Emergence of a time-independent population code in auditory cortex enables sound categorization and discrimination learning

S. Bagur[1,§,*], J. Bourg[1,*], A. Kempf[1], T. Tarpin[1], K. Bergaoui[1], Y. Guo[1], S. Ceballo[1], J. Schwenkgrub[1], J.L. Puel[2], J. Bourien[2], B. Bathellier[1,§]

[1] Institut Pasteur, Université Paris-Cité, INSERM, Institut de l'Audition, 63 rue de Charenton, F-75012 Paris, France.
[2] Institut des Neurosciences de Montpellier, Univ Montpellier, INSERM, Montpellier, France.
* equal contribution
§ Corresponding authors : brice.bathellier@pasteur.fr, sophie.bagur@pasteur.fr

Summary

Perception generates time-invariant objects and categories from time-varying streams of information. However, individual neuron responses, even in cortex, are not time-invariant as they usually track the temporal variations of the input. Here we show that representations of time-varying sounds remain decodable even after time-averaging at the level of neuronal populations in the mouse auditory cortex. This population-scale, time-invariant property is absent in subcortical auditory regions. By implanting light-sculpted artificial representations in the cortex with optogenetics, we show that robustness to time-averaging is a necessary property for rapid association of neural representations with behavioral output. Moreover, deep neural networks which perform sound recognition and categorization tasks generate population representations that become robust to time-averaging in their deeper layers. Hence, the auditory cortex implements a generic transformation that replicates temporal information into time-independent neural population dimensions and makes it available for learning and classification.

Keywords

Sensory processing, auditory system, auditory cortex, invariant representations, population codes, discrimination learning, category learning.

Introduction

A long standing idea in sensory processing is that object identification requires multiple features to be bound together. In hearing, the features defining a particular object include not only the sound frequencies extracted in the cochlea but also a variety of temporal modulations of sound intensity and frequency[1] that have a high prevalence in natural sounds [2,3]. Temporal modulations and in particular the direction of change (i.e. rising vs falling frequency or intensity) contribute to sound recognition and to global perceptual properties such as timbre [4] or loudness [5,6]. This can be directly experienced when listening to time-reversed versions of common sounds from which it can be difficult to recognize the original [4,7,8]. Beyond hearing, temporal variations are also key features in touch, in which object recognition is based on contact sequences [9], in olfaction, in which smells sequentially activate olfactory receptors [10] or in the visual identification of actions.

At the neurophysiological level, time-varying sounds produce temporal activity sequences throughout the auditory system including the auditory cortex. In an individual auditory neuron, these temporally structured firing patterns provide specific information about sound identity which is not conveyed by the neuron's mean firing rate [11,12]. Therefore, individual auditory system neurons do not represent sounds in a time-independent manner, raising the question of whether a neuronal correlate of time-invariant auditory object perception exists in the auditory system. Some studies have demonstrated the existence of specific neuronal activations to particular temporal features [13,14] or to the direction of temporal modulations [15,16] in the auditory cortex, but also as early as the auditory midbrain or even in the brainstem [14,17–19]. This led to the common view that the tuning of subcortical neurons provides all the basic building-blocks to construct auditory objects [20]. However, cortical inactivation experiments during sound discrimination behaviors strikingly contrast with this view. Indeed, they show that whereas auditory cortex is dispensable for simple tone frequency discriminations[21,22] it is necessary for discriminating even basic temporal features including sound duration [23,24] or frequency modulations [21,22,25]. These results point towards a specific and unidentified transformation of temporal feature representations in the auditory cortex that enables discrimination.

To isolate this transformation, we reasoned that population-scale measurements are less sensitive to experimental variability and sampling biases than quantifications on single cells. We therefore combined, in a large-scale effort, temporally deconvolved calcium imaging and single-unit electrophysiology in the awake mouse with detailed biophysical modeling of the cochlea to extensively and consistently sample responses to a wide range of spectral and temporal sound features in the auditory nerve, inferior colliculus, auditory thalamus and auditory cortex. Using a noise-corrected metric to measure representation distances, we established that the distinctive property of the auditory cortex is that population representations of sounds remain similar whether or not the temporal details of neuronal responses are removed by time-averaging. Remarkably, this form of time-invariance provided by cortical representations does not reflect a rate coding scheme in individual neurons whose responses are temporally structured. Rather, time-invariant representations are an emergent property of the neuronal population.

Combining a reinforcement learning model and behavioral discrimination of optogenetically engineered activity patterns in the cortex, we established that the speed at which the discrimination of two sounds can be learnt in a task is principally determined by the time-averaged representations of these two sounds and not by their temporal sequence representations. Hence, the robustness of cortical representations to time-averaging has a direct functional impact on the association of different time-varying sounds to various behaviors. In addition, this accounts for the results of cortical inactivation experiments [21,22,25]. Finally, deep networks performing sound categorization implement representations that are robust to time-averaging in their deeper layers. Moreover, representations were more similar between the auditory system and artificial networks for networks performing sound categorization than for networks performing other tasks. Together these results show that the emergence of time-independent representations is a key cortical computation that enables efficient association of temporally structured sounds to behavior and their categorization as auditory objects.

Results

Emergence of time-independent representations in the auditory cortex

In order to precisely measure how the representations of core auditory features evolve across the auditory hierarchy, we performed large-scale recordings in three successive regions: the inferior colliculus (IC), the auditory thalamus (TH) and the auditory cortex (AC) (**Fig. 1A-G, Supplemental Table 1**). In each region, we measured the responses to a set of 140 sounds, mainly of 500 ms duration, which were chosen to cover simple, widely studied spectral and temporal features, including amplitude and frequency modulations (**Fig 1B, Supplemental Table 2**).

To rapidly obtain large datasets from these structures, we used GCAMP6s-based two-photon calcium imaging of either cell bodies (AC and IC, **Fig. 1C & F**) or axonal projections (TH, imaged in AC) (**Fig. 1D**). Collecting data simultaneously from around 1000 AC neurons or TH axonal boutons and from 100 to 200 neurons in IC, we could extensively sample representations in each region (**Fig. 1A-F**). In AC, all 60.822 ROIs were mapped to functional subfields based on tonotopic gradients [28] and to the cortical layer from imaging depth (**Fig. S1A-F).** 70% of ROIs were in primary auditory cortex (A1), the largest subfield of AC, but the anterior, suprarhinal and dorsal posterior auditory fields were also covered **(Fig. 1C & Fig. S1E**). Moreover, with recording depth reaching up to 600 µm, we sampled neurons from layers 1 to 5 with an emphasis on layers 2 and 3 (**Fig. S1F**). Therefore, with the exception of layer 6 and of the small ventro-posterior subfield, the whole of primary and secondary AC was extensively covered with a total number of neurons of about one fifth the estimated number per hemisphere [29]. Inputs from TH were sampled with 39.191 putative TH axonal boutons spread across AC (75% of ROIs in A1) (**Fig. 1D**) and validated post-hoc with the thalamic marker VGLUt2 (**Fig. S1G,H**)[30]. In addition, we recorded 15.132 ROIs in the dorsal IC down to 250µm depth (**Fig. 1F**).

Calcium signals were temporally deconvolved using a linear algorithm to retrieve estimates of neuronal firing rate variations that are robust to parametrization errors [31] and previously verified in cortical neurons [32]. This allowed us to reach a ~150 ms temporal precision as estimated from responses to amplitude modulated sounds (**Fig. 1C,D,F**). The temporal modulations of our sounds were chosen to evolve at timescales compatible with this resolution of calcium imaging.

Since deconvolution has not been verified for TH axons, we performed electrophysiological recording in primary and secondary auditory thalamus (498 single units, **Fig. 1E**). Electrophysiology was also used to cover the central inferior colliculus (563 single units), the main primary subregion of this structure [33] (**Fig. 1G**). Electrode locations were identified with post-hoc histology and short-latency responses (**Fig. 1E,G**). Finally, we used a detailed biophysical model of the cochlea calibrated against auditory nerve recordings[34] (AN), to provide insight into the information entering the auditory system (**Fig 1H, Fig S1I,J**).

Based on this rich dataset, we first measured classical, single cell, feature-tuning indexes, including preference to frequency or intensity modulation direction (e.g. **Fig. 2A**), speed and frequency. Consistent with previous reports [14,18,20] these measures indicated that tuning to all

these features is weak in the AN but then appears as early as IC (**Fig. S2**). They did not evidence any further evolution of tuning strength along the auditory hierarchy.

We reasoned that neuronal variability and measurement noise may impact single cell measurements and obscure changes in encoding, given that visual inspection of sample neurons suggested a higher specificity of population patterns in the cortex than subcortical stages (e.g. **Fig. 2B**). Inspired by recent reports that population scale measures efficiently circumvent noise-related biases[35], we used a noise-corrected population measure to systematically compare sound representations between areas. This metric quantifies the similarity between population vectors evoked by a pair of sounds by calculating the Pearson correlation between the two (**Fig. 2A**). Correlation typically decreases when data is corrupted by variability (**Fig. 2C**). By exploiting population vectors sampled from multiple single-trials and in the limit of a large neuron number, a simple formula allows us to provide an unbiased estimate of the correlation in absence of variability, as we verified analytically and by simulations (**Supplemental Mathematical Derivations, Fig. 2C**)[36]. This noise correction enabled us to compare datasets with widely different variability levels (**Fig. 2D**). Applying the noise-corrected correlation metric to the population representations of all pairs of sounds, we constructed Representational Similarity Analysis (RSA)[37] matrices that capture the relations between all sounds in the space of neural activity (**Fig. 2E**).

Sounds are encoded in neural activity along the temporal dimension (*when* neurons are active) and neural population dimensions (*which* neurons are active). To identify their relative contributions, we calculated noise-corrected RSA matrices based on these two encoding strategies (**Fig. 2E**). The first one takes into account the full sequence of activity observed in the neuronal population during and immediately after sound presentation (sequence code, **Fig. 2A**). The second one evaluates the information that can be retrieved solely from the time-independent activity level of neurons by time-averaging neuronal responses (time-averaged code, **Fig. 2A**). Low correlations between two sounds for the sequence code indicate that they are coded by fairly different patterns of temporal activity. Low correlations for the time-averaged code indicate that the two sounds activate different neural populations, irrespective of the sequence of activity, making information available in a time-independent manner (see examples **Fig. 2A,B**).

These noise-corrected RSA matrices shown in **Fig. 2E** capture multiple aspects of how sound representation evolves throughout the auditory system. Contrary to sound feature tuning indexes (**Fig. S2**), these measures clearly delineated robust changes of representations across stages. First, overall pattern similarity levels decreased from AN to IC for both sequence and time-averaged codes, indicating a sharpening of population tuning in the brainstem (**Fig. 2F**). Second, population response similarity increased in the TH before decreasing again in AC (**Fig. 2F**). This surprising non-monotonic evolution of tuning sharpness has never been reported and corresponds to a densification of the representation in TH that can be quantified with sparseness measures (**Fig. S3A,B**).

Strikingly, the AC displayed strong decorrelation of the time-averaged code relative to all prior areas, leading to a value very close to that of the sequence code (**Fig. 2F**). This unique convergence of sequence and time-averaged codes is observed both in the mean RSA correlation values (**Fig. 2G**) and in the structure of the RSA matrices (**Fig. 2H**). It is also confirmed by the similar accuracy of population decoders trained and tested with the sequence

or time-averaged code in AC (**Fig. 2I, S3C**). All these metrics indicate that the information present in the full temporal sequence of activity is still largely accessible after time-averaging in the AC but not in IC and TH. These results hold in all subfields of AC (**Fig. 2F**) and are robust to neuron number (**Fig. S3D,E**). Interestingly, the dorsal IC which receives cortical feedback shows an intermediary profile, more similar to AC than central IC (**Fig. 2F**).

A possible explanation of these results could be a loss of temporal resolution along the auditory hierarchy, resulting in a less-informative, because less-resolved, sequence code [38]. Two observations demonstrate that our results strongly differ from this scenario. First, decoding accuracy does not decrease for the sequence code, while it increases for the time-averaged code from IC to AC (**Fig. S3C**). Second, we decomposed neuronal responses into Fourier components which capture the information content at specific timescales. We observed that, in AC, accuracy is already very close to plateau value at 0Hz (time-averaged activity level), whereas in subcortical areas it increases when adding faster timescales (**Fig. S3F-H**). Therefore, AC, contrary to earlier stages, implements a time-independent representation of sounds at neural population scale, which emerges without loss of the temporal information contained in the time-sequences of neuronal responses. This could be seen as a hybrid coding scheme in which temporal information is made available along neuronal dimensions.

Auditory cortex specifically separates the time-averaged representations of time-varying sounds

To better understand the convergence of sequence and time-averaged codes in AC, we quantified representation similarity for particular groups of sounds as a measure of population tuning to particular features. For example, averaging correlations across pairs of pure tones for specific frequency intervals allowed building population tuning curves for frequency (**Fig. 3A,B**). This showed that frequency tuning is sharper in IC and AC than in AN and TH (**Fig. 3C**). This is in line with the overall densification of the auditory code observed in TH since sharp tuning corresponds to a sparse code and broad tuning to a dense code (**Fig. 2F & S3A,B)**. We also quantified population intensity tuning and observed the same level of correlation in AC and IC between representations of pure tones differing in intensity (**Fig. 3D**). This is in agreement with previous descriptions of single neuron intensity tuning both in IC and AC [16,39,40]. Hence, for simple tones, neither intensity nor frequency tuning are sharpened between IC and AC. Moreover, for these stationary sounds, sequence and time-averaged codes provide the same levels of correlation (**Fig. 3C,D**). This indicates that activity sequence does not play a role in coding frequency or intensity of stationary pure tones which therefore do not contribute to the convergence of time-averaged and sequence representations in AC.

In contrast, population representations of time-varying sounds are changed in the cortex. Most strikingly, the correlation of time-averaged representations of time-symmetric sounds drops specifically in the cortex compared to subcortical structures, although the imprecision of activity measurements in thalamic axons weakens the conclusion for intensity ramps (**Fig 3E,F)**. Therefore, while AC does not improve frequency and intensity tuning, it clearly sharpens population tuning to the direction of modulations. This appears as an important driver of the convergence of time-averaged and sequence codes, but other temporal aspects also contribute. Similar but smaller effects are observed for non-directional features such as sinusoidal amplitude modulations or frequency sweeps differing in speed (**Fig. S3J,K**). Time-averaged representations of frequency-modulated sweeps differing only in intensity are also

more decorrelated in AC (**Fig. S3I**) in contrast to pure tones of different intensity. This suggests an interaction in the coding of frequency modulations and intensity, which may relate to perceptual observations made in humans[6].

Overall, these measurements demonstrate that the key transformation of sound representations from the subcortical to the cortical stage is the decorrelation of time-averaged representations of sounds which differ by their temporal variations.

Time-averaged representations determine associative learning speed

We therefore interrogated the possible functional advantages provided by decorrelated time-averaged representations. To associate a sound to a rewarding or defensive action, it is necessary to associate its neuronal representation to motor circuits by specific synapses. If two sounds have representations that differ only by activity sequences and not by the pattern of neurons they recruit, one intuitively expects that simple synaptic plasticity mechanisms that are local in time will not allow discriminative associations with these two sounds. Hence, we reasoned that high correlations for time-averaged representations should impair discriminative learning.

To quantify this idea, we upgraded a previously published feedforward neural network model of auditory discrimination learning [41,42] with synaptic learning rules including both Hebbian plasticity and an eligibility trace mechanism previously described in the mouse striatum (**Fig. 4A**). Striatum was chosen as it is the key site of auditory reinforcement learning [43,44], but our conclusions depend little on the specific learning rule. We trained the model to discriminate between the population responses to pairs of sounds taken from the AC, TH or IC datasets. This allowed us to measure learning duration for a broad range of time-averaged and sequence correlation values (**Fig. 4A**). In line with our intuition, time-averaged correlation and not sequence correlation predicted the duration of discrimination learning (**Fig. 4B)**. Moreover, we observed that learning duration steeply rises with increasing correlation of the time-averaged representations, following a monotonic, but strongly non-linear relationship (**Fig. 4C**).

To directly test the importance of time-averaged representations for discriminative learning and evaluate the predictions of the model, we trained mice expressing ChR2 in cortical pyramidal cells to discriminate between different spatio-temporal optogenetic stimulations, patterned at the mesoscopic level in the AC (**Fig. 4D-G, FigS4A-C**). This strategy allowed us to control neural activity patterns, reliably positioned in identified tonotopic fields of AC (**Fig. 4E, S4A**), and to evaluate how encoding strategy influences learning in a Go-Nogo discrimination task. We compared, in the same mice, the learning duration for representations that differ by the identity of the active neurons (low correlations for time-averaged and sequence codes) with that for representations that differ only by the sequence of active neurons (low correlation for the sequence code but high correlation for the time-averaged code). Optogenetic stimuli were of the same 500 ms duration as the previously studied sounds. Concretely, during the time-independent task, mice had to discriminate the optogenetic activation of two spatially separate spots, A vs B. During the sequence task they had to discriminate the successive activation of spot A then spot B against the time-symmetric sequence (A-B vs B-A, **Fig. 4F**). Task order was counterbalanced across mice. Incorrect licks to the Nogo stimulus were punished by a time-out and correct licks to the Go stimulus were

rewarded. Rewards were provided by an intracranial stimulation of the medial forebrain bundle (**Fig. 4G**), a protocol that yields similar learning curves to water rewards in deprived animals[45].

In line with our model, mice learnt the time-independent discrimination much faster than the sequence discrimination for which only a few mice succeeded to perform above chance level after several thousands of trials (**Fig. 4H,I, Fig. S4D,E**). This therefore corroborates the proposition that decorrelation of time-averaged representations in cortex is crucial to accelerate discriminative learning, in particular of time-symmetric sounds.

Interestingly, this proposition also provides an explanation of why AC's involvement in sound discrimination depends on the pair of sounds that is discriminated, in particular for time-symmetric sounds[21,22,25]. Indeed, time-averaged representations of time-symmetric sounds are highly correlated subcortically (>0.9), and clearly less in the cortex (0.74, **Fig. 3E,F**). The non-linear relationship between correlation and learning speed in our model predicts a ~3 fold decrease in learning duration with cortical compared to subcortical representations (**Fig. 4C**). By contrast, both in cortex and subcortically, the correlation between representations of pure tones of different frequencies is below 0.75 (**Fig. 3C**). For this range of correlation values, learning occurs quickly and the impact of representation similarity on learning speed is marginal (**Fig. 4C**). Our model therefore predicts that cortical lesions performed before discrimination training will dramatically increase learning duration for time-symmetric sounds but not for pure tones, as observed experimentally [22]. In the intact brain, cortical and subcortical representations may compete for associations with decisions. In this case, their roles will depend on how fast discriminative associations are learnt, and therefore will crucially depend on the correlation of time-averaged representations. Based on this assumption, the model predicts that post-training lesions of the AC have a much stronger impact on discrimination of time-symmetric sounds than of distant pure tones, as also observed experimentally [21,46,47].

Convergence of time-averaged and sequence representations in deep neural networks for sound categorization

Our results so far indicate that time-independent population representations as observed in cortex are important for associating specific sounds to a binary behavioral output. However in natural situations, sound-driven behaviors rely on multiple associations with broad stimulus categories or auditory objects. We therefore hypothesized that models which produce complex stimulus categories also implement a convergence between both types of representation.

To test this, we first analyzed the responses from a previously published convolutional neural network (CNN), whose time-averaged representations were shown to be similar to human AC representations measured by functional magnetic resonance imaging [26]. This network robustly identifies a wide range of words and music styles using a two-branch architecture with one word and one music branch (**Fig. 5A**). In line with our hypothesis, we observed that this network generated decorrelated time-averaged representations and convergent sequence and time-averaged representations when reaching deeper layers. However, convergence was only observed for the range of stimuli categorized by a specific branch. For example, only music sounds and not words had convergent time-averaged and sequence representation in the branch dedicated to music, and vice versa (**Fig. S5A**). Moreover, this CNN did not implement convergent time-averaged and sequence representations for the 140 simple sounds that we played to mice, transposed to match the frequency range of words and music

(**Fig. 5A**). This indicates that the emergence of time-independent representations in deep networks relates tightly to the sound training set and the target stimulus categorization.

We therefore investigated CNNs categorizing key features of the stimuli presented to our mice: the frequency and intensity range, and the type of frequency and amplitude modulations present in the sounds (**Fig. 5B, S5B,C**). Networks were trained on this multicategorization task with an augmented set of sounds that homogeneously covered all these features and combinations thereof. Sounds were embedded in natural noise from various backgrounds to complexify categorization. We observed that time-averaged and sequence-based representations also converged in deep layers of this network after training (**Fig. 5B**), but not in untrained networks (**Fig. S5D**). This corroborates the observations for the word and music categorization task, now for the same stimuli as those used to probe the mouse auditory system.

Typical CNNs are designed to reduce the precision of sensory receptive fields in deeper layers, thereby reducing the number of parameters to fit in the model. In our case, this leads to a shrinkage of the temporal dimension which forces the sequence and time-averaged code to converge. However, if we implemented the same sound feature categorization task in CNNs which did not shrink the temporal dimension across layers, we still observed a clear convergence of the two coding strategies (**Fig. 5C**). The main effect of temporal shrinking in our simulations was to accelerate learning (**Fig. 5C, Fig. S5B,C**). This demonstrates that convergence of sequence and time-averaged codes is not the consequence of structural constraints but rather the consequence of the computations performed by the network, in particular the fact that sounds are assigned to specific labels. This idea is corroborated by the observation that networks performing single sound identification (assigning one label per sound) also implemented a convergence of time-averaged and sequence codes (**Fig. 5D, S5E**). To compare with networks that do not perform labeling, we trained autoencoders which must reconstruct the denoised input stimulus through a small central bottleneck (**Fig. S5F)**. This network did not show convergence between the two coding strategies (**Fig. 5E**).

In addition, representations of the categorization network qualitatively reproduced all aspects of the convergence of time-averaged and sequence codes observed in the auditory system (compare **Fig. 5F-H** and **Fig. 2G-I**). In particular, like in cortex, we observed an absence of time resolution loss in the deeper layers of the artificial networks, especially when the architecture preserves time resolution (compare **Figs. S5G,H** and **Figs S3G,H**). This underlines the computational homology between the transformations observed in categorizing deep convolutional networks and in the mouse auditory system.

Signatures of task-driven categorization in the geometry of auditory representations

We next investigated whether the geometry of sound representations could be further used to probe the underlying perceptual tasks that the mouse auditory system performs. This is difficult to determine in the absence of subjective experience and limited ethological surveys of mouse sound perception. We therefore systematically compared RSA matrices of CNNs trained on all previously described tasks with RSA matrices measured in the auditory system (**Fig. 6A-E**), reasoning that the structure of the RSA matrices may reflect the categorization task. We first observed that early layers of the auditory system have representations that largely differ from any of the CNNs (**Fig. 6E, Fig. S5I**). This indicates that these CNNs poorly emulate computations that occur in the early stages of the auditory system. Similarity of RSA matrices

between CNNs and the mouse auditory system increased when considering deeper structures and layers (**Fig. 6E, Fig. S5I**). The task leading to highest similarity with IC, TH and AC was the multi-categorization task (**Fig. 6B,E**). By contrast, the simple identification of the same sounds without any categorization led to weaker similarity between RSA matrices of the auditory system and of the CNN (**Fig. 6C,E**). The mismatch was even larger for the autoencoder network performing sound compression and denoising (**Fig. 6D,E**). This result suggests that parallel categorization of multiple features is an important function of the computations that shape representations in the mouse auditory system. Representations in the CNN performing word and music categorization also tended to outperform identification and compression networks (**Fig. 6E**), further supporting the idea that categorization is a key computation for the mouse auditory system.

For the multi-categorization task, we further determined which categories were important to account for the neural data. For example, if we removed frequency modulation categories and trained a network on the reduced version of the task, we did not observe decorrelation of time-averaged representations for frequency sweeps of opposite direction, unlike in the full task or in the auditory system (**Fig. 6F,G** see **Fig. S5J** for the same analysis with each of the four categories). This confirms that the detailed structure of the task is directly reflected in RSA matrices. We found that, except in the AN, removing frequency modulation, amplitude modulation and intensity categories strongly reduced the match between CNN and auditory system RSA matrices (**Fig. 6H)**. Removing frequency categories had little effect, likely because this information was explicitly available in the structure of the input, but the removal of the intensity categories had a major effect, underscoring the importance of this feature in the mouse brain (**Fig. 6H)**.

Despite the strong analogies, none of the networks fully reproduced RSA matrices observed experimentally and further discrepancies were observed. First, CNNs tended to implement a stronger decorrelation of representations in their deeper layers than those observed in AC (**Figs. 5A-D** vs **2F**). Second, no re-correlation of representations was found in CNNs unlike what we observed in TH (**Figs. 5A-D** vs **2F**). Hence, the monotonous transformations implemented in CNNs differ from those in the subcortical part of the auditory system. In line with this, the CNN layer that resembled most IC, TH and AC representations was generally the same intermediate layer (**Fig. S5I**).

Overall, comparison of CNN and auditory system representations indicate a crucial role of sound categorization both in the decorrelation of time-averaged representations and in the fine structure of representations.

Discussion

Our findings show that a key distinguishing property of auditory cortical representations with respect to subcortical levels is that time-averaged representations recapitulate time-dependent representations at the population level. So far, single cell level analyses have opposed a temporal code in neurons that follow temporal fluctuations of sounds and a rate code in neurons that lack temporal accuracy but whose time-averaged activity changes with the stimulus [11]. Several studies have noted an increase of rate coding neurons between subcortical and cortical structures although mostly for rapid acoustic fluctuations [13,38]. However, neurons with temporal coding properties still exist in the AC even for fine time scales[48], making it difficult to establish what specific coding scheme emerges at this level. Our results solve this conundrum by establishing that the code specifically reorganizes in the cortex to resist time-averaging at the population level, while temporal properties are largely preserved (**Fig. 2E-I, Fig. S3G,H**). This hybrid coding scheme also naturally emerges in deep neuronal networks that perform different types of sound categorizations (**Fig. 5**), indicating that this is likely a generic mechanism to extract meaning from multidimensional time-varying signals.

At the single neuron level, a first necessary condition for the representation to be robust to time-averaging at population level is the existence of neurons which respond specifically to particular directions and/or speeds of temporal modulations (e.g. **Fig. 2A-B**). A second necessary condition is that these specific responses are sufficiently diverse and complementary across neurons to cover all the information that is known to be contained in the temporal dimension (ex: first spike timing, temporal multiplexing)[12,49]. While the first condition was extensively studied [14,17–19,23,24,40,50], the second was never investigated. By using new noise corrected population analysis tools, our study addresses this question to demonstrate that temporal information efficiently transfers to the neural population dimensions, however only in the cortex. It is important to note that low temporal resolution neuronal recording methods, such as functional magnetic resonance or ultrasound imaging [2,26,51], assume that temporal information transfers to the population level. Our results validate this assumption for cortical representations but disprove it for subcortical levels. Further experiments are necessary to precisely demonstrate this point in other species and for more complex sounds than the ones used in our study.

Using modeling and causal optogenetic manipulation (**Figs. 4-6**), we also show that constructing auditory representations that resist time-averaging is functionally important for transforming time-varying inputs into decisions, perceptual categories or meaningful auditory objects. Biologically realistic reinforcement learning linking temporally structured representations of stimuli to decisions is strongly accelerated when the time-averaged representations are decorrelated (**Fig. 4**). Previous studies [52,53] have shown that rats can detect whether 100ms intra-cortical electrical stimulations of two loci are synchronous or presented with a relative delay between 3 to 100ms. The stimulation patterns themselves are highly correlated after time-averaging. However, the overlap of the two stimuli at short timescales can easily introduce various spike count modifications in the generated activity patterns through synaptic interactions in the cortical circuit. These could contribute to the perceptual discrimination[54] beyond the available temporal cues. Our time-symmetric stimulations with temporal gaps of 25 ms across sequence elements (see Methods) avoid this

issue by limiting local processing at synaptic time scales. Hence, these results do not contradict ours.

It has long been proposed that appetitive or aversive discriminative learning can occur without cortex through direct thalamic projections to the amygdala and the striatum [55,56] when stimuli are sufficiently simple[57]. Our findings indicate that the degree of stimulus simplicity tightly relates to the dissimilarity of time-averaged representations in thalamus and cortex which will determine which pathway drives faster learning. This is in line with the fact that discriminations involving overlapping frequency modulations, more decorrelated in AC than in TH (**Fig. 3**), are cortex-dependent [16–18]. This kinetic competition between cortex and thalamus is supported by evidence of learning in cortico-striatal projections even for simple discriminations[58].

An important observation of our study is that artificial networks which efficiently perform perceptual decisions to identify sounds categories rely on representations that become resistant to time-averaging (**Fig. 5**). Our results also illustrate how the task which artificial networks are trained to perform tightly dictates representation structure (**Fig. 6**). This is in line with recent findings that natural constraints on perceptual tasks generate representations similar to human brain representations [26,59]. Our study already suggests that the mouse auditory system structures sound information in a manner compatible with broad categorization purposes (**Fig. 6**), but this approach could be extended towards more precise inferences of ecologically relevant stimulus categories in particular species. Beyond analogies, our systematic comparison of artificial neural networks also allowed identifying major differences between the auditory system and deep networks, as a few studies have started to indicate[60]. Most strikingly, CNNs produce a gradual, step-by-step decorrelation whereas in the auditory system the transformation is non-monotonic with a denser, more correlated representation in the TH. This may reflect additional functional or anatomical constraints that are not taken into account by models and that will also need to be disentangled.

Acknowledgments: We thank Maia Brunstein of the Hearing Institute Bioimaging Core Facility of C2RT/C2RA for help in acquiring Airyscan images of thalamocortical boutons in the auditory cortex and Alexander Kell for help implementing the word and music classification network. We also thank Yves Boubenec, Yves Frégnac, Andrew King, Srdjan Ostojic and Christine Petit for their feedback on the manuscript. We acknowledge funding from Fondation pour l'Audition, FPA IDA02 (BB) and APA 2016-03 (BB) ; European Research Council, ERC CoG 770841 DEEPEN, (BB) ; Fondation pour la Recherche Médicale SPF202005011970 (SB) ; European Union's Horizon 2020 research and innovation programme under grant agreement No 964568, project Hearlight (BB). We acknowledges the support of the Fondation pour l'Audition to the Institut de l'Audition.

Author contributions: S.B., Ja.B., A.K. and B.B. conceived experiments, designed the study and interpreted data. S.B., Ja.B., A.K., T.T., J.S. and B.B. collected data and S.B., Ja.B., A.K., S.C. and B.B. performed data analysis. Je.B. and J.L.P conceived and implemented the cochlear model. B.B. and S.B. implemented the reinforcement learning model. S.B., K.B. and Y.G. implemented the deep learning models. S.B., Ja.B. and B.B. prepared figures. S.B. and B.B. wrote the manuscript. S.B and B.B. managed the project.

Declaration of interests: The authors declare no competing interests.

Figures and legends

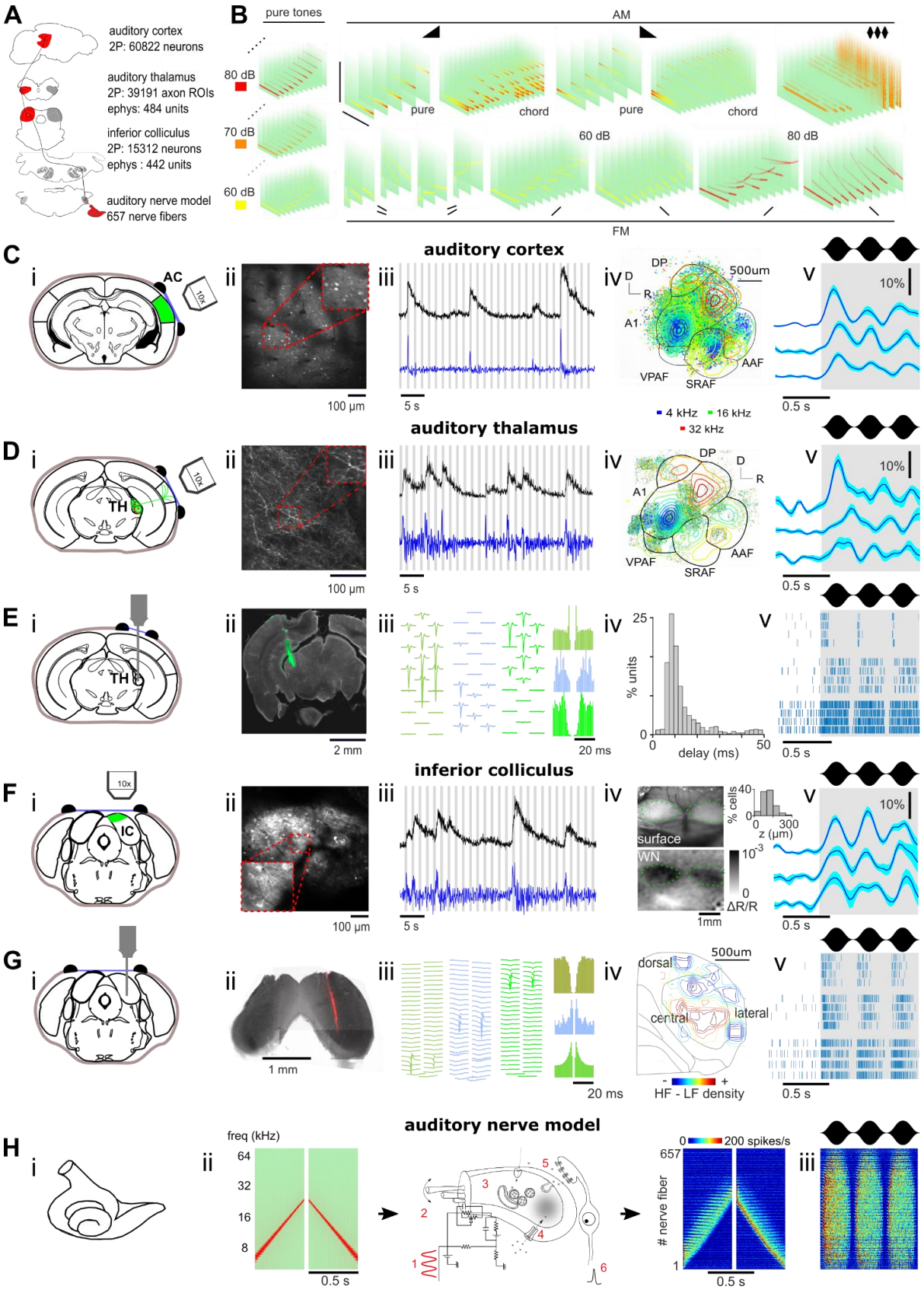

Figure 1: Extensive sampling of the auditory hierarchy. A. Sketch of the auditory system and sample sizes at each level. **B.** Spectrograms of the sound set. **C.** (i) Schematic of imaging

strategy, (ii) sample field of view, and (iii) raw (black) or deconvolved (blue) calcium traces (gray bar: sound presentation) for a sample neuron in AC. (iv) Location of all recorded neurons, color-coded according to their preferred frequency at 60dB, overlayed with the tonotopic gradients obtained from intrinsic imaging. (v) Response of 3 neurons to 3Hz amplitude modulated white noise. **D.** Same as in C for thalamic axon imaging. **E.** (i) Schematic of recording strategy, (ii) sample histology with di-I strained electrode track, (iii) average waveforms and auto-correlograms of three single units, (iv) response latencies of all single units, (v) raster plot of 5 trials from 3 sample units in response to 3Hz modulated white noise for auditory thalamus. **F.** Same as C for dorsal IC except for (iv): view of the cranial window and intrinsic imaging response to white noise. Inset histogram shows distribution recording depths. **G.** Same as E for central IC, except for (iv): reconstructed of IC tonotopy from single units. **H.** (i) Schematic of the cochlea and (ii) of the biophysical model taking a spectrogram as input and providing the responses of auditory nerve fibers. (iii) Response to 3Hz amplitude-modulated white noise.(A1 : primary auditory cortex, DP: dorsal posterior field, AAF: anterior auditory field, VPAF : ventral posterior auditory field, SRAF : suprarhinal auditory field)

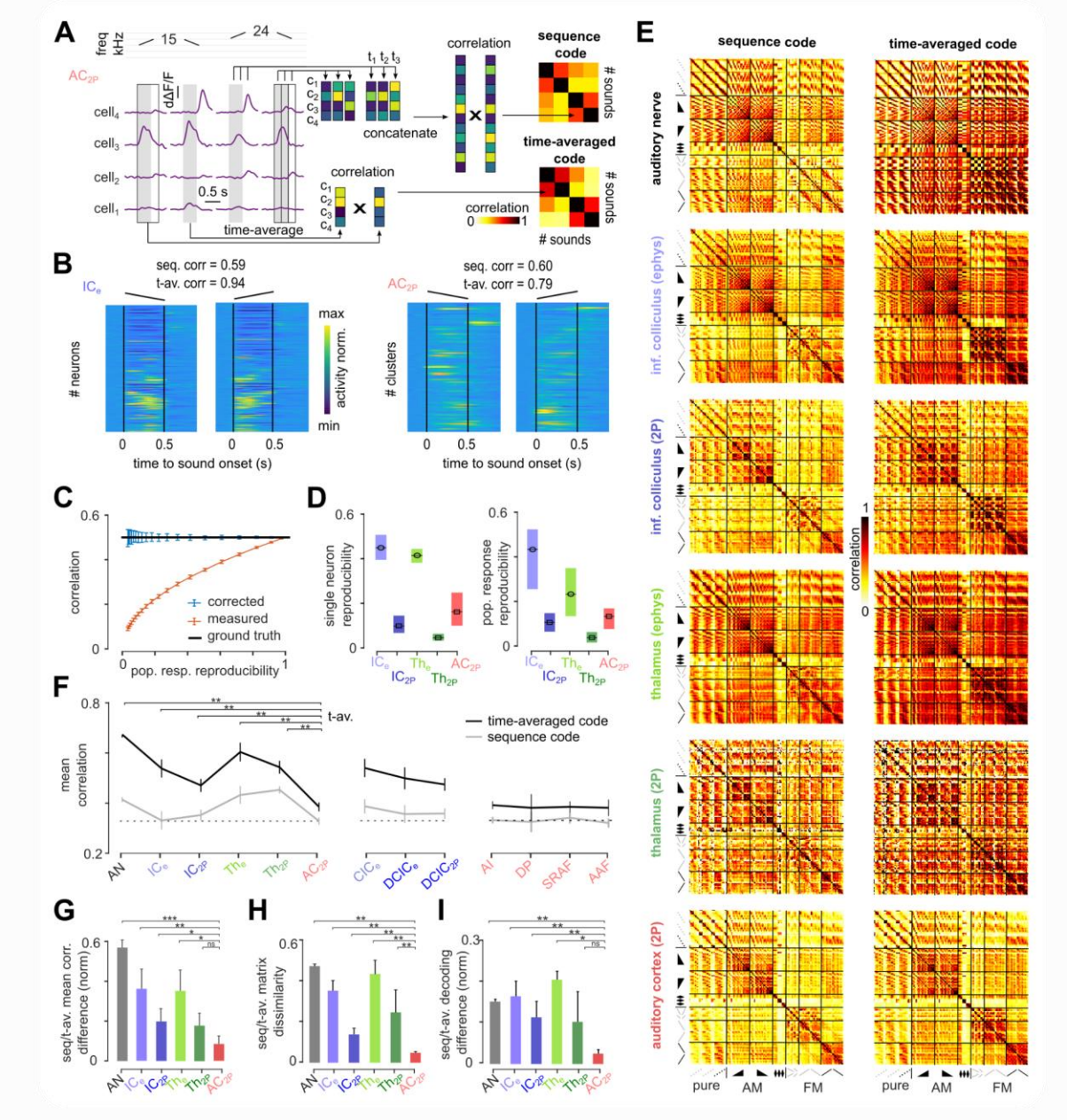

Figure 2 : Emergence of a time-independent cell identity code in the auditory cortex. A. Responses of 4 AC neurons to different up and down frequency sweeps illustrating how sequence and time-averaged correlation is calculated to compose the RSA matrices. **B.** Sample responses to up and down frequency sweeps from IC and AC neurons ordered by response amplitude. **C.** Measured correlation of simulated data with low to high response reproducibility before (orange) or after (blue) noise-correction. **D.** Reproducibility of single neuron (left) or population (right) responses measured as the mean inter-trial correlation between responses across sounds (left : n=number of neurons per area, right : n=140 sounds, error bars are quantiles). **E.** Noise-corrected RSA matrices for all sound pairs for sequence (left) or time-averaged (right) codes. **F.** Mean noise-corrected correlation by area. (p-value for 100 bootstraps comparing time-averaged correlation of each region to AC, error bars are bootstrapped S.D). **G.** Normalized difference between mean noise-corrected correlation for time-averaged and sequence codes. (p-value for 100 bootstraps, errors bars are S.D). **H.** Noise-corrected dissimilarity between RSA matrix structure of time-averaged and sequence

codes. (p-value for 100 bootstraps, error bars are S.D). **I.** Normalized difference between mean sound decoding accuracy for time-averaged and sequence codes. (p-value for 100 bootstraps, error bars are S.D).

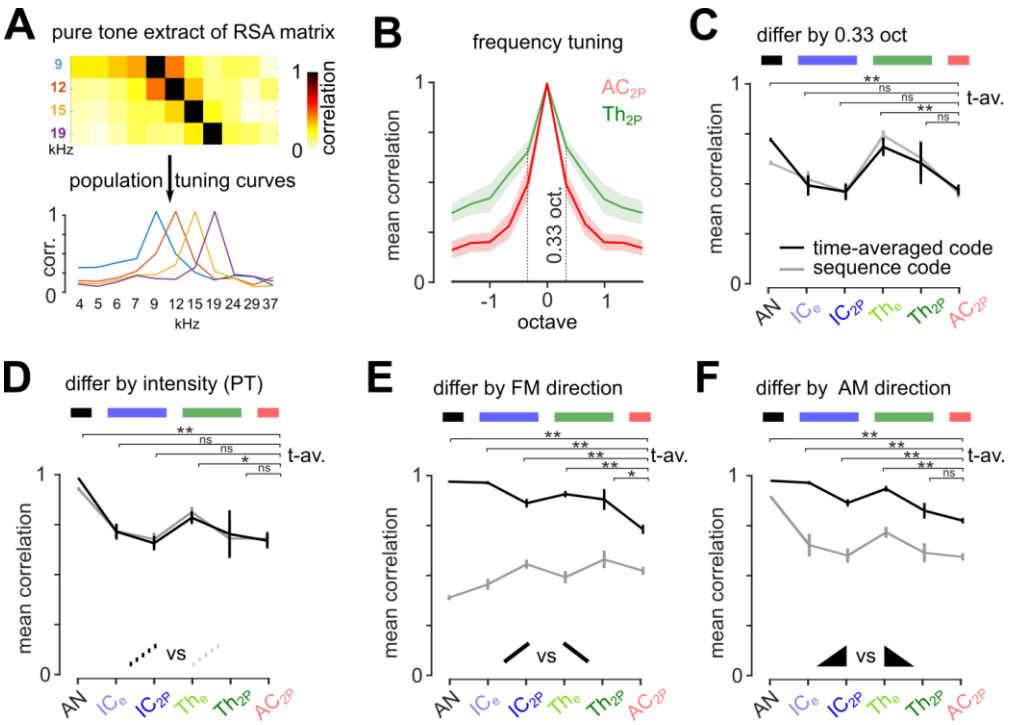

Figure 3: Time-averaged representations of time-symmetric sounds decorrelate in AC. **A.** Illustration of method to calculate population tuning curves shown in B from RSA matrix. **B.** Mean noise-corrected correlation between pure tones as a function of their frequency separation. **C–F.** Mean noise-corrected correlation between sound pairs differing by only one acoustic property : **C**. pure tones at 70dB differing by 0.33 octaves, **D**. pure tones at the same frequency differing by intensity, **E**. frequency sweeps with same start and end frequency at same intensity differing by direction, **F**. amplitude ramps at same frequency differing by direction. For sounds without temporal structure, correlation of representations are similar in AC and IC, whereas for time-symmetric sounds, all brain areas show larger time-averaged correlations than in the cortex, except for TH2P in F likely due to teh high variability of thalamic responses. p-value for 100 bootstraps comparing time-averaged correlation of each region to AC, error bars are S.D. Statistical test details are given in the **Supplemental Table 3**.

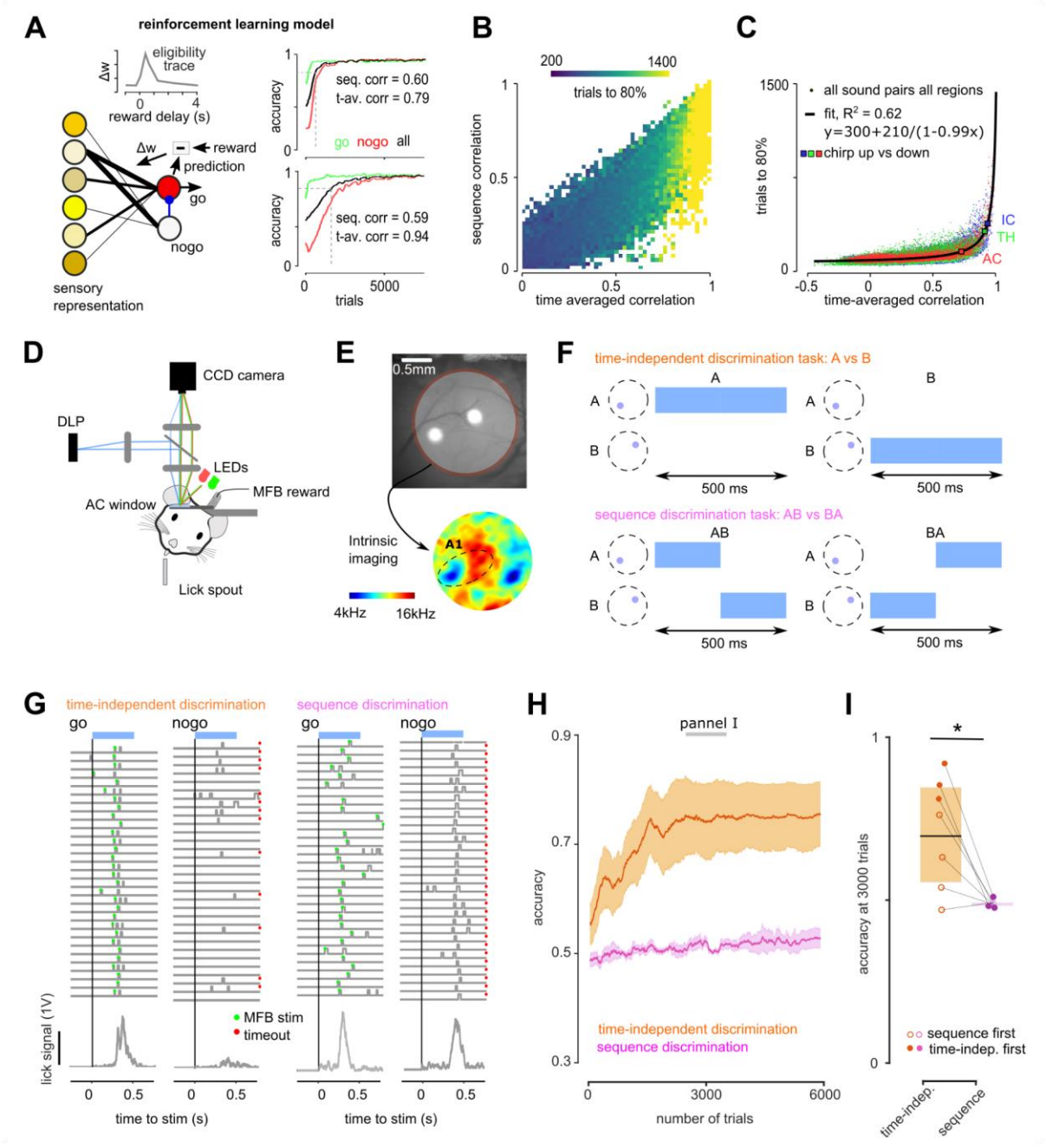

Figure 4 : Time-independent sound representations in AC supports faster learning. **A.** Sketch of the reinforcement learning model (bottom left), eligibility trace dynamics (top left) and example learning curves for two recorded representations that have similar sequence code correlations but different time-averaged code correlations. **B.** Heatmap of the number of trials needed to reach 80% accuracy at discriminating between a pair of sounds as a function of the time-averaged and sequence code correlations between the representations of these sounds (averaged over all pairs of representations for all brain regions). **C.** Number of trials to 80% accuracy as a function of the correlations of time-averaged representations. Large square dots show the mean correlation and learning time for time-symmetric frequency sweeps in IC, TH and AC and the black line shows the fit to data. **D.** Sketch of patterned optogenetic experiment in AC (MFB: medial forebrain bundle). **E.** Cortical window from an example mouse showing the location of the stimulation spots and the corresponding tonotopic map. **F.** Sketch

of the optogenetic stimulation time courses for each discrimination task. **G.** Sample lick traces (top) and mean lick signal (bottom) for Go and NoGo trials in the time-independent (left) and sequence (right) discrimination tasks. Green dots: reward times. Red dots: timeouts. **H.** Learning curves for all mice performing each task (n=7, error bars are sem). **I.** Accuracy at 1500 trials for all mice. (paired Wilcoxon test, p = 0.032, signed rank value = 27, n=7).

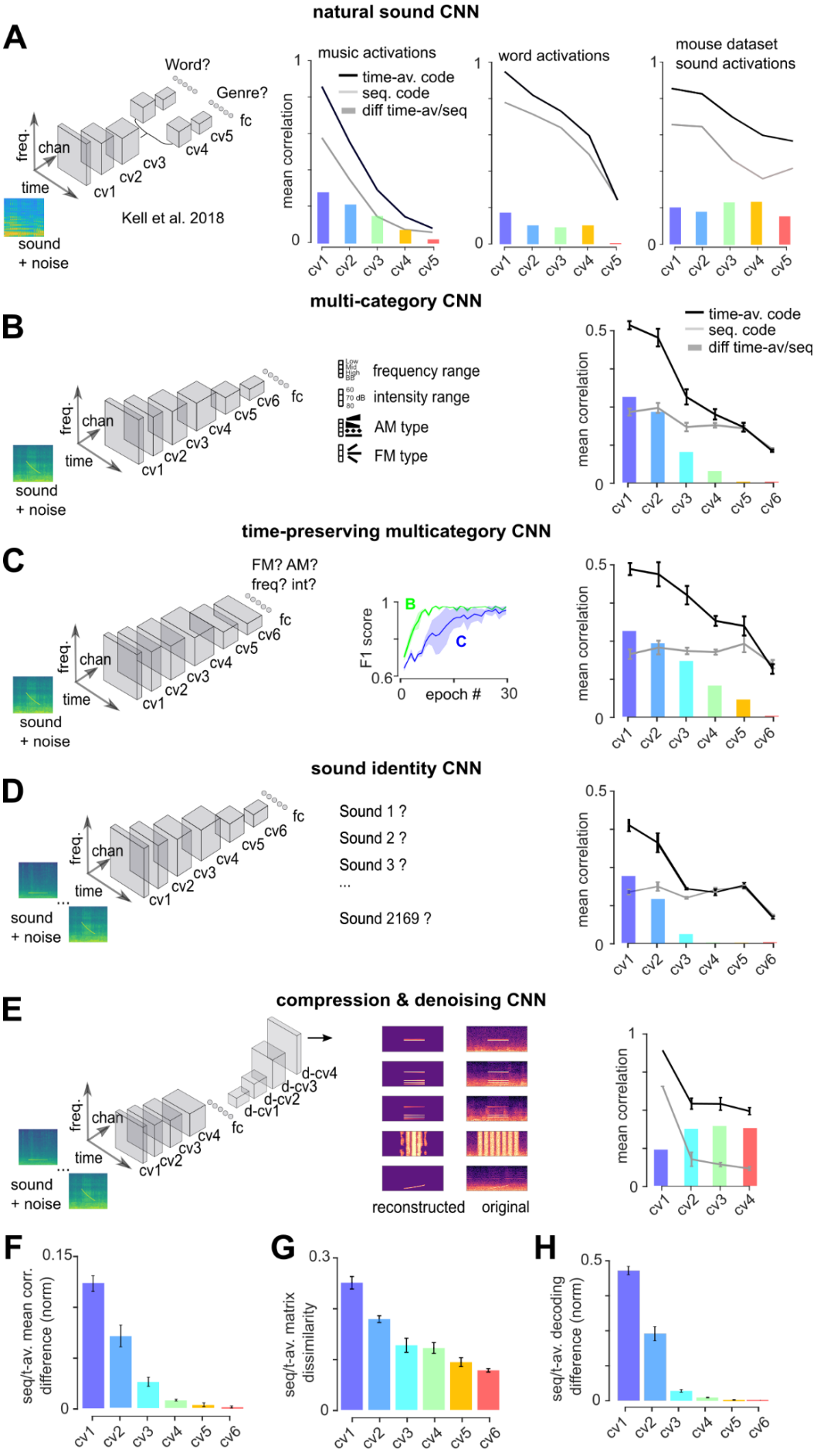

Figure 5 : Categorization deep networks implement a time-independent code in the deep layers.
A-E. (left) Schematic of CNN architectures and target categories. **B-E** (right) Mean response correlations for the sequence and time-averaged codes from RSA matrices constructed with the set of 140 sounds presented to mice (line) and difference between the two (bars). **A.** Model adapted from [26] using RSA matrices constructed using musical snippets (left), words (center) or the set of 140 sounds presented to mice transposed to human hearing range (right). **B.** Multi-category CNN (n=8 networks). **C.** Multi-category CNN without shrinking of the temporal dimension (n=8 networks). Inset shows learning curves from training epochs for networks in B and C. **D.** CNN performing sound identification **E.** Autoencoder CNN performing sound compression and denoising through a 20-unit bottleneck. **F-H.** All graphs refer to the time-preserving categorization CNN and reproduce analysis shown in **Fig 2G-I** : **F.** Normalized difference between mean noise-corrected correlation for time-averaged and sequence codes. **G.** Noise-corrected dissimilarity between RSA matrix structure of time-averaged and sequence codes. **H.** Normalized difference between mean sound decoding accuracy for time-averaged and sequence codes. (error bars are sem over trained networks). (cv : convolution block, d-cv : deconvolution block - see methods for architecture details)

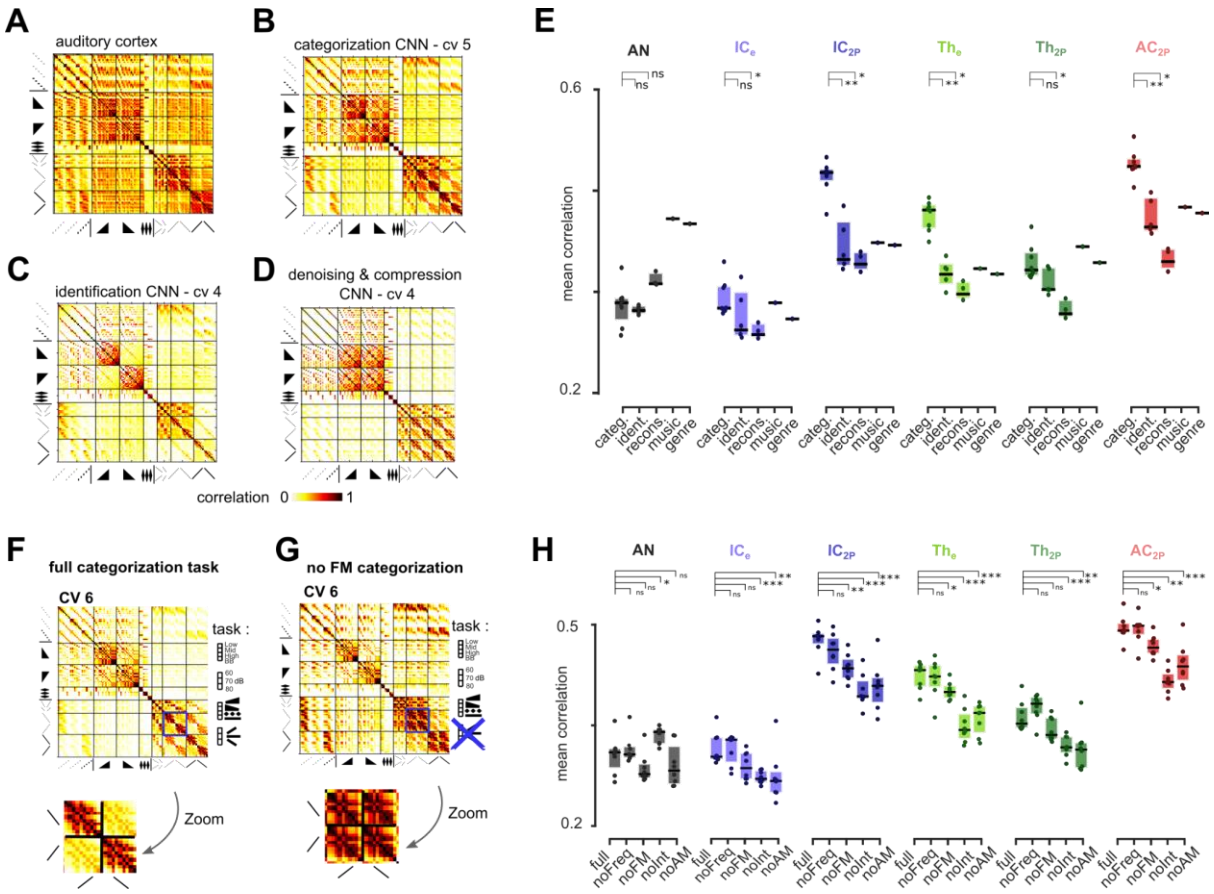

Figure 6: Signatures of sound categorization in the mouse auditory system
A-D. Time-averaged code RSA matrix from AC (A) and the closest resembling layer of CNNs performing multi-category (B), identification (C), denoising and compression (D) tasks. **E.** Correlation between the RSA matrices from each region of the mouse auditory system and the closest resembling layer of CNNs performing different tasks. Each point represents one network trained on the task either with different architecture or different random initialization. (Statistics are sign rank tests, n=8,8,4,1,1, p-values in **Supplemental Table 3**) **F.G** RSA matrices from a CNN trained perform the full multi-category task with four different category types (F) or with only three category types excluding the frequency modulation (FM) type (G). The magnified part of the matrix shows the presence or absence of FM sweeps decorrelation depending on whether FM stimuli are classified or not. **H.** Correlation between the RSA matrices in each brain area and the most similar layer of CNNs trained on the full multi-category task and partial multi-category tasks that exclude one out of the four category types (Statistics are sign rank tests, n=8, p-values in **Supplemental Table 3**).

MATERIALS AND METHODS

RESOURCE AVAILABILITY

Data availability

All datasets are freely available at 10.12751/g-node.sz67di, hosted by G-Node Infrastructure.

Code availability

Custom codes used in this study are freely available at 10.12751/g-node.sz67di, hosted by G-Node Infrastructure.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All mice used for imaging and electrophysiology were 6 to 14 weeks old male and female C57Bl6J mice that had not undergone any other procedures. For optogenetic stimulation, we used Emx1-IRES-Cre (Jax #005628) crossed with Ai27 (Jax #012567) mice. Mice were group-housed (2–6 per cage) before and after surgery, had ad libitum access to food and water and enrichment (running wheel, cotton bedding and wooden logs) and were maintained on a 12-hour light-dark cycle in controlled humidity and temperature conditions (21-23°C, 45-55% humidity). All experiments were performed during the light phase. All experimental and surgical procedures were carried out in accordance with the French Ethical Committee the French Ethical Committees #59 and #89 (authorizations APAFIS#9714-2018011108392486 v2 and APAFIS#27040-2020090316536717 v1).

METHOD DETAILS

Surgery

Mice were injected with buprenorphine (Vétergesic, 0,05-0,1 mg/kg) 30 min prior to surgery. Surgical procedures were carried out using either intraperitoneal ketamine (Ketasol) and medetomidine (Domitor) which was antagonized with atipamezole (Antisedan, Orion pharma) at the end of the surgery) or 3% isoflurane delivered via a mask. After induction, mice were kept on a thermal blanket during the whole procedure and their eyes were protected with Ocrygel (TVM Lab). Lidocaine was injected under the skin of the skull 5 minutes prior to incision.

For calcium imaging, craniotomies of either 3 (IC) or 5 (AC) mm were performed above the IC or the AC. Injections of 150nL of AAV1.Syn.GCaMP6s.WPRE (Vector Core, Philadelphia, PA; 10^13 viral particles per ml; used pure for TH and diluted 30x for AC and IC) were made at 30 nL/min with pulled glass pipettes at a depth of 500µm and spaced every 500 µm to cover the whole surface of the IC or AC. The craniotomy was sealed with a circular glass coverslip. The coverslip and head post were fixed to the skull using cyanolite glue and dental cement (Ortho-Jet, Lang).

For electrophysiology recordings, the skull above the IC or above the cortex dorsal to the TH was exposed for ulterior craniotomy. A well was formed around it using dental cement in order to retain saline solution during recordings and the head post was fixed to the skull using cyanolite glue and dental cement. To protect the skull, the well was filled with a waterproof

silicone elastomer (Kwikcast, WPI) that could be removed prior to recording. The head post was fixed to the skull using cyanolite glue and dental cement (Ortho-Jet, Lang).

For patterned optogenetic stimulation of the cortex, a cranial window was placed above the AC as for calcium imaging but without viral injection. For MFB stimulation, a bipolar stimulation electrode (60-µm-diameter twisted stainless steel, PlasticsOne) was implanted using stereotaxic coordinates (AP -1.4, ML +1.2, DV +4.8). It was then fixed along with the headplate to the skull using dental cement (Ortho-Jet, Lang).

After surgery, mice received a subcutaneous injection of 30% glucose and metacam (1 mg/kg). Mice were subsequently housed for one week with metacam delivered via drinking water or dietgel (ClearH20). Mice were given one week to recover from surgery without any manipulation. Then, for four days before recording, mice were habituated to head restraint for increasing periods of time (30 min - 2 hours). For electrophysiological experiments, the day before recording animals were briefly anesthetized using isoflurane anesthesia (2%) in order to perform craniotomy and durectomy for electrode descent.

Two photon calcium imaging in the awake mouse

Imaging was performed using a two-photon microscope (Femtonics, Budapest, Hungary) equipped with an 8kHz resonant scanner combined with a pulsed laser (MaiTai-DS, SpectraPhysics, Santa Clara, CA) set at 900 nm. We used a 10x Olympus objective (XLPLN10XSVMP), which provided a field of view of up to 1x1 mm. For AC, a 1x1mm field of view was used. For IC, the field of view was adjusted to the size of the structure (~0.5x0.5 mm). For thalamic axons, the field of view was reduced to 0.22x0.22 mm. Images were acquired at 31.5 Hz.

Electrophysiology in the awake mouse

Electrophysiology was performed using Neuronexus probes : (1x32 linear probe for IC and 4*8 comb for TH). For track reconstruction, the electrodes were dipped in dil, diO or diD (Vybrant™ Multicolor Cell-Labeling Kit, Thermofisher) prior to recording and allowed to dry at least 15 min before insertion. Recordings were performed using warmed saline filling the cyanolite glue well and in contact with the reference electrode. After each recording the well was amply flushed and then refilled with Kwickast. A maximum of three recordings were performed per site. Data was sampled at 20kHz using an Intan RHD2000 amplifier board.

Sound delivery

Sounds were generated with Matlab (The Mathworks, Natick, MA) and were delivered at 192 kHz with a NI-PCI-6221 card (National Instruments) driven by the software Elphy (G. Sadoc, UNIC, France) and feeding an amplified free-field loudspeaker (SA1 and MF1-S, Tucker-Davis Technologies, Alachua, FL) positioned 15 to 20 cm from the mouse ear. Sound intensity was cosine-ramped over 10 ms at the onset and offset to avoid spectral splatter. The head fixed mouse was isolated from external noise sources by sound-proof boxes (custom-made by Femtonics, Budapest, Hungary or Decibel France, Miribel, France) providing 30 dB attenuation above 1 kHz. Sounds were calibrated in intensity at the location of the mouse ear using a probe microphone (Bruel & Kjaer, type 4939-L-002). For two-photon calcium imaging,

the resonant scanner generated a harmonic background noise at 8kHz (intensity at the mouse ear, 45 dB SPL).

During a recording session, each of the 140 sounds (sketched in **Fig. 1B**) was presented 15 times in random order. In order to be compatible with 2 photon image acquisition, sounds were presented in 120 blocks of 32s each, interleaved by a 15s pause in a 94 min protocol. The list of all sound parameters can be found in the **Supplemental Table 2**.

Intrinsic optical imaging recordings in anesthetized mouse

Intrinsic imaging was performed to localize AC in mice under light isoflurane anesthesia (1% delivered with SomnoSuite, Kent Scientific) on a thermal blanket. Images were acquired at 20Hz using a 50mm objective (1.2 NA, NIKKOR, Nikon) with a CCDcamera (GC651MP, Smartek Vision) equipped with a 50 mm objective (Fujinon, HF50HA-1B, Fujifilm) through the cranial window implanted 1-2 weeks before the experiment (4-pixel binning, field of view between 3.7 x 2.8 mm or 164 x 124 pixels at 5.58 mm/pixel). Signals were obtained under 780 nm LED illumination (M780D2, Thorlabs). Images of the vasculature over the same field of view were taken under 530 nm LED illumination (NSPG310B, Conrad). Sequences of short pure tones at 80 dB SPL were repeated for 2 s every 30 s with 10 trials per sound. Acquisition was triggered and synchronized using a custom made GUI in MATLAB. For each sound, we computed baseline and response images, 3 s before and 3 s after sound onset, respectively. The change in light reflectance $\Delta R/R$ was calculated for each repetition of each sound frequency (4, 8, 16, 32 kHz, white noise) as the difference between the baseline and response image and was then averaged across all repetitions of a given tone frequency. Response images were smoothed applying a 2D Gaussian filter (sd = 3 pixels). Auditory cortex activity appeared as regions with reduced light reflectance changing with frequency, revealing the tonotopic maps of its different subfields. To align intrinsic imaging responses across different animals, the 4 kHz response was used as a functional landmark. The spatial locations of maximal amplitude responses in the 4 kHz response map for the A1, A2 and AAF (three points) was extracted for each mouse and a Euclidean transformation matrix was calculated by minimizing the sum of squared deviations (RMSD) for the distance between the three landmarks across mice. This procedure yielded a matrix of rotation and translation for each mouse that was applied to compute intrinsic imaging responses averaged across a population of mice.

Histology and immunostainings

In order to extract the brain for histology, mice were deeply anesthetized using a ketamine-medetomidine mixture and perfused intracardially with 4% buffered paraformaldehyde fixative. The brains were carefully dissected and left in paraformaldehyde overnight and then sliced into fifty micrometer sections using a vibratome. Slices were either stained with cytochrome oxidase or directly mounted using a mounting medium with DAPI. Analysis of the fluorescence band diI, diO or diD allowed isolating up to 3 tracks per mouse for electrophysiological experiments.

For Vglut2 immunostainings, after fixation, tissues were rinsed in PBS and blocked in Tris-Buffered Saline (TBS) supplemented with 5 % (vol/vol) Normal Donkey Serum (Jackson Immunoresearch) and 0.3 % (wg/vol) Triton X-100. Then, sections were incubated for 48h at 4°C while rocking with a primary antibody: guinea pig anti-Vglut2 (1:500, Synaptic Systems

#135404), followed by a 4 h incubation with a secondary donkey anti-guinea pig IgG [F(ab')2fragments] (1:500, Jackson ImmunoResearch #706606148). Tissues were rinsed and mounted using Prolong diamond antifade (Life Technologies). Pictures of the brain sections were taken with LSM 900 confocal microscope (Zeiss Microsystems) using 20x objective, whereas the magnified view of the thalamocortical boutons was obtained with Airyscan acquisition and 63x objective.

The labeled boutons (GCaMP alone in green; GCaMP with Vglut2 in yellow) were counted manually using ZEISS ZEN 2 microscope software in 12 sample regions selected within layer 1 AC in 3 different Airyscan images. The number of boutons was then calculated per volume tissue.

Behavioral discrimination of patterned optogenetic stimuli

For patterned optogenetic activation in the mouse AC, we used a video projector (DLP LightCrafter, Texas Instruments) powered by a blue LED (center wavelength 460 nm). To project a two-dimensional image onto the AC surface. The image of the micromirror chip was collimated through a 150 mm cylindrical lens (Thorlabs, diameter: 2 inches) and focused through a 50 mm objective (NIKKOR, Nikon). Light collected by the objective passes through a dichroic beam splitter (long pass, > 640nm, FF640-FDi01, Semrock) and is collected by a CCD camera (GC651MP, Smartek Vision) equipped with a 50 mm objective (Fujinon, HF50HA-1B, Fujifilm).

The behavioral task aimed to teach mice to discriminate between two optogenetically induced patterns of activity in AC. The reinforcement used for the task used medial forebrain bundle (MFB) stimulation in non-deprived mice. This protocol leads to similar learning speed, motor response timing and psychometric measurements as water rewards in deprived animals [45]. In the "time-independent task", the two stimuli were composed of 500 ms illumination of 300 µm diameter spots placed at different locations of AC. In the "sequence discrimination task", the two stimuli were composed of a succession of two 250 ms illuminations of 300 µm diameter spots at different locations in the cortex in one order (AB) or in the reversed order (BA). All light stimuli were temporally modulated at 20 Hz (25 ms ON, 25 ms OFF). To prevent visual perception of the optogenetic stimuli a constant and strong background illumination provided by a white LED lamp was used and a cache was placed in front and close to the eyes to limit visual inputs. Mice were trained on both tasks in random order. The spots used in the first task they learnt were positioned at the two extremes of the tonotopic axis of A1 and the spots in the second task were positioned at equal distance, orthogonal to this axis. Alignment of optogenetic stimulus locations across days was done using blood vessel patterns at the surface of the brain manually aligned to a reference blood vessel image taken at the beginning of the experiment.

Behavioral experiments were monitored and controlled using a custom Matlab software controlling an input-output board (PCIe-6351, National Instruments) and the images delivered by the video projector. Mice performed behavior for one hour per day. During the entire behavioral training period, food and water were available *ad libitum* as rewards were provided through the stimulation of the medial forebrain bundle (MFB).

MFB stimulation was delivered via a pulse train generator (PulsePal V2, Sanworks) that produced 2ms biphasic pulses at 50Hz for 100ms at a voltage calibrated for each individual mouse to the minimal level that evoked sustained responding, using the protocol in [45]. The stimulation was controlled with a solenoid valve (LVM10R1-6B-1-Q, SMC). A voltage of 5V was applied through an electric circuit joining the lick tube and an aluminum foil on which the mouse was sitting. Lick events could be monitored by measuring the voltage across a series resistor in this circuit.

Training was broken down into three phases. (i) Lick training: On the first day, mice were presented with the lick tube and any licking was rewarded with immediate MFB stimulation. Mice generally began licking at high rates after 1-2minutes and the session was continued until mice reliably collected around 300 rewards. (ii) Go training: On the following day, Go trials were presented with 80% probability, while the remaining trials were blank trials (no stimulus). A trial consisted of a random inter-trial interval (ITI : 0.5 to 1 s), a random 'no lick' period (duration adjusted, see below) and a fixed response window of 1.5 s. The first lick occuring during the response window on a Go trial was scored as a 'hit' and triggered immediate MFB stimulation. During initial go training the 'no lick' period was between 2 and 5 s in order to discourage non-specific licking. When mice achieved >80% accuracy for the Go stimulus, a final Go session was performed during which a cache was placed over the window to verify that animals were not licking to remnant visual cues from the video projector (**Fig. S4**). On this day and for subsequent Go/NoGo sessions, the no lick period was shortened to 1.5 to 3 s in order to obtain more trials per session. (iii) Go/NoGo training: After Go training, the second stimulus (NoGo) was introduced. During presentation of the NoGo sound, the absence of licking for the full response window was scored as a 'correct rejection' (CR) and the next trial immediately followed. Any licking during NoGo trials was scored as a 'false alarm' (FA), no stimulation was given, and the animal was punished with a random time-out period between 5 and 7 s. Each session contained 45% Go stimuli, 45% NoGo stimuli and 10% blank stimuli. Note that the Go training was used to ensure high motivation of the animal during the Go/Nogo training by establishing an association between the optogenetic stimulus and the reward. For the time-independent task, this association was generalized to the NoGo stimulus, as seen through very high false alarm rates at the beginning of the Go/NoGo training (e.g. **Fig. S4**). This indicates that faster learning for the time-independent task is not due to an absence of generalization between the Go and NoGo stimulus when transitioning from the Go to the Go/NoGo training phases.

Learning curves were obtained by calculating the fraction of correct responses over blocks of 150 trials. Discrimination performance over one session was calculated as (hits + correct rejections)/total trials.

Data pre-processing

For calcium imaging, regions of interest corresponding to putative neurons (AC and IC) or axons and boutons (TH) were identified by using Autocell [16] (https://github.com/thomasdeneux/Autocell). Briefly, each frame of the recording was corrected for horizontal motion using rigid body registration.This step was visually controlled and all sessions with visible z motion were eliminated. A hierarchical clustering algorithm, based on pixel covariance over time, agglomerated pixels up to a user-selected number of

clusters corresponding to regions of the size of neurons of axons. Clusters were automatically filtered according to size and shape criteria. This step was controlled by a detailed visual inspection of selected regions of interest (ROIs) during which ROIs without visually identifiable cell body shape were discarded.

For each region of interest, the mean fluorescence signal F(t) was extracted together with the local neuropil signal $F_{np}(t)$. Then 70% of the neuropil signal was subtracted from the neuron signal to limit neuropil contamination. Baseline fluorescence F_0 was calculated with a sliding window computing the 3rd percentile of a Gaussian-filtered trace over the imaging blocks. Fluorescence variations were then computed as $f(t) = \Delta F/F = (F(t) - F_0)/F_0$. An estimate of firing rate variations r(t) was then obtained by linear temporal deconvolution of f(t): $r(t) = f'(t) + f(t)/\tau$, f'(t) being the first derivative of f(t) and $\tau = 2s$, the estimated decay of the GCAMP6s fluorescent transients. This simple method efficiently corrects the strong discrepancy between fluorescence and firing rate time courses due to the slow decay of spike-triggered calcium transients. It does not correct for the rise time of GCAMP6s, leading to remnant low pass filtering of the firing rate estimate and a delay of ~100ms between the firing rate peaks and the peaks of the deconvolved signal. Finally, response traces were smoothed with a Gaussian filter ($\sigma = 31ms$).

Electrophysiological signals were high-pass filtered and spike sorting was performed using the CortexLab suite (https://github.com/cortex-lab, UCL, London, England). Single unit clusters were identified using kilosort 2.5 followed by manual corrections based on the interspike-interval histogram and the inspection of the spike waveform using Phy (https://github.com/cortex-lab/phy).

Both for imaging and electrophysiology data, single trial sound responses were extracted (0.5s before and 1s after sound onset) and the average activity over the prestimulus period (0.5s - 0s before sound onset) was subtracted for each trial.

Reproducibility index and cell selection

To quantify the noise levels in the data, we calculated the mean inter-trial correlation across all pairs of trials. The single neuron reproducibility is then defined for each neuron as the average of the inter-trial correlation for that neuron's response to all 140 sounds. The population response reproducibility for each sound is defined as the average of the inter-trial correlations of the full sequence of response of the whole neural population to that sound. Region of interests (ROIs) or single units with reproducibility below 0.12 were classified as non-responsive and were excluded from all analyses except population sparseness. As detailed in the **Supplemental Table 1**, the number of responsive units and the corresponding fraction of the total number of units/ROIs recorded are: AC, 19414 (32%), TH, 3969 (12%), THE, 484 (97%), 5936 (39%), 442 (78%).

Noise-corrected correlation

For each dataset, population representations were estimated after pooling all recording sessions in a virtual population. We used the correlation between population vectors as a metric of similarity between representations. The areas and techniques used to estimate neuronal ensemble representations yielded different levels of trial-to-trial variability due to intrinsic neuronal response variability and measurement noise. Most representation metrics are biased by variability, even after trial averaging, due to variability residues. For example, the correlation between two population representations (population vectors) will tend to

decrease with respect to a variability-free estimate [36]. When multiple observations of the same representations are available, it is possible to account for the impact of variability, by using specific estimators [36]. Here we showed analytically (see **Supplemental Mathematical Derivations**) that the value of the Pearson correlation coefficient $\rho_{\vec{\nu}_s\vec{\nu}_{s'}}$ between population vectors for two sounds $\vec{\nu}_s$ and $\vec{\nu}_{s'}$ in absence of variability can be exactly estimated from noise-corrupted single-trial observations $\vec{\nu}_{s,r}$ and $\vec{\nu}_{s',r'}$ of $\vec{\nu}_s$ and $\vec{\nu}_{s'}$ when their dimension N approaches infinity, based on the formula:

$$\rho_{\vec{\nu}_s\vec{\nu}_{s'}} \approx \frac{\frac{1}{R^2}\sum_{r,r'}\rho_{\vec{\nu}_{s,r}\vec{\nu}_{s',r'}}}{\sqrt{\frac{1}{R^2(1-R)^2}\left(\sum_{r\neq r'}\rho_{\vec{\nu}_{s,r}\vec{\nu}_{s,r'}}\right)\left(\sum_{r\neq r'}\rho_{\vec{\nu}_{s',r}\vec{\nu}_{s',r'}}\right)}}$$

in which r and r' are single trial indices and R is the total number of trials. This analytical result is confirmed by simulations for finite N, indicating that our estimator converges to the correlation value of the noise-free vectors (**Fig. 2C**). Code for calculating this estimator is provided with the online data set.

Simulations for finite N show as expected that the estimator displays substantial deviations around the true correlation which however average to zero. This leads to values of the estimator that can be outside [-1,1] in some cases. Our estimator displays extremely large deviations when $\frac{1}{R(R-1)}\sum_{r\neq r'}\rho_{\vec{\nu}_{s,r}\vec{\nu}_{s,r'}}$ approaches 0, i.e. for representations that are dominated by noise. This occurred more often in datasets obtained by imaging, in particular in the thalamic axonal boutons dataset (TH). To limit imprecisions from these extreme values we excluded from all datasets sounds for which $\frac{1}{R(R-1)}\sum_{r\neq r'}\rho_{\vec{\nu}_{s,r}\vec{\nu}_{s,r'}<0.01}$. In typical neural data, there are significant noise correlations across simultaneously recorded neurons within a trial. Therefore, the effective N can be much lower than the number of neurons. We minimized this contribution by shuffling trial identity for each neuron independently.

To evaluate the significance of mean correlation differences across all sound pairs for time-averaged or sequence representations, we used a bootstrap procedure over the independently recorded sessions. This procedure had the advantage of providing a statistical assessment for biological replicability based on strictly independent measurements (neurons of the same recording are not fully independent statistically). The noise-corrected correlation measure was estimated 100 times after a random resampling of sessions with replacement. Based on this distribution, we measured the standard deviation and calculated p-values down to 0.01.

Sequence correlation was measured on vectors formed by concatenating the responses of all neurons throughout time (vector dimension = $N_{Neurons}$ x $N_{TimeBins}$). Time-averaged correlation was measured first by time-averaging the responses of each neuron and then concatenating these values for all neurons (vector dimension = $N_{Neurons}$). In both cases, we used data from the sound onset to 250ms after the sound offset. To normalize the difference between sequence and time-averaged correlation when comparing between areas we use the formula :

$$\rho_{diff} = \frac{\rho_{t-av} - \rho_{seq}}{1 - 0.5 \times (\rho_{t-av} + \rho_{seq})}$$

Noise-corrected sparseness measure

There exist several sparseness measures which are all biased by variability in neuronal activity measurements [61–63]. The most classical measure as defined in [61,62] is not appropriate for baseline-corrected, linearly deconvolved calcium data because it requires positive response values. We show in the **Supplemental Mathematical Derivations** that kurtosis, the 4th order moment of a distribution, is a sparseness measure [63] which can be corrected for variability-related biases and is appropriate for all our datasets. This metric quantifies the "long-tailedness" of the distribution. Sparse response properties correspond to rare and strong responses which generate long-tailed response distributions as opposed to dense response properties which correspond to more compact response distributions. For lifetime sparseness, measured for each neuron separately, Kurtosis is defined as:

$$K_n = \frac{< (\nu_{n,s} - < \nu_{n,s} >_s)^4 >_s}{< (\nu_{n,s} - < \nu_{n,s} >_s)^2 >_s^2} - 3$$

in which $<>_s$ indicates averaging over sounds and $\nu_{n,s}$ is the noise-free response of neuron n to sound s. In the case of population sparseness, which is measured for each sound separately, $<>_s$ should be replaced by $<>_n$ which indicates averaging over neurons. The Kurtosis formula can be developed into the moments of order 1 to 4 of $\nu_{n,s}$.

$$K_n = \frac{< \nu_{n,s}^4 >_s - 4 < \nu_{n,s} >_s < \nu_{n,s}^3 >_s + 6 < \nu_{n,s} >_s^2 < \nu_{n,s}^2 >_s - < \nu_{n,s} >_s^4}{(< \nu_{n,s}^2 >_s - < \nu_{n,s} >_s^2)^2} - 3$$

Starting from the second order, estimates of these moments based on trial-averaged response include noise-related bias terms, which skew the kurtosis estimates for limited trial counts. We analytically demonstrated and numerically verified that these biases can be suppressed using noise corrected formulae of all moments that are detailed in the **Supplemental Mathematical Derivations**. Code for these calculations is provided with the online data set.

When calculating population sparseness, we analyzed all neurons including non-responsive neurons. Non-responsive neurons with aberrant response levels (>5 times the maximal value of responsive neurons) were excluded. Based on this, the percentages of units used were : ICE : 92%, IC: 80%, TH: 61%, THE: 97%, AC:92%).

Population activity classifiers

To evaluate the accuracy of sound identification based on single-trial population responses, we trained a nearest-neighbor classifier on a subset of trials and cross-validated it on a distinct subset of trials. Training and testing sets were constructed by randomly selecting half of the trials for each unit. For each sound, we correlated the population response averaged over the training trials for this sound with the population response averaged over the testing trials for all the other sounds. The sound with the highest correlation was assigned as the prediction. Decoding accuracy is defined as the proportion of correctly assigned sounds.

Sequence and time-averaged codes were defined as for the correlation measures. Statistical significance was evaluated using the same bootstrap procedure as for the correlation

measures. Importantly, decoding depends inherently on trial-to-trial noise which limits the possibility of comparing between areas. This analysis serves to contrast sequence and time-averaged codes within an area.

To measure the information contained at different timescales, the temporal sequence of population activity was decomposed into its Fourier coefficients corresponding to a discrete set of timescales ranging from T, the 750 ms sound response duration, down to $2\Delta t$, where Δt is the discretization time of the dataset ($1/2\Delta t = f$ the Nyquist frequency ; Δt = $T/24$ = 31.25 ms for 2P-imaging data and Δt = $T/96$ = 7.81 ms for electrophysiology data).

The Fourier coefficient $C_{n,r}$ for frequency n/T and neuron r is defined as

$$C_{n,r} = \sum_{k=1}^{2K} v_r(k)\, e^{\frac{i2\pi kn}{Tf}}$$

where $v_r(k)$ is the activity of neuron r at timestep k, $i = \sqrt{-1}$ and $K = Tf$. Each coefficient is a complex number or, equivalently, a two-dimensional vector. Hence the activity sequence for a given neuron is either represented by a vector of $2K$ data points or of $2K$ Fourier coefficients.

To measure the information present at a given time scale, we applied the population activity classifier on the population vector containing the 2N Fourier coefficients for this time scale for the N neurons of the dataset (**Fig. S3G**). To measure information present above a particular time scale T_{max}, we used the Fourier coefficients from 1 to T_{max} for each neuron and concatenated them into a $2NT_{max}$ population vector (**Fig. S3H**). Of note, when evaluating information at particular time scales, we did not apply any temporal filtering steps to avoid artifacts due to the finite size of the filter and preserve the full bandwidth of the data.

Tuning analysis

To quantify the number of neurons significantly tuned to a specific property, we first performed a parametric ANOVA test to identify the neurons which respond significantly more to one of the sounds of interest (e.g. 60, 70 or 80 dB levels across all pure tones for intensity tuning, up vs down modulations in a given frequency range for frequency modulation). We used a threshold of p=0.05. We do not compare the absolute number of neurons tuned to a given property between areas since this will largely reflect the different levels of noise in the data sets and we focus on the properties of significantly tuned neurons.

To measure the tuning of individual units to classes of stimuli (for example up chirps vs down chirps) we used the following modulation index:

$$MI = \frac{\underline{v_a} - \underline{v_b}}{0.5 * (|\underline{v_a}| + |\underline{v_b}|)}$$

Reinforcement learning model

We adjusted a previously published reinforcement learning model [41], to learn discriminations between pairs of temporal inputs. The model receives as inputs the temporal responses for two sounds: ($X_{\mathrm{Go}}(t)$) for the rewarded sound and ($X_{\mathrm{NoGo}}(t)$) for the non-rewarded sound. The model learns the synaptic weights between these input representations and a downstream decision circuit (**Fig. 4A**). This circuit is composed of a Go-unit which outputs the decision

(synaptic weights : w_E) and an inhibitory neuron that provides immediate linear inhibition to the reward neuron (synaptic weights : w_I). The temporal output, y(t), of the model can therefore be described as :

$y(t) = w_E.X(t) - w_I.X(t) - \xi$ where θ is the Heaviside step function, ξ is a time - independent Gaussian random noise process that models stochasticity of behavioral choices. The decision to go is made if the mean activity of the Go-unit within the response window $< y(t) >_t$ is larger than 0.2 ($<.>_t$. denotes time averaging over 0.5s).

The synaptic weights are updated according to a learning rule which compares the reward prediction to the actual reward, assuming that reward prediction corresponds to the mean input received by the Go-unit. The learning rule has three particularities that have been previously shown to be important to account for mouse behavior[41] and compatible with our knowledge of synaptic plasticity rules. First, it is asymmetric : the learning rate is larger when an unexpected reward occurs than when an expected reward does not. Second, it is multiplicative : the learning rate at a given synapse depends on the current weight of that synapse. Finally, it takes into account the known dynamics of the eligibility trace in the striatum [43,64] which is a key target of both AC and TH in discrimination learning[44]. The eligibility trace is a key mechanism in the “neohebbian framework” that aims to explain how synaptic plasticity can accommodate delays between action initiation and environmental feedback. This theory proposes that synapses that undergo pre-post coincidence prior to feedback are tagged via a long-lasting (~ few seconds) eligibility trace. Weight changes will only occur at these tagged synapses if they are subsequently exposed to neuromodulatory feedback before this eligibility trace decays [64]. In line with this, in the striatum, potentiation of synapses is conditioned on dopamine release within a ~3s time window following coincidence of pre- and post-synaptic activity [43]. To implement this in our model, the temporal signal for the model input is convolved with a kernel corresponding to the temporal profile of dopaminergic plasticity gating taken from Yagishita et al [43] before calculation of the weight update.

The learning rule is implemented as :

$$\delta w_E = \lambda f(R - \sigma(w_E - w_I).X)ElTr$$

$$\delta w_I = -\lambda f(R - \sigma(w_E - w_I).X)ElTr$$

where λ the learning rate, R is the action outcome (R = 1 for reward, R = -1 for no reward, σ is the behavioral noise level parameter that sets the models peak performance, $f()$ is the function that implements asymmetric learning such that

$$f(u) = u, \quad \text{if } u < 0$$

$$f(u) = \nu u, \quad \text{if } u \geq 0$$

$\nu > 1$ is the learning rate asymmetry ratio,

$$ElTr = \int_0^{T_{ElTr}} X(u)y(u)D(t-u)du$$

where $D(u)$ is the temporal function shown in **Fig. 4A** and taken from Yagishita et al [43] and T_{ElTr} = 0.5s.

In order to estimate the speed at which the model learns to discriminate between different neural representations, we used as input the population vector time series for two different sounds from a given area. For calcium imaging, we first performed clustering of the response to reduce dimensionality. The model was then run for three independent simulations to average out the stochastic contribution and we evaluated the number of trials to reach 80% based on the average learning curve over these three repeats.

For dimensionality reduction of the population vector, we performed agglomerative hierarchical clustering based on the euclidean distance between each neuron's full temporal response to all stimuli. The number of clusters was established by increasing the number of clusters until the sound-pair RSA matrix constructed from the clusters explained 95% of the variance of the matrix constructed from the full neural population. Clustering was performed independently for each data set and yielded approximately 150 clusters in all areas. AC data displayed in **Fig. 2B** represent clusters rather than single neurons.

Convolutional neural networks

Augmented sound set. In order to train deep neural networks, we created an augmented sound set that covered all the basic parameters explored by the original 140 sound set used in experiments. We first augmented the basic sounds composing the sound set from 140 to 2169. This first step generated the sounds by independently varying all features defining the sounds (frequency, intensity, amplitude modulation direction or period, frequency modulation direction, chord composition). Thereby, a given feature cannot be predicted based on other features as in the experimental sound set. We further augmented the sound set using the approach from [26]. Each 500ms sound is embedded at a random time in a randomly chosen 1.5 s snippet taken from an auditory scene (bus station, park, street…) with a random intensity (average : 53db, std : 7dB). We thus generated a total of 150.000 sounds for the test (6.000), train (110.000) and validation (34.000) sets respectively.

Task definitions. The multi-category task required the network to output a 14-element binary category vector in which 1 indicates that the sound presented belongs to one of 14 categories, divided into 4 groups within which categories are mutually exclusive: frequency range, intensity range, frequency modulation type, and amplitude modulation type. However, all sounds had to receive one label from each group. The group structure was not provided to the network which therefore had to learn that a sound could not be simultaneously high and mid frequency for example. The categories were defined as follows: Frequency range group: high frequency (4-8 kHz) / mid frequency (9-17 kHz) / low frequency 18-38 kHz) / broadband (white noise only). For chords and frequency modulated chirps, the frequency value used for categorization was the average of all frequencies (i.e. middle of the chirp). Intensity range group: high time-averaged intensity (80dB) / mid time-averaged intensity (70dB) / and low time averaged intensity (60 dB). Amplitude modulated sounds were assigned to their closest time-averaged range group. We obtained different overall intensities by ramping sounds sublinearly, linearly or supralinearly. Amplitude-modulation group: Up-ramping/ down-ramping / sinusoidal modulation / no modulation. Frequency-modulation group: Up chirp / Down chirp / no modulation.

We also implemented reduced versions of the multi-category task in which some category groups were excluded. In order to probe the effect of changing the category structure on representations of specific sounds, we selected a subset of sound pairs for each auditory feature that differed only according to one of the 14 categories (**Fig. 6** and **S5**).

The sound identification task required the network to output the identity of each of the 2169 different sounds without any category.

The convolutional autoencoder is a network trained to reproduce with minimal loss its input with the constraint of passing all information through a small, central bottleneck layer. It is composed of an encoder sub-network that processes the input to allow for compression in the bottleneck layer and a decoder sub-network that reconstructs the output from the low-dimensional bottleneck representation.

Architecture definition and training All networks take as input a 2D (time x frequency) matrix of the log-scaled spectrogram of the sound and must produce as output the labels described above. In order to achieve this, a series of convolutional blocks is applied to transform the input. All classification networks were built from a series of 6 blocks composed of the same layers :

- convolution : the input is convolved by a filter whose weights the network must learn, each layer applies multiple filters, generating a 3D matrix (time x frequency channel) from the initial 2D input (free parameters : kernel size, kernel stride, channel number)
- activation : the output of the convolution is passed through a Relu non-linear activation function
- maxpooling : the output of activation is downsampled by taking the maximal value of neighboring values (free parameters : pool size, pool stride)
- dropout : in order to improve the robustness of training, during each training batch a random 50% selection of connections are eliminated. During testing and validation, all connections are active.

After these convolutional blocks, a final 64-node fully connected layer with a Relu non-linearity allows to aggregate information across time, frequency and channel dimensions. The output layer is obtained for the multilabel task by applying a sigmoid function to the fully connected output and for the identification task by applying a softmax function.

The output of the last layer allowed us to calculate the value of the loss function that comprises the error the network makes (categorical cross entropy loss function) and a L1 regularization term in order to improve network robustness. This loss was then back-propagated during training in order to optimize the weights of the connections using the Adam optimizer.

Any given architecture requires arbitration across a wide range of free parameters, most notably the kernel and max pooling size and stride as well as the number of channels in each block. One approach to this problem is to perform a search across architectures to obtain optimal performance on the task. This has allowed optimization on ecologically-relevant tasks to be proposed as a criteria for building deep networks that function like the brain. However we focused on general properties of CNNs and were using a simple task without natural sounds. We therefore chose to assess the generality of our results on various architectures instead of performing an exhaustive search. We also verified the reliability of our results for a

given architecture by using 2 different initialization weights per architecture. The four architectures we evaluated are defined as follows (CV : convolution layer, MP : max pooling layer, FC : fully connected layer, Ker : kernel size) :
(1) Input : 109 x 150; Cv1 : 109 x 150 x 18, Ker(3,3); MP; CV2 : 55 x 75 x 20, Ker(5,5); CV3 : 55 x 75 x 24, Ker(6,6) ; MP; CV4 : 28 x 38x 28, Ker(7,7) ; CV5 : 28 x 38 x 32, Ker(8,8); MP; CV6 : 14 x 19 x 32, Ker(9,9); FC : 64
(2) Input : 109 x 150; Cv1 : 55 x 75 x 18, Ker(3,3); CV2 : 55 x 75 x 20, Ker(5,5); CV3 : 28 x 38 x 24, Ker(6,6); CV4 : 28 x 38x 28, Ker(7,7); CV5 : 14 x 19 x 32, Ker(8,8); CV6 : 14 x 19 x 32, Ker(9,9); FC : 64
(3) Input : 109 x 150; Cv1 : 55 x 75 x 1, Ker(7,7)8; CV2 : 55 x 75 x 20, Ker(7,7); CV3 : 28 x 38 x 24, Ker(7,7); CV4 : 28 x 38x 28, Ker(7,7); CV5 : 14 x 19 x 32, Ker(7,7); CV6 : 14 x 19 x 32, Ker(7,7); FC : 64
(4) Input : 109 x 150; Cv1 : 55 x 75 x 24, Ker(3,3); CV2 : 55 x 75 x 24, Ker(5,5); CV3 : 28 x 38 x 24, Ker(6,6); CV4 : 28 x 38x 24, Ker(7,7); CV5 : 14 x 19 x 24, Ker(8,8); CV6 : 14 x 19 x 24, Ker(9,9); FC : 64

One prominent consequence of the choice of CNN architecture is the way in which the input volume evolves throughout the network. Choosing a large stride in the convolutional or a large window size in the max pooling layer will lead to a shrinkage of the input dimensions (time and frequency). Given that the temporal dimension is preserved in the brain, we examined an architecture in which there is no shrinkage at all of the temporal dimension. To do this, we used the 4 same architectures described above, with the temporal dimension kept constant by setting all strides to 1 and eliminating max pooling. This results in a large expansion of the parameters in the network and impacts training speed although asymptotic performance remains the same (**Fig. 5**).

The convolutional autoencoder receives as input the 2D spectrogram and must output a denoised spectrogram (spectrogram of the central sound without the background noise). The autoencoder was composed of 4 convolutional blocks as previously described in the encoding part and decoding networks, the bottleneck is a fully-connected, 20 node layer. Training was performed with an Adam optimizer, L1 and L2 regularization and MSE as a loss function.

The convolutional neural network trained on word and musical genre recognition was previously published[26] and parameters have been made available at (https://github.com/mcdermottLab/kelletal2018). This network is composed of a central branch that splits into two branches, with one branch trained to identify musical genres and the other branch trained to identify words. In the original paper, the network was shown to achieve human-like performance and to qualitatively reproduce psychophysical measures during these tasks.

Analysis of CNN activations Once the networks had been trained, we analyzed the responses of all nodes in each activation layer to the 140 sounds that were presented during experimental sessions. Each sound generates at a given layer a 3D matrix (time x frequency x channels). By considering the temporal response of each frequency x channel combination we obtained analogs to the temporal response of individual neurons. We then applied the same analysis techniques to these artificial responses as described above for neural recordings. In order to perform decoding which requires multiple presentations of the same

sound, we presented to the network multiple copies of each sound embedded in different noise backgrounds.

Cochlear model

A computational model of the mouse cochlea was implemented based on the seminal model of Meddis[65,66]. The model consists of a cascade of six stages recapitulating stapes velocity, basilar membrane velocity, inner hair cell (IHC) receptor potential, IHC presynaptic calcium currents, transmitter release events at the ribbon synapse, and firing response in auditory nerve fibers (ANFs) including refractory effects. The input model is a sound stimulus (in Pascals). The output is a train of spiking events (in spikes/s) in 590 ANFs innervating 40 IHCs with a characteristic frequency (CF) distributed at regular intervals along the cochlear tonotopic from 5 to 50 kHz, 12 IHCs per octave. This distribution covered 82.8% of the basilar membrane length from 1.2% (apex) to 83.9% (base) in 2.07% increments. According to experimental data, the number of ANFs per IHC (N) was controlled by the relationship N=-0.0038x^2+0.375x+7.9 where x is the IHC location along the basilar membrane such that x=-56.5+82.5 log(CF), with x in percent from the apex and CF in kHz[67]. By adjusting the time constant of the calcium clearance τ_Ca within each IHC synapse[66], ANFs with different spontaneous discharge rate (SR=91.1 $\tau_{Ca}^{2.66}$, with τ_{Ca} in ms and SR in spikes/s) were simulated from 0.5 to 95 spikes/s (21 ± 19.8 spikes/s, mean ± SD) to match the SR distribution reported in mouse auditory nerve.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical results (degrees of freedom, p-values and statistical values) are reported in figure legends or in **Supplemental Table 3**. For statistical analysis of neural data, we performed a bootstrap analysis as detailed above. For statistical analysis of behavioral data provided in the manuscript, the Kolmogorov–Smirnov normality test was first performed on the data. If the data failed to meet the normality criterion, statistics relied on non-parametric tests. We therefore represent the median and quartiles of data in boxplots in all figures, in accordance with the use of non-parametric tests. Ranksum and signed rank: we report the signed rank statistic if the number of replicates is too weak to provide the normal Z statistic.

Supplemental figures and tables

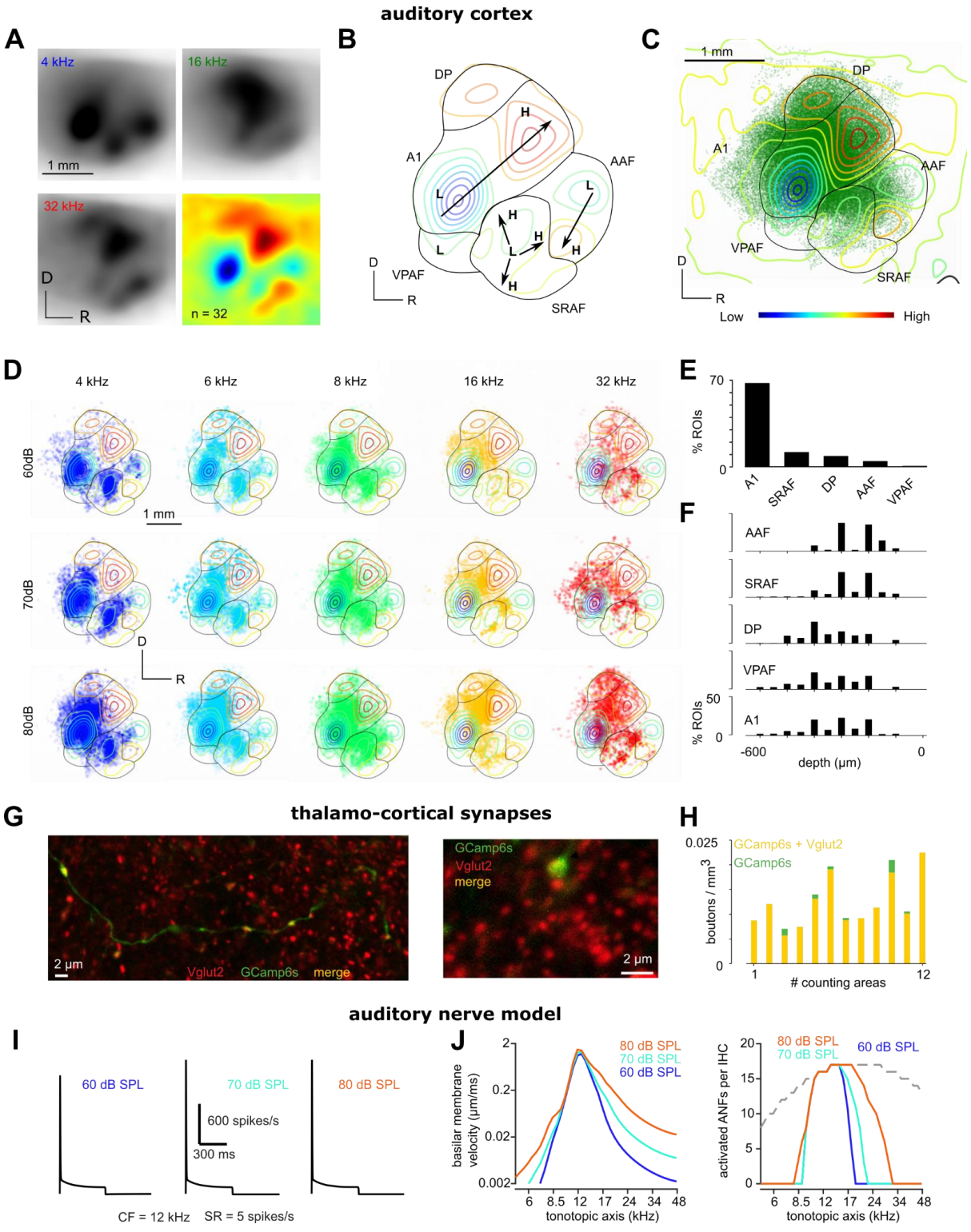

Figure S1. Details of auditory system sampling. A. Mean intrinsic imaging responses (n=32 mice) for 4, 16 and 32 kHz sounds (black) and the subtraction of 32kHz and 4kHz maps (color). This extended data set allowed us to construct a consensus map to align mice included in the study. **B.** Illustration of method used to identify AC subregions based on the tonotopic gradients established in[28]. **C.** Localization of all recorded ROIs on the consensus tonotopic

map with AC subregions. **D.** Localization of responsive neurons to increasing frequency and intensity. Note the larger recruitment with stronger intensity and the spatial shift with frequency. **E.** Proportion of units per subarea. **F.** Depth distribution of units per subarea. **G**. Example thalamocortical axon expressing GCaMP6s merged with Vglut2. Thalamic axonal boutons expressing Vglut2 appear yellow as shown in the magnified region (right). **H.** Density of labeled boutons (Vglut2^{+};GCaMP6s-expressing in yellow; GCaMP6s alone in green) in layer 1 of the AC (12 sample regions; 4 regions per confocal image; means and STD: 0.0122±0.0052, 0.0005±0.0008, density of co-labelled and green only boutons, respectively). **I**. Peristimulus time histogram of an auditory nerve fiber (ANF) with a characteristic frequency equal to that of the presented 12-kHz tone burst (10-ms rise/fall, 500-ms duration) with increasing level from 60, 70 and 80 dB SPL. Note the rapid adaptation of the firing. **J**. Basilar membrane velocity and sound-activated auditory nerve fibers per inner hair cell (IHC) along the tonotopic axis. Note the reduced frequency selectivity with the increasing intensity. Gray dashed line shows the mouse synaptic cochleogram. The criterion for sound-activated auditory nerve fibers was 10 spikes/s above the spontaneous rate.

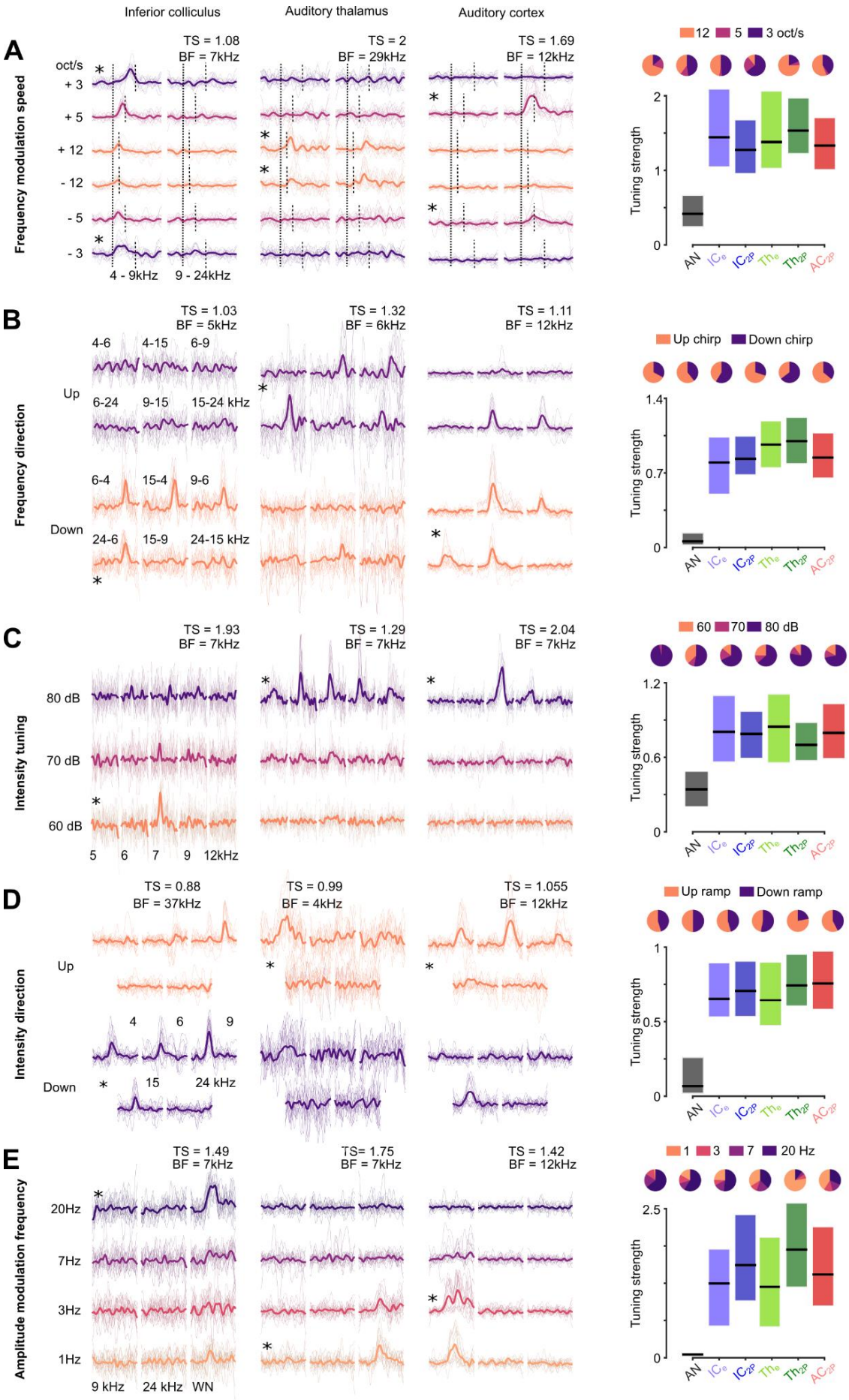

Figure S2. Single cell tuning to diverse acoustic features from cochlea to auditory cortex

A-E. Right: For each tuning property we show the responses of example neurons from the IC, TH and AC to sounds that differ according to that property and provide the tuning strength (TS) and best frequency (BF) for that neuron. Asterisks indicate significant tuning of the neuron to a specific value, for example the leftmost neuron in A is an IC neuron that is significantly tuned to frequency modulation speed with a maximum response for decreasing frequency at 3oct/s. **Left** : Boxplot giving the distribution of tuning strengths across the whole population and piecharts showing the proportion of neurons maximally tuned to each parameter value for significantly tuned neurons.

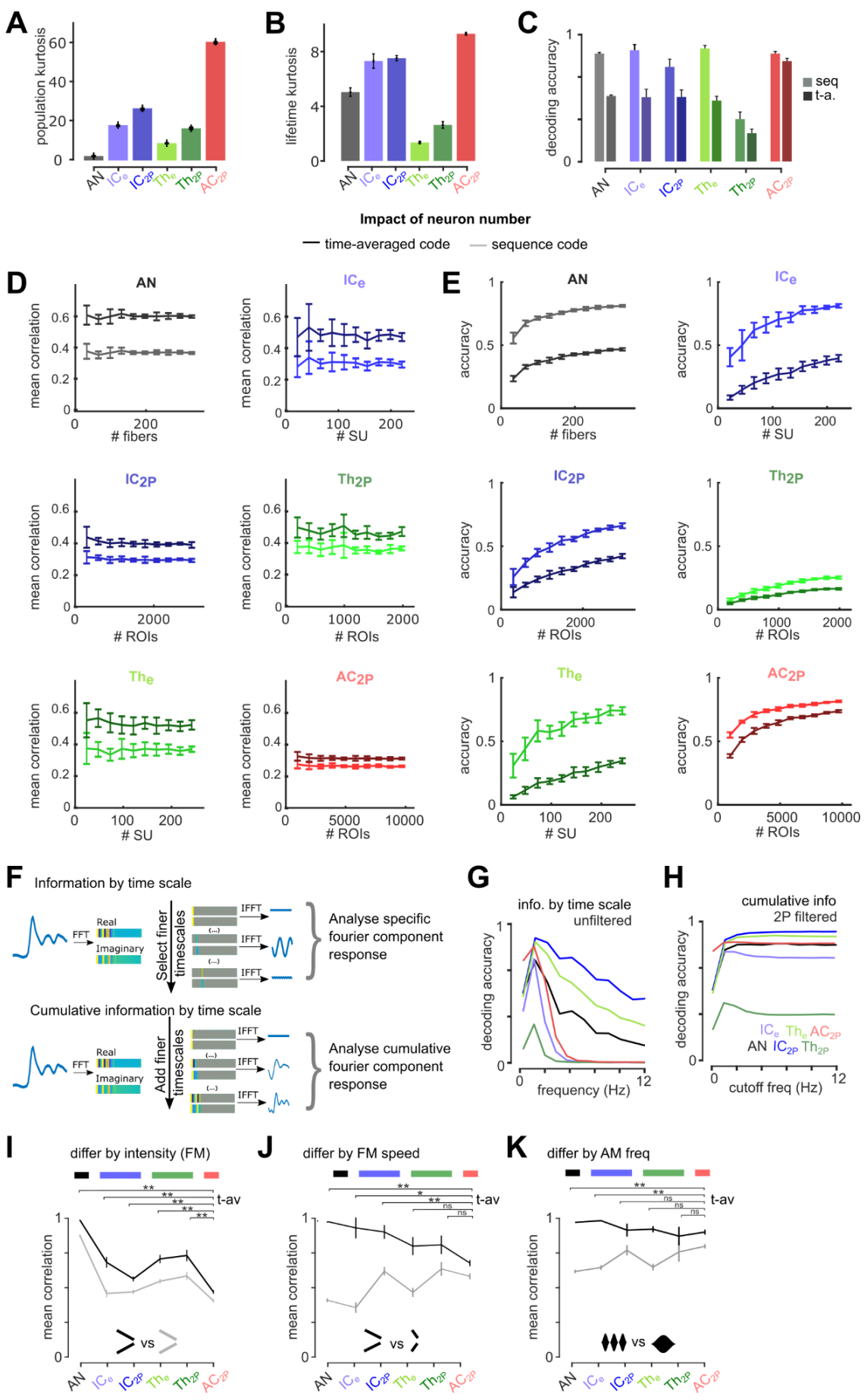

Figure S3. Robustness of correlation and accuracy measures

A-B. Noise-corrected sparseness measured using kurtosis (n=140 sounds for population kurtosis (**A**) and n=all neurons for lifetime kurtosis (**B**)). **C.** Mean sound decoding accuracy for time-averaged and sequence codes. **D.** Noise-corrected correlation for time-averaged and sequence code in each area with varying numbers of sub-selected neurons. **E.** Decoding accuracy for time-averaged and sequence code in each area with varying numbers of sub-selected neurons. **F.** Sketch illustrating the decomposition of population responses by timescale as in G and H. **G.** Mean decoding accuracy based on successive Fourier coefficients of neural responses. 0Hz = time-averaged code. **H.** Same as **G** but for the concatenation of successive Fourier coefficients. The robustness of AC representations to time averaging can be seen in this figure as the fact that accuracy is already very close to plateau value at 0Hz (time-averaged activity level), contrary to other areas which show an increase when adding faster timescales. As expected, 2 photon data only contained information up to 3Hz whereas electrophysiology data was informative even above 12Hz. Importantly, in all brain areas the cumulative information saturated around 3Hz, which is much lower than the known 30 Hz frequency cutoff for AC [11,38]. **I-K.** Mean noise-corrected correlation between sound pairs differing by only one acoustic property : **I.** frequency sweeps with identical frequency content and duration at 60dB vs 80dB, **J**. frequency sweeps with identical frequency content of different duration, **K**. amplitude modulated sounds with same carrier frequencies modulated at 1Hz vs 3Hz (p-value for 100 bootstraps comparing time-averaged correlation of each region to AC, error bars are S.D). P-values and details of statistical tests are given in the **Supplemental Table 3**.

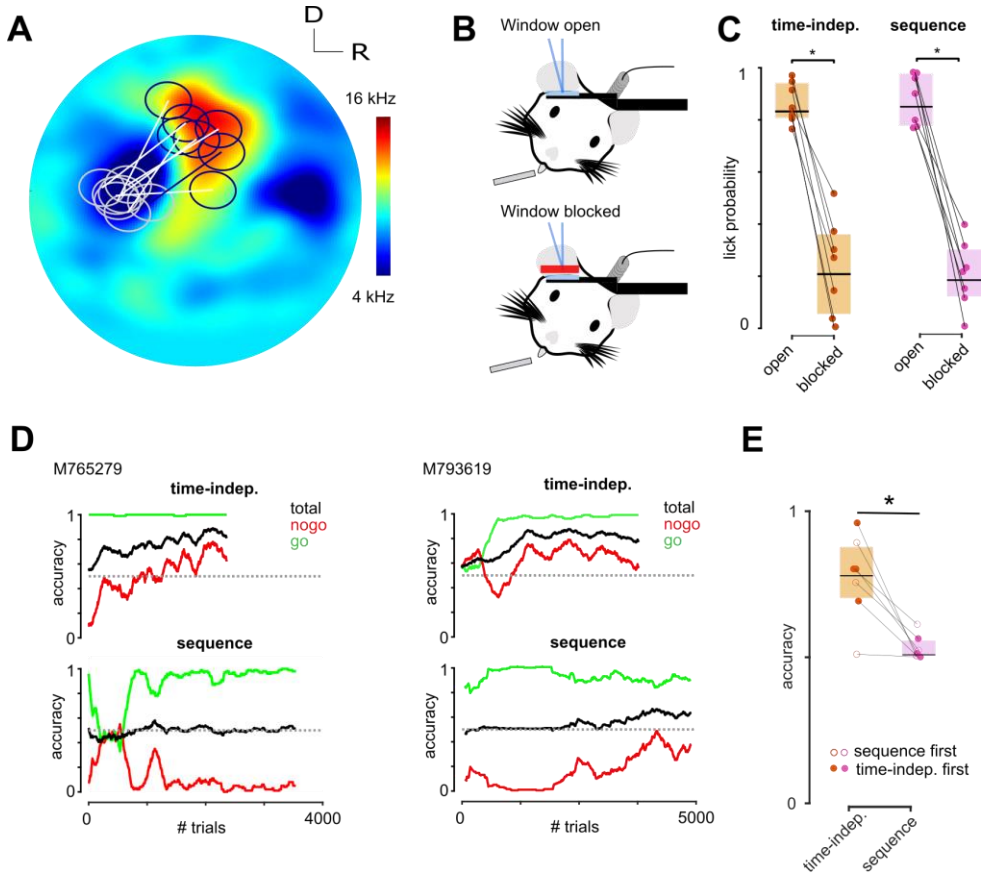

Figure S4. Details of optogenetic cortical stimulation protocol

A. Population average intrinsic imaging map of tonotopic areas in AC showing the localization of all spots used for optogenetic stimulation. **B-C.** Control experiment showing that response to optogenetic stimulation is specific to cortical activation : mice ceased responding to light stimulation when the cranial was blocked by a small cache that left all other light cues intact. Note also that the lick probability for time-averaged or sequence patterns is identical during this initial phase. (paired Wilcoxon test, p = 0.0156, signed rank value = 28, n=7) **D.** Learning curves from two example mice in both tasks. **E.** Accuracy over the last 300 trials for all mice. (paired Wilcoxon test, p = 0.015, signed rank value = 28, n=7).

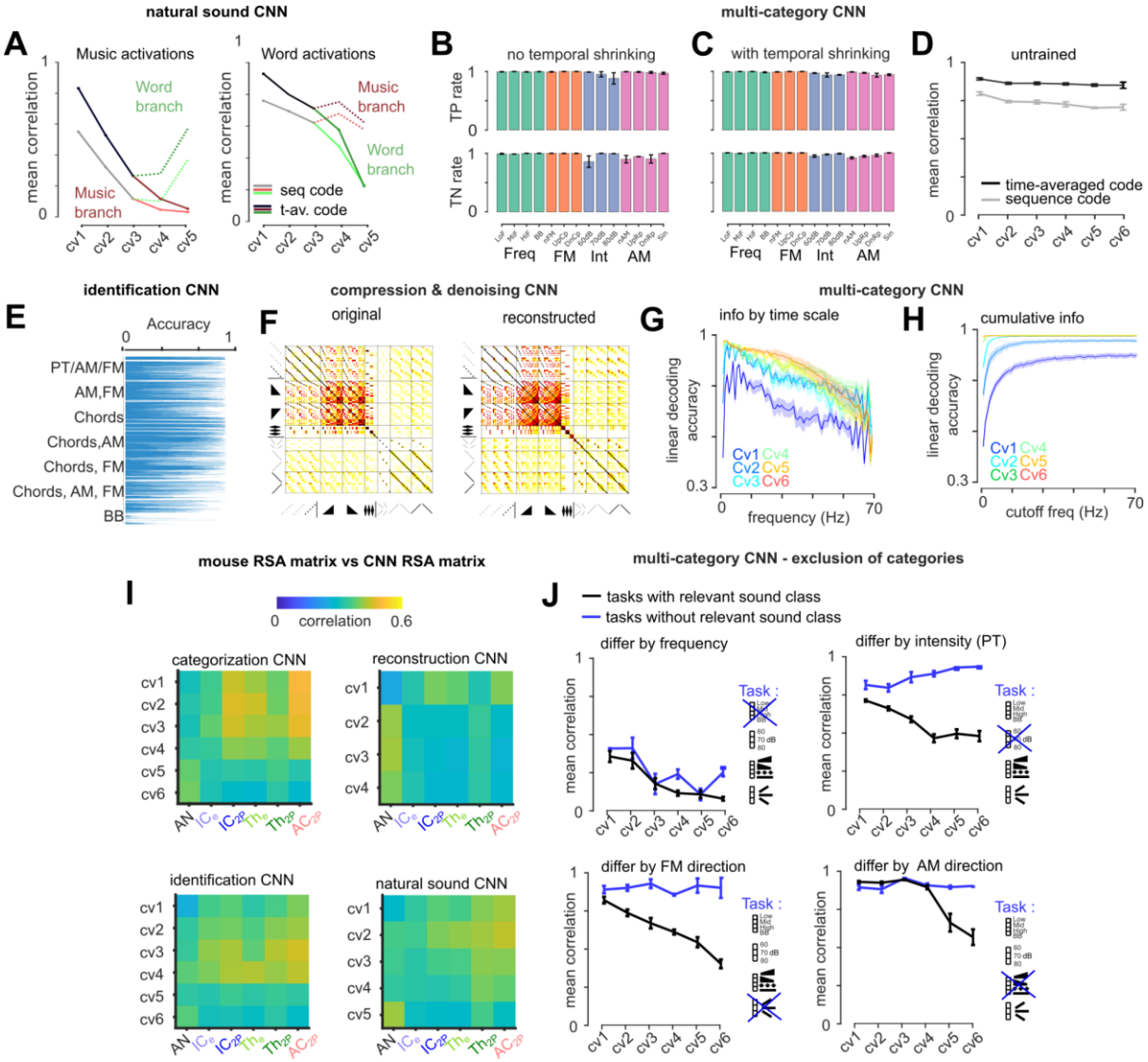

Figure S5. Details of CNN training and task specificity

A. Mean correlations from the network trained on natural sounds from Kell et al for musical snippets (left) or words (center). Note that the branch trained to perform musical genre recognition (music branch - red) decorrelates music but not words and vice versa. **B-C**. Category by category performance of CNNs trained without shrinking of the temporal dimension (B) or with (C) (n=8, error bars are sem). **D.** Mean response correlations from RSA matrices from untrained networks with the same architecture as those trained on the multi-category task (n=8, error bars are sem) **E.** Accuracy of trained network at identifying each individual sound out of 2169. **F.** RSA matrix of original sounds and reconstructed sounds showing that the autoencoder fully preserved the relations between all the sounds. **G.** Mean decoding accuracy based on successive Fourier coefficients of CNN responses. 0Hz = time-averaged code (n=8, shaded areas are sem). **H.** Same as **G** but for the concatenation of successive Fourier coefficients. **I.** Average correlation between the RSA matrices from each region of the mouse auditory system and the different layers of networks performing different tasks. **J**. Mean time-averaged correlation between representations of sounds differing by a specific property throughout networks trained on variations of the multi-class task. Blue curves correspond to networks not trained on the class relevant for the sounds (ex not trained on

frequency for sounds differing only by frequency (top left)) whereas black curves correspond to networks trained on all variations of the task that include the relevant class.

Supplemental Table 1. Details of dataset

Brain region	***Recording method***	***Units recorded***	***Responsive units***	***Number of animals***	***Number of sessions***	***Recorded units per animal (min, mean, max)***	***Recorded units per session (min, mean, max)***
Auditory cortex	Cell body 2 photon calcium imaging	60822	19414 (32%)	7	60	2164 / 8688 / 20631	57 / 1013 / 1782
Auditory thalamus	Axonal bouton 2 photon calcium imaging	39191	3969 (12%)	4	24	1280 / 9287 / 19870	477 / 1632 / 3120
	Single unit electrophysiology	498	484 (97%)	10	33	4 / 49 / 113	2 / 15 / 32
Inferior colliculus	Cell body 2 photon calcium imaging	15312	5936 (39%)	30	101	25 / 510 / 2975	25 / 151 / 495
	Single unit electrophysiology	563	442 (78%)	11	30	10 / 56 / 119	4 / 18 / 54

Supplemental table 2. Sound parameters

			Start freq. (kHz)	Stop freq. (kHz)	Start int. (dB)	Stop int. (dB)	Dur. (ms)
1		blank	NaN	NaN	NaN	NaN	500
2	Pure tones	tono60dB_4kHz	4	4	60	60	500
3		tono60dB_5kHz	5	5	60	60	500
4		tono60dB_6kHz	6	6	60	60	500
5		tono60dB_7kHz	7	7	60	60	500
6		tono60dB_9kHz	9	9	60	60	500
7		tono60dB_12kHz	12	12	60	60	500
8		tono60dB_15kHz	15	15	60	60	500
9		tono60dB_19kHz	19	19	60	60	500
10		tono60dB_24kHz	24	24	60	60	500
11		tono60dB_29kHz	29	29	60	60	500
12		tono60dB_37kHz	37	37	60	60	500
13		tono70dB_4kHz	4	4	70	70	500
14		tono70dB_5kHz	5	5	70	70	500
15		tono70dB_6kHz	6	6	70	70	500
16		tono70dB_7kHz	7	7	70	70	500
17		tono70dB_9kHz	9	9	70	70	500
18		tono70dB_12kHz	12	12	70	70	500
19		tono70dB_15kHz	15	15	70	70	500
20		tono70dB_19kHz	19	19	70	70	500
21		tono70dB_24kHz	24	24	70	70	500
22		tono70dB_29kHz	29	29	70	70	500
23		tono70dB_37kHz	37	37	70	70	500
24		tono80dB_4kHz	4	4	80	80	500
25		tono80dB_5kHz	5	5	80	80	500
26		tono80dB_6kHz	6	6	80	80	500
27		tono80dB_7kHz	7	7	80	80	500
28		tono80dB_9kHz	9	9	80	80	500
29		tono80dB_12kHz	12	12	80	80	500
30		tono80dB_15kHz	15	15	80	80	500
31		tono80dB_19kHz	19	19	80	80	500
32		tono80dB_24kHz	24	24	80	80	500

33		tono80dB_29kHz	29	29	80	80	500
34		tono80dB_37kHz	37	37	80	80	500
35	Pure up ramps	Up4kHz	4	4	60	80	500
36		Up6kHz	6	6	60	80	500
37		Up9kHz	9	9	60	80	500
38		Up15kHz	15	15	60	80	500
39		Up24kHz	24	24	60	80	500
40	Chord up ramps	Up4+6kHz	4, 6	4, 6	60	80	500
41		Up4+9kHz	4, 9	4, 9	60	80	500
42		Up4+15kHz	4, 15	4, 15	60	80	500
43		Up4+24kHz	4, 24	4, 24	60	80	500
44		Up6+9kHz	6, 9	6, 9	60	80	500
45		Up6+15kHz	6, 15	6, 15	60	80	500
46		Up6+24kHz	6, 24	6, 24	60	80	500
47		Up9+15kHz	9, 15	9, 15	60	80	500
48		Up9+24kHz	9, 24	9, 24	60	80	500
49		Up15+24kHz	15, 24	15, 24	60	80	500
50		Up4+6+9+15kHz	4, 6, 9, 15	4, 6, 9, 15	60	80	500
51		Up4+6+9+24kHz	4, 6, 9, 15, 24	4, 6, 9, 15, 24	60	80	500
52		Up4+6+15+24kHz	4, 6, 15, 24	4, 6, 15, 24	60	80	500
53		Up4+9+15+24kHz	4, 9, 15, 24	4, 9, 15, 24	60	80	500
54		Up6+9+15+24kHz	6, 9, 15, 24	6, 9, 15, 24	60	80	500
55		UpmultiHz	4, 6, 9, 15, 24	4, 6, 9, 15, 24	60	80	500
56	Pure down ramps	Down4kHz	4	4	80	60	500
57		Down6kHz	6	6	80	60	500
58		Down9kHz	9	9	80	60	500
59		Down15kHz	15	15	80	60	500
60		Down24kHz	24	24	80	60	500
61	Chord down ramps	Down4+6kHz	4, 6	4, 6	80	60	500
62		Down4+9kHz	4, 9	4, 9	80	60	500
63		Down4+15kHz	4, 15	4, 15	80	60	500
64		Down4+24kHz	4, 24	4, 24	80	60	500
65		Down6+9kHz	6, 9	6, 9	80	60	500
66		Down6+15kHz	6, 15	6, 15	80	60	500
67		Down6+24kHz	6, 24	6, 24	80	60	500

68		Down9+15kHz	9, 15	9, 15	80	60	500
69		Down9+24kHz	9, 24	9, 24	80	60	500
70		Down15+24kHz	15, 24	15, 24	80	60	500
71		Down4+6+9+15kHz	4, 6, 9, 15	4, 6, 9, 15	80	60	500
72		Down4+6+9+24kHz	4, 6, 9, 15, 24	4, 6, 9, 15, 24	80	60	500
73		Down4+6+15+24kHz	4, 6, 15, 24	4, 6, 15, 24	80	60	500
74		Down4+9+15+24kHz	4, 9, 15, 24	4, 9, 15, 24	80	60	500
75		Down6+9+15+24kHz	6, 9, 15, 24	6, 9, 15, 24	80	60	500
76		DownmultiHz	4, 6, 9, 15, 24	4, 6, 9, 15, 24	80	60	500
77	Sinusoid AM modulation	Sin1Hz9kHz	9	9	60 - 80	60 - 80	500
78		Sin3Hz9kHz	9	9	60 - 80	60 - 80	500
79		Sin7Hz9kHz	9	9	60 - 80	60 - 80	500
80		Sin20Hz9kHz	9	9	60 - 80	60 - 80	500
81		Sin1Hz24kHz	24	24	60 - 80	60 - 80	500
82		Sin3Hz24kHz	24	24	60 - 80	60 - 80	500
83		Sin7Hz24kHz	24	24	60 - 80	60 - 80	500
84		Sin20Hz24kHz	24	24	60 - 80	60 - 80	500
85		Sin1HzWhitenoise	WN	WN	60 - 80	60 - 80	500
86		Sin3HzWhitenoise	WN	WN	60 - 80	60 - 80	500
87		Sin7HzWhitenoise	WN	WN	60 - 80	60 - 80	500
88		Sin20HzWhitenoise	WN	WN	60 - 80	60 - 80	500
89	Up chirp varying speed	ChirpUp4kHz60dB100ms	4	9	60	60	100
90		ChirpUp4kHz60dB250ms	4	9	60	60	250
91		ChirpUp4kHz60dB500ms	4	9	60	60	500
92		ChirpUp24kHz60dB100ms	9	24	60	60	100
93		ChirpUp24kHz60dB250ms	9	24	60	60	250
94		ChirpUp24kHz60dB500ms	9	24	60	60	500
95	Down chirp varying speed	ChirpDown4kHz60dB100ms	9	4	60	60	100
96		ChirpDown4kHz60dB250ms	9	4	60	60	250
97		ChirpDown4kHz60dB500ms	9	4	60	60	500
98		ChirpDown24kHz60dB100ms	24	9	60	60	100
99		ChirpDown24kHz60dB250ms	24	9	60	60	250
100		ChirpDown24kHz60dB500ms	24	9	60	60	500
101	Up chirp - 60 dB	ChirpUpclose4kHz60dB	4	6	60	60	500
102		ChirpUpclose4to9kHz60dB	4	9	60	60	500

103		ChirpUpclose4to15kHz60dB	4	15	60	60	500
104		ChirpUpclose4to24kHz60dB	4	24	60	60	500
105		ChirpUpclose6kHz60dB	6	9	60	60	500
106		ChirpUpclose6to15kHz60dB	6	15	60	60	500
107		ChirpUpclose6to24kHz60dB	6	24	60	60	500
108		ChirpUpclose9kHz60dB	9	15	60	60	500
109		ChirpUpclose9to24kHz60dB	9	24	60	60	500
110		ChirpUpclose15kHz60dB	15	24	60	60	500
111	Down chirp - 60 dB	ChirpDownclose6kHz60dB	6	4	60	60	500
112		ChirpDownclose9to4kHz60dB	9	4	60	60	500
113		ChirpDownclose15to4kHz60dB	15	4	60	60	500
114		ChirpDownclose24to4kHz60dB	24	4	60	60	500
115		ChirpDownclose9kHz60dB	9	6	60	60	500
116		ChirpDownclose15to6kHz60dB	15	6	60	60	500
117		ChirpDownclose24to6kHz60dB	24	6	60	60	500
118		ChirpDownclose15kHz60dB	15	9	60	60	500
119		ChirpDownclose24to9kHz60dB	24	9	60	60	500
120		ChirpDownclose24kHz60dB	24	15	60	60	500
121	Up chirp - 80 dB	ChirpUpclose4kHz80dB	4	6	80	80	500
122		ChirpUpclose4to9kHz80dB	4	9	80	80	500
123		ChirpUpclose4to15kHz80dB	4	15	80	80	500
124		ChirpUpclose4to24kHz80dB	4	24	80	80	500
125		ChirpUpclose6kHz80dB	6	9	80	80	500
126		ChirpUpclose6to15kHz80dB	6	15	80	80	500
127		ChirpUpclose6to24kHz80dB	6	24	80	80	500
128		ChirpUpclose9kHz80dB	9	15	80	80	500
129		ChirpUpclose9to24kHz80dB	9	24	80	80	500
130		ChirpUpclose15kHz80dB	15	24	80	80	500
131	Down chirp - 80 dB	ChirpDownclose6kHz80dB	6	4	80	80	500
132		ChirpDownclose9to4kHz80dB	9	4	80	80	500
133		ChirpDownclose15to4kHz80dB	15	4	80	80	500
134		ChirpDownclose24to4kHz80dB	24	4	80	80	500
135		ChirpDownclose9kHz80dB	9	6	80	80	500
136		ChirpDownclose15to6kHz80dB	15	6	80	80	500
137		ChirpDownclose24to6kHz80dB	24	6	80	80	500

138		*ChirpDownclose15kHz80dB*	*15*	*9*	*80*	*80*	*500*
139		*ChirpDownclose24to9kHz80dB*	*24*	*9*	*80*	*80*	*500*
140		*ChirpDownclose24kHz80dB*	*24*	*15*	*80*	*80*	*500*

Supplemental Table 3. Detail of statistical comparisons

Fig 2F & G. Bootstrap comparison of sequence and time-averaged mean correlations in structure X vs in AC						
	AN	ICE	IC	THE	TH	
Sequence	<0.01	0.43	0.12	0.01	<0.01	
Time averaged	<0.01	<0.01	<0.01	<0.01	<0.01	
(Seq - T.A) norm	<0.01	<0.01	0.032	0.01	0.086	
Fig 2H. Bootstrap comparison of RSA matrix similarity in structure X vs in AC						
	AN	ICE	IC	THE	TH	
Seq vs T.A	<0.01	<0.01	0.01	<0.01	0.01	
Fig 2I. Bootstrap comparison of difference between sequence and time-averaged accuracy in structure X vs in AC						
	AN	ICE	IC	THE	TH	
(Seq - T.A) norm	<0.01	<0.01	0.01	0.04	0.15	
Fig 3. Bootstrap comparison of time-averaged mean correlations in structure X vs in AC						
	AN	ICE	IC	THE	TH	
3C - freq	<0.01	0.25	0.61	<0.01	0.27	
3D - int PT	<0.01	0.27	0.71	0.03	0.52	
3E - FM direction	<0.01	<0.01	<0.01	<0.01	0.04	
3F - AM direction	<0.01	<0.01	<0.01	<0.01	0.1	
Fig 6E. Wilcoxon sign rank test of difference RSA matrix correlation between area X and network Y (p-value / signed rank value)						
	AN	ICE	IC	THE	TH	AC
Cat vs ID	0.28 / 64	0.28 / 64	0.003 / 75	0.0016 / 76	0.12 / 67	0.0016 / 76
Cat vs Rec	0.08 / 39	0.01 / 60	0.01 / 60	0.0121 / 60	0.012 / 60	0.012 / 60
Fig 6H. Wilcoxon sign rank test of difference RSA matrix correlation between area X and network Y (p-value / signed rank value)						
	AN	ICE	IC	THE	TH	AC
Full vs NoFreq	0.79 / 65	0.87 / 66	0.19 / 81	0.87 / 70	0.23 / 56	0.87 / 66
Full vs noFM	0.16 / 82	0.10 / 84	0.0047 / 94	0.02 / 90	0.10 / 84	0.01 / 92
Full vs noInt	0.015 / 45	0.0002 / 100	0.0003 / 99	0.0003 / 99	0.0002 / 100	0.0002 / 100
Full vs noAM	0.28 / 79	0.007 / 93	0.0006 / 98	0.0002 / 100	0.007 / 93	0.0047 / 94
Fig S3 I-K Bootstrap comparison of time-averaged mean correlations in structure X vs AC						
S3I - int FM	<0.01	<0.01	<0.01	<0.01	<0.01	
S3J - FM speed	<0.01	0.04	<0.01	0.12	0.22	
S3K - AM freq	<0.01	<0.01	0.66	0.27	0.81	

Supplemental documents

Supplemental Mathematical Derivations (pdf)

References

1. Bizley, J.K., and Cohen, Y.E. (2013). The what, where and how of auditory-object perception. Nat. Rev. Neurosci. *14*, 693–707. 10.1038/nrn3565.
2. Landemard, A., Bimbard, C., Demené, C., Shamma, S., Norman-Haignere, S., and Boubenec, Y. (2021). Distinct higher-order representations of natural sounds in human and ferret auditory cortex. eLife *10*, e65566. 10.7554/eLife.65566.
3. Nelken, I., Rotman, Y., and Yosef, O.B. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. Nature *397*, 154–157. 10.1038/16456.
4. Berger, K.W. (1964). Some Factors in the Recognition of Timbre. J. Acoust. Soc. Am. *36*, 1888–1891. 10.1121/1.1919287.
5. Stecker, G.C., and Hafter, E.R. (2000). An effect of temporal asymmetry on loudness. J. Acoust. Soc. Am. *107*, 3358–3368. 10.1121/1.429407.
6. Neuhoff, J.G. (1998). Perceptual bias for rising tones. Nature *395*, 123–124. 10.1038/25862.
7. Saberi, K., and Perrott, D.R. (1999). Cognitive restoration of reversed speech. Nature *398*, 760–760. 10.1038/19652.
8. Malcolm Slaney and Richard F. Lyon, Apple Hearing Demo Reel, Apple Computer Technical Report #25, Cupertino, CA 95014 (c) 1991 (shorturl.at/lrWX4).
9. Isett, B.R., Feasel, S.H., Lane, M.A., and Feldman, D.E. (2018). Slip-Based Coding of Local Shape and Texture in Mouse S1. Neuron *97*, 418-433.e5. 10.1016/j.neuron.2017.12.021.
10. Chong, E., and Rinberg, D. (2018). Behavioral readout of spatio-temporal codes in olfaction. Curr. Opin. Neurobiol. *52*, 18–24. 10.1016/j.conb.2018.04.008.
11. Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. Nat. Neurosci. *4*, 1131–1138. 10.1038/nn737.
12. Nelken, I., Chechik, G., Mrsic-Flogel, T.D., King, A.J., and Schnupp, J.W.H. (2005). Encoding Stimulus Information by Spike Numbers and Mean Response Time in Primary Auditory Cortex. J. Comput. Neurosci. *19*, 199–221. 10.1007/s10827-005-1739-3.
13. Gao, X., and Wehr, M. (2015). A Coding Transformation for Temporally Structured Sounds within Auditory Cortical Neurons. Neuron *86*, 292–303. 10.1016/j.neuron.2015.03.004.
14. Kopp-Scheinpflug, C., Sinclair, J.L., and Linden, J.F. (2018). When Sound Stops: Offset Responses in the Auditory System. Trends Neurosci. *41*, 712–728. 10.1016/j.tins.2018.08.009.
15. Aponte, D.A., Handy, G., Kline, A.M., Tsukano, H., Doiron, B., and Kato, H.K. (2021). Recurrent network dynamics shape direction selectivity in primary auditory cortex. Nat. Commun. *12*, 314. 10.1038/s41467-020-20590-6.
16. Deneux, T., Kempf, A., Daret, A., Ponsot, E., and Bathellier, B. (2016). Temporal asymmetries in auditory coding and perception reflect multi-layered nonlinearities. Nat. Commun. *7*, 12682. 10.1038/ncomms12682.
17. Hage, S.R., and Ehret, G. (2003). Mapping responses to frequency sweeps and tones in the inferior colliculus of house mice. Eur. J. Neurosci. *18*, 2301–2312. 10.1046/j.1460-9568.2003.02945.x.
18. Kuo, R.I., and Wu, G.K. (2012). The Generation of Direction Selectivity in the Auditory System. Neuron *73*, 1016–1027. 10.1016/j.neuron.2011.11.035.
19. Pressnitzer, D., Winter, I.M., and Patterson, R.D. (2000). The responses of single units in the ventral cochlear nucleus of the guinea pig to damped and ramped sinusoids. Hear Res *149*, 155–166.
20. King, A.J., and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. *12*, 698–701. 10.1038/nn.2308.
21. Ceballo, S., Piwkowska, Z., Bourg, J., Daret, A., and Bathellier, B. (2019). Targeted Cortical Manipulation of Auditory Perception. Neuron *104*, 1168-1179.e5.

10.1016/j.neuron.2019.09.043.
22. Ohl, F.W., Wetzel, W., Wagner, T., Rech, A., and Scheich, H. (1999). Bilateral Ablation of Auditory Cortex in Mongolian Gerbil Affects Discrimination of Frequency Modulated Tones but not of Pure Tones. Learn. Mem. *6*, 347–362. 10.1101/lm.6.4.347.
23. Solyga, M., and Barkat, T.R. (2021). Emergence and function of cortical offset responses in sound termination detection. eLife *10*, e72240. 10.7554/eLife.72240.
24. Li, H., Wang, J., Liu, G., Xu, J., Huang, W., Song, C., Wang, D., Tao, H.W., Zhang, L.I., and Liang, F. (2021). Phasic Off responses of auditory cortex underlie perception of sound duration. Cell Rep. *35*. 10.1016/j.celrep.2021.109003.
25. Dalmay, T., Abs, E., Poorthuis, R.B., Hartung, J., Pu, D.-L., Onasch, S., Lozano, Y.R., Signoret-Genest, J., Tovote, P., Gjorgjieva, J., et al. (2019). A Critical Role for Neocortical Processing of Threat Memory. Neuron *104*, 1180-1194.e7. 10.1016/j.neuron.2019.09.025.
26. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron *98*, 630-644.e16. 10.1016/j.neuron.2018.03.044.
27. Cadieu, C.F., Hong, H., Yamins, D.L., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., and DiCarlo, J.J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol *10*, e1003963. 10.1371/journal.pcbi.1003963.
28. Romero, S., Hight, A.E., Clayton, K.K., Resnik, J., Williamson, R.S., Hancock, K.E., and Polley, D.B. (2020). Cellular and Widefield Imaging of Sound Frequency Organization in Primary and Higher Order Fields of the Mouse Auditory Cortex. Cereb. Cortex *30*, 1603–1622. 10.1093/cercor/bhz190.
29. Murakami, T.C., Mano, T., Saikawa, S., Horiguchi, S.A., Shigeta, D., Baba, K., Sekiya, H., Shimizu, Y., Tanaka, K.F., Kiyonari, H., et al. (2018). A three-dimensional single-cell-resolution whole-brain atlas using CUBIC-X expansion microscopy and tissue clearing. Nat. Neurosci. *21*, 625–637. 10.1038/s41593-018-0109-1.
30. Nahmani, M., and Erisir, A. (2005). VGluT2 immunochemistry identifies thalamocortical terminals in layer 4 of adult and developing visual cortex. J. Comp. Neurol. *484*, 458–473. 10.1002/cne.20505.
31. Yaksi, E., and Friedrich, R.W. (2006). Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. Nat. Methods *3*, 377–383. 10.1038/nmeth874.
32. Pachitariu, M., Stringer, C., and Harris, K.D. (2017). Robustness of spike deconvolution for calcium imaging of neural spiking. 156786. 10.1101/156786.
33. Winer, J.A., and Schreiner, C.E. eds. (2005). The Inferior Colliculus (Springer-Verlag) 10.1007/b138578.
34. Taberner, A.M., and Liberman, M.C. (2005). Response Properties of Single Auditory Nerve Fibers in the Mouse. J. Neurophysiol. *93*, 557–569. 10.1152/jn.00574.2004.
35. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K.D. (2019). High-dimensional geometry of population responses in visual cortex. Nature *571*, 361–365. 10.1038/s41586-019-1346-5.
36. Spearman, C. (1904). The Proof and Measurement of Association between Two Things. Am. J. Psychol. *15*, 72. 10.2307/1412159.
37. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 10.3389/neuro.06.004.2008.
38. Wang, X., Lu, T., Bendor, D., and Bartlett, E. (2008). Neural coding of temporal information in auditory thalamus and cortex. Neuroscience *154*, 294–303. 10.1016/j.neuroscience.2008.03.065.
39. Polley, D.B., Heiser, M.A., Blake, D.T., Schreiner, C.E., and Merzenich, M.M. (2004). Associative learning shapes the neural code for stimulus magnitude in primary auditory cortex. Proc. Natl. Acad. Sci. *101*, 16351–16356. 10.1073/pnas.0407586101.

40. Liu, Y., Zhang, G., Yu, H., Li, H., Wei, J., and Xiao, Z. (2019). Robust and Intensity-Dependent Synaptic Inhibition Underlies the Generation of Non-monotonic Neurons in the Mouse Inferior Colliculus. Front. Cell. Neurosci. *13*, 131. 10.3389/fncel.2019.00131.
41. Bathellier, B., Tee, S.P., Hrovat, C., and Rumpel, S. (2013). A multiplicative reinforcement learning model capturing learning dynamics and interindividual variability in mice. Proc. Natl. Acad. Sci. *110*, 19950–19955. 10.1073/pnas.1312125110.
42. Ceballo, S., Bourg, J., Kempf, A., Piwkowska, Z., Daret, A., Pinson, P., Deneux, T., Rumpel, S., and Bathellier, B. (2019). Cortical recruitment determines learning dynamics and strategy. Nat Commun *10*, 1479. 10.1038/s41467-019-09450-0.
43. Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C.R., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science *345*, 1616–1620. 10.1126/science.1255514.
44. Znamenskiy, P., and Zador, A.M. (2013). Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. Nature *497*, 482–485. 10.1038/nature12077.
45. Verdier, A., Dominique, N., Groussard, D., Aldanondo, A., Bathellier, B., and Bagur, S. (2022). Enhanced perceptual task performance without deprivation in mice using medial forebrain bundle stimulation. Cell Rep. Methods, 100355. 10.1016/j.crmeth.2022.100355.
46. Dalmay, T., Abs, E., Poorthuis, R.B., Hartung, J., Pu, D.-L., Onasch, S., Lozano, Y.R., Signoret-Genest, J., Tovote, P., Gjorgjieva, J., et al. (2019). A Critical Role for Neocortical Processing of Threat Memory. Neuron *104*, 1180-1194.e7. 10.1016/j.neuron.2019.09.025.
47. O'Sullivan, C., Weible, A.P., and Wehr, M. (2019). Auditory Cortex Contributes to Discrimination of Pure Tones. eNeuro *6*, ENEURO.0340-19.2019. 10.1523/ENEURO.0340-19.2019.
48. Kayser, C., Logothetis, N.K., and Panzeri, S. (2010). Millisecond encoding precision of auditory cortex neurons. Proc. Natl. Acad. Sci. *107*, 16976–16981. 10.1073/pnas.1012656107.
49. Walker, K.M.M., Bizley, J.K., King, A.J., and Schnupp, J.W.H. (2011). Multiplexed and Robust Representations of Sound Features in Auditory Cortex. J. Neurosci. *31*, 14565–14576. 10.1523/JNEUROSCI.2074-11.2011.
50. Mittmann, D.H., and Wenstrup, J.J. (1995). Combination-sensitive neurons in the inferior colliculus. Hear. Res. *90*, 185–191. 10.1016/0378-5955(95)00164-X.
51. Norman-Haignere, S., Kanwisher, N.G., and McDermott, J.H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron *88*, 1281–1296. 10.1016/j.neuron.2015.11.035.
52. Yang, Y., DeWeese, M.R., Otazu, G.H., and Zador, A.M. (2008). Millisecond-scale differences in neural activity in auditory cortex can drive decisions. Nat. Neurosci. *11*, 1262–1263. 10.1038/nn.2211.
53. Yang, Y., and Zador, A.M. (2012). Differences in Sensitivity to Neural Timing among Cortical Areas. J. Neurosci. *32*, 15142–15147. 10.1523/JNEUROSCI.1411-12.2012.
54. Musall, S., von der Behrens, W., Mayrhofer, J.M., Weber, B., Helmchen, F., and Haiss, F. (2014). Tactile frequency discrimination is enhanced by circumventing neocortical adaptation. Nat Neurosci *17*, 1567–1573. nn.3821 [pii] 10.1038/nn.3821.
55. LeDoux, J.E., Farb, C.R., and Romanski, L.M. (1991). Overlapping projections to the amygdala and striatum from auditory processing areas of the thalamus and cortex. Neurosci. Lett. *134*, 139–144. 10.1016/0304-3940(91)90526-Y.
56. Chen, L., Wang, X., Ge, S., and Xiong, Q. (2019). Medial geniculate body and primary auditory cortex differentially contribute to striatal sound representations. Nat. Commun. *10*, 418. 10.1038/s41467-019-08350-7.
57. LeDoux, J. (1996). Emotional networks and motor control: a fearful view. Prog. Brain Res. *107*, 437–446. 10.1016/s0079-6123(08)61880-4.
58. Xiong, Q., Znamenskiy, P., and Zador, A.M. (2015). Selective corticostriatal plasticity during acquisition of an auditory discrimination task. Nature *521*, 348–351. 10.1038/nature14225.

59. Francl, A., and McDermott, J.H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nat. Hum. Behav. *6*, 111–133. 10.1038/s41562-021-01244-z.
60. Xu, Y., and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. Nat. Commun. *12*, 2065. 10.1038/s41467-021-22244-7.
61. Friedrich, R.W., and Laurent, G. (2001). Dynamic Optimization of Odor Representations by Slow Temporal Patterning of Mitral Cell Activity. Science *291*, 889–894. 10.1126/science.291.5505.889.
62. Rolls, E.T., and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol. *73*, 713–726. 10.1152/jn.1995.73.2.713.
63. Willmore, B., and Tolhurst, D.J. (2001). Characterizing the sparseness of neural codes. Netw. Bristol Engl. *12*, 255–270.
64. Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., and Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. Front. Neural Circuits *12*, 53. 10.3389/fncir.2018.00053.
65. Lopez-Poveda, E.A. (2003). An approximate transfer function for the dual-resonance nonlinear filter model of auditory frequency selectivity. J. Acoust. Soc. Am. *114*, 2112–2117. 10.1121/1.1605389.
66. Bourien, J., Tang, Y., Batrel, C., Huet, A., Lenoir, M., Ladrech, S., Desmadryl, G., Nouvian, R., Puel, J.-L., and Wang, J. (2014). Contribution of auditory nerve fibers to compound action potential of the auditory nerve. J. Neurophysiol. *112*, 1025–1039. 10.1152/jn.00738.2013.
67. Müller, M., Hünerbein, K. von, Hoidis, S., and Smolders, J.W.T. (2005). A physiological place–frequency map of the cochlea in the CBA/J mouse. Hear. Res. *202*, 63–73. 10.1016/j.heares.2004.08.011.