline window of −900 to −100 ms. This measure represents HGA; for each channel, HGA samples that deviated more than five times the interquartile (IQR) range from the median were

labeled as outliers and discarded, where IQR and median were calculated using all timepoints from all trials for a given channel.

Anatomy. The pipeline for determining channel locations and generating anatomy plots consisted of the following steps. For each patient, a cortical surface was reconstructed from a preoperative T1-weighted MRI using Freesurfer (69). The output files were then imported into Brainstorm (70), a third-party tool developed in MATLAB that was used for all subsequent steps in this anatomy pipeline. A postoperative CT was coregistered to the MRI and then used to mark individual channel locations. Nonlinear MNI normalization was then performed in Brainstorm, which internally uses SPM12 for the procedure (71).Finally, the Julich Brain (v3.0) volumetric atlas (72) was imported for ROI analysis, using the MNI reverse field deformation to transform the atlas into the patient-specific spaces. The core/belt/parabelt taxonomy was used with the following definitions: core – Te1.0 and Te1.1; medial belt –TeI; lateral belt – Te1.2, Te2.1, and Te2.2; parabelt – Te3, STS1, and TPJ (37, 73–75). The combination of these regions constituted our definition of the auditory cortex. Due to sparse coverage in medial belt, this region was excluded from most analyses.

To define a channel's precise location within an ROI along anatomically relevant dimensions (e.g., in the analysis for Fig. 3), we calculated a set of three orthogonal axes for each of core, lateral belt, and parabelt (one for each hemisphere, resulting in six total sets of axes). The axes for parabelt were determined using eigendecomposition. Using a three-dimensional point cloud of MNI voxels that belong to parabelt in the Julich atlas, we calculated the eigenvectors of this (demeaned) point cloud's covariance matrix. The eigenvector associated with the largest eigenvalue points along the axis of highest variance, which in this case is oriented from posterior to anterior along the length of STG. The other two eigenvectors point roughly from ventral to dorsal and from medial to lateral. Each axis was then defined as a line pointing along a given direction and centered at the point cloud's centroid. The edges of an axis were defined as 0–1 and were found by projecting the ROI's voxels to the axis and finding the furthest points from the centroid. Eigendecomposition was also used to find the primary axis in core; the eigenvector with the largest eigenvalue points from posteromedial to anterolateral along Heschl's gyrus. The secondary axis was calculated to run across Heschl's gyrus by finding a vector perpendicular to the primary axis and parallel to the STP (defined as Te1.0, Te1.1, Te1.2, Te2.1, Te2.2, TeI, and TI); the tertiary axis was thus perpendicular to STP. In lateral belt, the core axes directions were used, with the origin shifted to the lateral belt centroid. Each channel in a given ROI was then projected to that ROI's axes and normalized using the aforementioned 0 to 1 scaling.

MNI normalization for one patient produced abnormal results due to a prior stroke in the frontal lobe, mapping channels to incorrect locations on the MNI brain and producing incorrect ROI labels (since ROI labeling is dependent on the MNI transformation). For this patient, ROI labels were manually corrected to include in ROI analysis; his channels were not displayed on any MNI brain surface plots.

Encoding Models. Encoding models were used to predict neural responses to novel auditory stimuli. Using the MATLAB package *glmnet* (76, 77), we built L1-regularized regression models for each channel. First, HGA was averaged across all presentations of a stimulus (termed a channel's stimulus response). Instantaneous HGA values were sampled at 40 points per stimulus from $t_i = 50$ to 2,000 ms, sliding by 50 ms. Encoding model inputs consisted of DNN features spanning from t_i–975 ms to t_i (See the following section, *DNN features*, for more information.) Sparse random projection was used to reduce the dimensionality of the input features for computational tractability; this procedure projects a high dimensional matrix into a lower dimensional space while maintaining pairwise relationships between individual observations within some error bound E. The dimensionality of the projected feature matrix was determined using the Johnson–Lindenstrauss bound. This calculation takes the form of $k = 4\ln(N)/E^2$, where k is the new dimensionality, N is the number of observations, and E is the amount of acceptable error (set to 0.1 here). Note that k is not dependent on the dimensionality of the input but rather only depends on the number of observations. With 6,600 observations here (40 timepoints × 165 stimuli), this results in k = 3,518 features after sparse random projection for each DNN layer. The one factor that is dependent on the original feature matrix's dimensionality before projection (call this d) is the density s of the projection matrix, where $s = d^{-0.5}$. In words, a small proportion of elements in the projection matrix were randomly selected, with that proportion equal to $d^{-0.5}$. A random half of these elements were set to −1, with the other half set to 1. The original feature matrix was then multiplied by the projection matrix (size d × k) to generate a reduced dimensionality matrix that was used as the encoding model input. Separate random projections were used for each channel and layer; see *SI Appendix*, Table S8 for the dimensionalities of each DNN layer before projection and the densities of the projection matrices.

Encoding models were fit using a nested cross-validation, with the inner cross-validation (10-fold) used to select the regularization parameter lambda; the outer cross-validation (five-fold) was used to test the resultant model by generating HGA predictions on stimuli that were held out of the training set. These models are referred to as full models to indicate that they were built using all timepoints and to differentiate them from sliding models (see below).

Model accuracy was assessed by calculating the coefficient of determination (R^2) between observed and predicted HGA. This measure describes the fraction of the HGA variance explained by a model and is calculated by summing the squared residuals, dividing by the sum of squared residuals for a null model, and subtracting this ratio from one. We used a null model that predicted the time-varying mean across stimuli (rather than the global mean across stimuli and time) to discount any models that were simply capturing the generic auditory stimulus response. When characterizing which DNN layers produced the best models for each channel, we calculated the weighted mean of the R^2 curve across layers, which we refer to as the weighted DNN layer. Since R^2 can sometimes produce negative values (if a model performs worse than the null model), and a weighted mean requires all weights to be nonnegative, we set any negative R^2 values to zero when calculating the weighted DNN layer.

To investigate the temporal dynamics of encoding, sliding encoding models were built in the following way. For a given channel, we identified the DNN layer that produced its best performing full model (the layer that maximized R^2). Instantaneous HGA was sampled every 20 ms from 0 to 1,000 ms (51 timepoints) and served as the model output. To create the encoding model input features, we first extracted the best layer's node activations for acoustic spectrograms that spanned from −975 to 0 ms relative to each HGA sample. This feature matrix (that included all timepoints) then underwent sparse random projection using the same Johnson–Lindenstrauss bound formula, again with E = 0.1. With 8,415 total observations per channel (51 timepoints × 165 stimuli), each sparse random projected matrix had dimensionality k = 3,615. Next, separate models were built at each individual timepoint (by selecting that timepoint's HGA samples for each stimulus, as well as the relevant rows from the input feature matrix). All other encoding model details were identical to the full model analysis (i.e., L1 regularization, nested cross validation with 10 inner and 5 outer folds). Model performance was assessed separately at each timepoint using R^2, where the null model corresponded to the mean HGA across stimuli at that timepoint.

DNN Features. Stimulus features were generated from the DNN model YAMNet (27), inspired by the success of image classification models such as AlexNet (78). This machine learning model was trained to predict sound classes on thousands of hours of labeled audio taken from AudioSet (28), a large-scale library of sounds scraped from YouTube that have been manually tagged with sound categories of a hierarchical ontology of classes. For example, a given sound might be simultaneously tagged as Music, Soul music, Singing, and Female singing. Note that this DNN model was built and trained by an unaffiliated machine learning research team, was held static in this research (i.e., no further training was performed), and did not have access to any neural data, including the data used in this study.

The YAMNet model consists of 16 layers (depicted in Fig. 1*C*), starting with an input layer 0, which accepts mel-spectrograms of sounds of width 975 ms, and layer 1 that performs standard 2D convolutions. Layers 2 to 14 are depthwise separable convolutional layers, where each layer consists of two convolutions: first, a grouped depthwise convolution is performed, where each filter is associated with and learned on a single channel (here, channel refers to the third dimension, or depth, in each DNN layer). The second convolution in a depthwise separable layer is a pointwise convolution, where each filter is 1 × 1 × number of channels, allowing the layer to mix information across channels. Each individual convolution is followed by a batch normalization and a rectified linear unit (ReLU). The final layer consists of a global average pooling layer, a fully connected layer, and a softmax output that generates a probability distribution over the discrete sound classes.

For each of the 165 stimuli, 40 mel-spectrograms (width: 975 ms, leading edge: t_i = 50:50:2,000 ms) were extracted; for samples where t_i was less than 975 ms, the stimulus waveform was zero-padded as necessary. These spectrograms were provided as input to the YAMNet model, and activations across all nodes within a given layer were calculated. These layer activations served as stimulus features; repeating this process for all 16 layers thus produced 16 unique feature sets, numbered 0 (the input spectrogram) to 15 (the probability that the sound belongs to each semantic class).

The YAMNet model has not been trained to mimic human neural representations of sound features. Rather, the training objective function dictates that the model iteratively adjusts its weights to solve the task while minimizing the cross-entropy, a measure of the distance between the output class probabilities and the ground truth labels. Any similarities that arise between internal model representations and human neural responses are thus an emergent property of the model. The DNN learns to extract an optimal set of stimulus features for audio classification and can thus be viewed as a tool for data-driven feature extraction.

Furthermore, the model imposes a natural and interpretable gradient from early layers that represent low-level acoustic properties, through middle layers with more complex acoustic representations, to the deepest layers that represent wholly abstract stimulus features (i.e., the sound category). Each layer performs a nonlinear transformation, introducing further complexity in the stimulus representation with increasing layer depth. We can leverage these evolving stimulus representations to quantify the degree of representational complexity present at different cortical sites by identifying the DNN layers that are best able to explain the neural responses. Theoretically, an encoding model that can predict neural responses in the human auditory cortex with high fidelity using the activations (i.e., the responses) within a YAMNet hidden layer would demonstrate that there is shared information between the two systems.

Data, Materials, and Software Availability. Given the sensitive nature of patient data and the associated requirements for maintaining anonymity, the data cannot be shared publicly but will be made available upon request to the corresponding author.

ACKNOWLEDGMENTS. This work was funded by 1F30DC021342-01 awarded to J.L.H., R01MH132225-02 to ASG, R21DC019217-01A1 to T.J.A. and L.L.H., and R01DC013315-07 to T.J.A. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author affiliations: [a]Department of Neurological Surgery, University of Pittsburgh, PA 15213; [b]Department of Psychology, The University of Texas at Austin, TX 78712; and [c]Department of Bioengineering, University of Pittsburgh, PA 15261

1. T. R. Agus, S. Paquette, C. Suied, D. Pressnitzer, P. Belin, Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Sci. Rep.* **7**, 11526 (2017).
2. S. V. Norman-Haignere, J. H. McDermott, Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLOS Biol.* **16**, e2005127 (2018).
3. B. L. Giordano, S. McAdams, R. J. Zatorre, N. Kriegeskorte, P. Belin, Abstract encoding of auditory objects in cortical activity patterns. *Cereb. Cortex* **23**, 2025–2037 (2013).
4. M. Staib, S. Frühholz, Cortical voice processing is grounded in elementary sound analyses for vocalization relevant sound patterns. *Prog. Neurobiol.* **200**, 101982 (2021).
5. R. Santoro *et al.*, Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLOS Comput. Biol.* **10**, e1003412 (2014).
6. D. A. Depireux, J. Z. Simon, D. J. Klein, S. A. Shamma, Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J. Neurophysiol.* **85**, 1220–1234 (2001).
7. C. Humphries, E. Liebenthal, J. R. Binder, Tonotopic organization of human auditory cortex. *NeuroImage* **50**, 1202–1211 (2010).
8. L. M. Miller, M. A. Escabí, H. L. Read, C. E. Schreiner, Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J. Neurophysiol.* **87**, 516–527 (2002).
9. P. Albouy, L. Benjamin, B. Morillon, R. J. Zatorre, Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* **367**, 1043–1047 (2020).
10. P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, E. F. Chang, Human Superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci. Off. J. Soc. Neurosci.* **36**, 2014–2026 (2016).
11. H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 874 (2019).
12. R. Santoro *et al.*, Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4799–4804 (2017).
13. N. C. Singh, F. E. Theunissen, Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **114**, 3394–3411 (2003).
14. J. H. McDermott, M. Schemitsch, E. P. Simoncelli, Summary statistics in auditory perception. *Nat. Neurosci.* **16**, 493–498 (2013).
15. B. L. Giordano, M. Esposito, G. Valente, E. Formisano, Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nat. Neurosci.* **26**, 664–672 (2023).
16. S. Norman-Haignere, N. G. Kanwisher, J. H. McDermott, Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88**, 1281–1296 (2015).
17. M. Keshishian *et al.*, Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife* **9**, e53445 (2020).
18. J. Berezutskaya, Z. V. Freudenburg, U. Güçlü, M. A. J. van Gerven, N. F. Ramsey, Brain-optimized extraction of complex sound features that drive continuous auditory perception. *PLOS Comput. Biol.* **16**, e1007992 (2020).
19. K. Patil, D. Pressnitzer, S. Shamma, M. Elhilali, Music in our ears: The biological bases of musical timbre perception. *PLOS Comput. Biol.* **8**, e1002759 (2012).
20. M. Schönwiesner, R. J. Zatorre, Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 14611–14616 (2009).
21. P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, B. Pike, Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312 (2000).
22. Y. Li *et al.*, Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* **26**, 2213–2225 (2023).
23. M. Khosla, G. H. Ngo, K. Jamison, A. Kuceyeski, M. R. Sabuncu, Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Sci. Adv.* **7**, eabe7547 (2021).
24. A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).
25. G. Tuckute, J. Feather, D. Boebinger, J. H. McDermott, Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biol.* **21**, e3002366 (2023).
26. L. Reddy, R. M. Cichy, R. VanRullen, Representational content of oscillatory brain activity during object recognition: Contrasting cortical and deep neural network hierarchies. *eNeuro* **8**, 0362 (2021).
27. S. Hershey *et al.*, "CNN architectures for large-scale audio classification" in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP, 2017), pp. 131–135.
28. J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events" in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP, 2017), pp. 776–780.
29. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
30. K. V. Nourski, M. Steinschneider, H. Oya, H. Kawasaki, M. A. Howard, Modulation of response patterns in human auditory cortex during a target detection task: An intracranial electrophysiology study. *Int. J. Psychophysiol.* **95**, 191–201 (2015).
31. K. V. Nourski *et al.*, Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *NeuroImage* **101**, 598–609 (2014).
32. T. A. Hackett, T. M. Preuss, J. H. Kaas, Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. *J. Comp. Neurol.* **441**, 197–222 (2001).
33. P. Schneider *et al.*, Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nat. Neurosci.* **8**, 1241–1247 (2005).
34. A. M. Leaver, J. P. Rauschecker, Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *J. Neurosci.* **30**, 7604–7612 (2010).
35. C. Wessinger *et al.*, Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *J. Cogn. Neurosci.* **13**, 1–7 (2001).
36. M. M. Merzenich, J. F. Brugge, Representation of the cochlear partition on the superior temporal plane of the macaque monkey. *Brain Res.* **50**, 275–296 (1973).
37. M. Moerel, F. De Martino, E. Formisano, An anatomical and functional topography of human auditory cortical areas. *Front. Neurosci.* **8**, 225 (2014).
38. M. N. Wallace, P. W. Johnston, A. R. Palmer, Histochemical identification of cortical areas in the auditory region of the human brain. *Exp. Brain Res.* **143**, 499–508 (2002).
39. J. P. Rauschecker, B. Tian, M. Hauser, Processing of complex sounds in the macaque nonprimary auditory cortex. *Sci. Wash.* **268**, 111 (1995).
40. M. Chevillet, M. Riesenhuber, J. P. Rauschecker, Functional correlates of the anterolateral processing hierarchy in human auditory cortex. *J. Neurosci.* **31**, 9345–9352 (2011).
41. C. I. Petkov, C. Kayser, M. Augath, N. K. Logothetis, Functional imaging reveals numerous fields in the monkey auditory cortex. *PLOS Biol.* **4**, e215 (2006).
42. H. Blank, N. Wieland, K. von Kriegstein, Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neurosci. Biobehav. Rev.* **47**, 717–734 (2014).
43. K. von Kriegstein, D. R. R. Smith, R. D. Patterson, D. T. Ives, T. D. Griffiths, Neural representation of auditory size in the human voice and in sounds from other resonant sources. *Curr. Biol.* **17**, 1123–1128 (2007).
44. K. von Kriegstein, D. R. R. Smith, R. D. Patterson, S. J. Kiebel, T. D. Griffiths, How the human brain recognizes speech in the context of changing speakers. *J. Neurosci.* **30**, 629–638 (2010).
45. K. von Kriegstein, E. Eger, A. Kleinschmidt, A. L. Giraud, Modulation of neural responses to speech by directing attention to voices or verbal content. *Cogn. Brain Res.* **17**, 48–55 (2003).
46. E. Formisano, F. D. Martino, M. Bonte, R. Goebel, "Who" is saying "What"? brain-based decoding of human voice and speech *Science* **322**, 970–973 (2008).
47. P. Belin, R. J. Zatorre, Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport* **14**, 2105 (2003).
48. C. R. Pernet *et al.*, The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage* **119**, 164–174 (2015).
49. C. Bodin *et al.*, Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques. *Curr. Biol.* **31**, 4839–4844.e4 (2021), 10.1016/j.cub.2021.08.043.

50. I. DeWitt, J. P. Rauschecker, Phoneme and word recognition in the auditory ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E505–E514 (2012).
51. J. P. Rauschecker, S. K. Scott, Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat. Neurosci.* **12**, 718–724 (2009).
52. S. K. Scott, C. C. Blank, S. Rosen, R. J. S. Wise, Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* **123**, 2400–2406 (2000).
53. E. Formisano *et al.*, Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron* **40**, 859–869 (2003).
54. Y. Kajikawa, L. de La Mothe, S. Blumell, T. A. Hackett, A comparison of neuron response properties in areas A1 and CM of the Marmoset monkey auditory cortex: Tones and broadband noise. *J. Neurophysiol.* **93**, 22–34 (2005).
55. M. Moerel, F. D. Martino, E. Formisano, Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* **32**, 14205–14216 (2012).
56. B. H. Scott, B. J. Malone, M. N. Semple, Transformation of temporal processing across auditory cortex of awake macaques. *J. Neurophysiol.* **105**, 712–730 (2011).
57. M. Moerel *et al.*, Processing of natural sounds: Characterization of multipeak spectral tuning in human auditory cortex. *J. Neurosci.* **33**, 11888–11898 (2013).
58. S. V. Norman-Haignere *et al.*, Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nat. Hum. Behav.* **6**, 455–469 (2022), 10.1038/s41562-021-01261-y.
59. T. Overath, J. H. McDermott, J. M. Zarate, D. Poeppel, The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* **18**, 903–911 (2015).
60. M. López Espejo, S. V. David, A sparse code for natural sound context in auditory cortex. *Curr. Res. Neurobiol.* **6**, 100118 (2024).
61. E. E. Harford, L. L. Holt, T. J. Abel, Unveiling the development of human voice perception: Neurobiological mechanisms and pathophysiology. *Curr. Res. Neurobiol.* **6**, 100127 (2024).
62. J. K. Moore, Y.-L. Guan, Cytoarchitectural and axonal maturation in human auditory cortex. *J. Assoc. Res. Otolaryngol.* **2**, 297–311 (2001).
63. J. N. Giedd *et al.*, Brain development during childhood and adolescence: A longitudinal MRI study. *Nat. Neurosci.* **2**, 861–863 (1999).
64. N. Gogtay *et al.*, Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8174–8179 (2004).
65. D. A. Abrams *et al.*, Impaired voice processing in reward and salience circuits predicts social communication in children with autism. *eLife* **8**, e39906 (2019).
66. M. Bonte *et al.*, Development from childhood to adulthood increases morphological and functional inter-individual variability in the right superior temporal cortex. *NeuroImage* **83**, 739–750 (2013).
67. N. M. Raschle *et al.*, Investigating the neural correlates of voice versus speech-sound directed information in pre-school children. *PLOS ONE* **9**, e115549 (2014).
68. T. J. Abel *et al.*, Frameless robot-assisted stereoelectroencephalography in children: Technical aspects and comparison with Talairach frame technique. *J. Neurosurg. Pediatr.* **22**, 37–46 (2018).
69. B. Fischl *et al.*, Automatically parcellating the human cerebral cortex. *Cereb. Cortex* **14**, 11–22 (2004).
70. F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, R. M. Leahy, Brainstorm: A user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* **8**, 1–8 (2011).
71. S. B. Eickhoff *et al.*, A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* **25**, 1325–1335 (2005).
72. K. Amunts, H. Mohlberg, S. Bludau, K. Zilles, Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. *Science* **369**, 988–992 (2020).
73. P. Morosan *et al.*, Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage* **13**, 684–701 (2001).
74. P. Morosan, A. Schleicher, K. Amunts, K. Zilles, Multimodal architectonic mapping of human superior temporal gyrus. *Anat. Embryol. (Berl.)* **210**, 401–406 (2005).
75. D. Zachlod *et al.*, Four new cytoarchitectonic areas surrounding the primary and early auditory cortex in human brains. *Cortex* **128**, 1–21 (2020).
76. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
77. J. K. Tay, B. Narasimhan, T. Hastie, Elastic net regularization paths for all generalized linear models. *J. Stat. Softw.* **106**, 1–31 (2023).
78. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Proc. Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).