Music-selective cortex is sensitive to structure in both pitch and time

Dana Boebinger[1,2,3], Josh H. McDermott[1,2,3,4], *Nancy Kanwisher[2,3,4], *Sam Norman-Haignere[5,6,7]

[1] *Speech and Hearing Bioscience and Technology, Harvard University, Cambridge MA 02138 USA*
[2] *Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA 02139 USA*
[3] *McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge MA 02139 USA*
[4] *Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge MA 02139 USA*
[5] *Zuckerman Institute for Brain Research, Columbia University, New York NY 10027 USA*
[6] *Department of Biostatistics & Computational Biology, University of Rochester, Rochester NY 14642 USA*
[7] *Department of Neuroscience, University of Rochester, Rochester NY 14642 USA*

** = Co-senior authors*

ABSTRACT

Music is distinguished from other natural sounds by the presence of relatively discrete notes, which are then organized across pitch and time to convey melody, harmony, and rhythm. Growing evidence suggests that small clusters of neural populations within anterior and posterior human non-primary auditory cortex respond selectively to musical structure. However, it is unclear whether this selectivity reflects short-term musical structure at the level of individual notes, and/or the patterning of notes in pitch and time. We used fMRI voxel decomposition to measure the response of music-selective and non-selective auditory neural populations to synthetic music and drum stimuli whose notes were scrambled in pitch and/or time, disrupting musical pattern structure while largely preserving note-level structure. We observed reliably stronger responses to music with intact pitch and temporal pattern structure in both anterior and posterior music-selective regions bilaterally, but little difference between intact and scrambled music in non-selective populations. Further, only music-selective populations showed reliably stronger responses to note-scrambled music compared with non-music sounds. These results suggest that musical structure involving both individual notes and their patterning over time is specifically represented in localized music-selective neural populations of human non-primary auditory cortex.

INTRODUCTION

Music is a quintessentially human artifact that is distinguished from other sounds by many distinctive acoustic and structural properties. One striking property of music is the presence of discrete “notes” that are characterized by relatively stable pitch contours and spectral timbres, as well as characteristic temporal envelopes (e.g., attack, decay, sustain, release) (Agus et al., 2012; Lindblom & Sundberg, 2007; McAdams et al., 1995; Ogg et al., 2017; Ozaki et al., 2024; Savage et al., 2015). These acoustic properties substantially affect listeners’ perception of musicality. For example, manipulating speech syllables to have more sustained pitch contours causes the utterance to be perceived as more song-like (Sankaran et al., 2024), a well-known effect that is enhanced by repetition (Tierney et al., 2013, 2018). The patterning of notes in music – as for example conveyed by a musical score – is also highly structured in ways that are both common and unique to musical cultures across the world (Mehr et al., 2019; Savage et al., 2015). The pitch of musical notes is typically sampled from a stereotyped musical scale (Krumhansl & Kessler, 1982; Mehr, 2024) unlike speech and most other environmental sounds. The ordering and grouping of pitches into concurrent harmonies and melodic structures follows stereotyped statistical patterns and rules, some of which are common across cultures and others of which are culturally specific (Savage et al., 2015). The temporal patterning of notes is often strongly periodic

(Mehr, 2024; Mehr et al., 2019; Savage et al., 2015), unlike speech, and these periodicities are commonly organized into complex metrical hierarchies, which can be conveyed even in the absence of any pitch information (e.g., drumming). Prior work shows that listeners without any musical training are highly sensitive to all of these musical properties and are able to implicitly learn the statistical structure of their culture's music (Bigand & Poulin-Charronnat, 2006; Rohrmeier et al., 2011), often quite early in development (Hannon & Trainor, 2007).

Understanding how this panoply of rich and distinctive musical structure is encoded in the brain is one of the central challenges of auditory and cognitive neuroscience (Di Liberto et al., 2020; Koelsch et al., 2005; Leaver & Rauschecker, 2010; Norman-Haignere et al., 2015; Peretz et al., 2015; Sankaran et al., 2024). Animal research has identified tuning for many important acoustic primitives, such as pitch (Bendor & Wang, 2005; Bizley et al., 2010), pitch contour (Kilgard & Merzenich, 2002; Yin et al., 2008, 2014), interval timing (Merchant et al., 2013), and spectrotemporal modulation (DeCharms et al., 1998; Depireux et al., 2001; Miller et al., 2002; Woolley et al., 2005) that are likely critical for musical processing. Non-human animals also show greater responses to notes that are surprising given the recent context (Bianco et al., 2024; Yaron et al., 2012) and are able to learn artificial grammars through repeated exposure (Wilson et al., 2013), both of which may be precursors to sensitivity to musical pattern structure observed in the human brain (Di Liberto et al., 2020; Sankaran et al., 2024). Growing evidence suggests that the human brain also has unique adaptations for processing music-specific structure (Landemard et al., 2021). In particular, human non-primary auditory cortex contains small clusters in anterior and posterior non-primary auditory cortex that respond much more strongly to music than to all other natural sounds tested (Angulo-Perkins et al., 2014; Boebinger et al., 2021; Leaver & Rauschecker, 2010; Norman-Haignere et al., 2015; Norman-Haignere, Feather, et al., 2022). The clearest initial evidence for these neural populations came from the use of a statistical decomposition method applied to fMRI responses to natural sounds (Norman-Haignere et al., 2015), and has since been replicated in two other groups of fMRI participants (Boebinger et al., 2021) and confirmed by intracranial recordings that show individual electrodes with clear selectivity for music (Norman-Haignere, Feather, et al., 2022). The response of these music-selective neural populations is much lower to synthetic control stimuli matched to music in its spectrotemporal modulation statistics (Norman-Haignere & McDermott, 2018), unlike in ferrets who show nearly identical responses to natural and modulation-matched music (Landemard et al., 2021). However, the role of these neural populations in encoding different types of musical structure remains relatively unknown.

Recent work suggests that non-primary neural populations in human auditory cortex predominantly integrate information between approximately 200 and 500 ms (Norman-Haignere, Long, et al., 2022; Overath et al., 2015). This finding is consistent with prior studies showing that music selectivity emerges within a few hundred milliseconds after stimulus onset (Leaver & Rauschecker, 2010; Norman-Haignere, Feather, et al., 2022; Norman-Haignere, Long, et al., 2022) and suggests a representation of short-term musical structure on the timescale of a note or two. On the other hand, recent work using "encoding models" have provided some evidence for responses to surprising notes in the auditory cortex (Di Liberto et al., 2020) that are specific to music (Sankaran et al., 2024), suggesting some degree of pattern selectivity. Note surprisal, however, is just one aspect of musical pattern structure and thus only explains a very small fraction of the neural response to music. As a consequence, it remains unclear the extent to which music-selective neural populations respond predominantly to note-level structure or to note pattern structure in pitch and/or time.

A second key question concerns how representations of musical note and pattern structure are functionally and anatomically organized in the auditory cortex. More anterior regions of the human

auditory cortex show selectivity for lower temporal and higher spectral modulation rates and selectivity for tonal sounds compared with non-tonal sounds. More posterior regions of the auditory cortex show selectivity for high temporal and low spectral modulation rates that may play an important role in temporal structure (Hamilton et al., 2018; Hullett et al., 2016; Norman-Haignere et al., 2015; Santoro et al., 2014; Schönwiesner & Zatorre, 2009). A striking property of music-selective neural populations is that they show two distinct widely spaced clusters in anterior and posterior auditory cortex, and it is highly plausible that these two clusters might show distinct selectivity for temporal vs. pitch structure. Prior work has also proposed that there is hemispheric specialization for different spectrotemporal modulation rates (Albouy et al., 2020; Zatorre & Belin, 2001), with the left hemisphere biased towards computations that depend on fast temporal processing and the right hemisphere biased towards computations that depend on pitch and melody. While music selectivity appears to be largely bilateral, prior work measuring music selectivity has focused on relatively short-duration sounds (e.g., 2 seconds) and has not explicitly manipulated temporal and pitch structure.

Here, we used a scrambling approach for examining neural selectivity, inspired in part by prior work characterizing the selectivity of high-level visual cortex (Liu et al., 2010). We reasoned that if music-selective neural populations predominantly encode note-level structure, then responses should be similar for music that has been effectively scrambled in the domain of the score, removing all consistent note pattern structure, but largely preserving the acoustic structure of individual notes. In contrast, if note pattern structure is represented in music-selective neural populations, we reasoned that there might be a stronger response to the intact compared with note-scrambled music and that this effect should be much stronger in music-selective neural populations. To test these predictions, we used functional MRI to measure human auditory cortical responses to synthetic MIDI-generated music clips that were scrambled in pitch and/or time, thus removing virtually all pitch and temporal pattern structure while largely preserving note-level structure. We measured responses to MIDI drum stimuli (without any pitch) which were scrambled in time, thus cleanly isolating temporal pattern structure. We also tested several non-music stimuli to examine whether responses to the note-scrambled stimuli were still greater than those for other non-music sounds. We used component methods developed in prior work (Norman-Haignere et al., 2015) to isolate neural populations that respond selectively to music and compare their selectivity with those of nearby neural populations that respond selectively to other sound categories or acoustic features. We repeated this analyses in both anterior and posterior auditory cortex as well as in the left and right hemisphere to investigate any anatomical or functional organization.

METHODS

Participants

Fifteen young adults (7 self-identified female, 8 self-identified male, 0 self-identified non-binary; mean age = 25.6 years, SD = 4.5 years) participated in the experiment. Of these participants, 12 had some degree of musical training (mean = 10.1 years of training, SD = 6.7). All participants were fluent English speakers and reported normal hearing. The study was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT, and written informed consent was obtained from all participants. Participants were compensated with an hourly wage for their time.

Overview of experimental design

For the current study, participants completed multiple scanning sessions, which included a set of fMRI "localizer" scans to estimate each participant's voxel weights for the components from our previous study (Norman-Haignere et al., 2015), and the main fMRI experiment in which responses were measured to different music conditions. For the component localizer, participants were scanned while they listened to a set of 30 natural sounds. Based on the voxel responses to these sounds, we then estimated each participant's voxel weights that best approximated the previously inferred response profiles from the original study (Norman-Haignere et al., 2015). We then applied these weights to participants' voxel responses in the main experiment, in which they listened to synthetic music clips that had been manipulated by scrambling either the note pitches and/or onset times. This process allowed us to examine how the response components from our previous study (Norman-Haignere et al., 2015) respond to the new stimulus conditions, and thus the extent to which neural selectivity for music reflects note-level vs. pattern structure.

Component localizer

Overview of voxel decomposition

Because a single fMRI voxel pools responses from hundreds of thousands of neurons, voxel responses may be modeled as the sum of the responses of multiple underlying neuronal populations. In our previous work, we found that voxel responses to natural sounds in auditory cortex can be approximated by a weighted sum of six canonical response profiles, or "components:"

$$\boldsymbol{v}_i \approx \sum_{k=1}^{6} \boldsymbol{r}_k w_{k,i}$$

(Eq. 1)

where $\boldsymbol{v}_i$ is the response (a vector) of a single voxel to the sound set, $\boldsymbol{r}_k$ represents the kth component response profile across sounds, which is shared across all voxels, and $w_{k,i}$ represents the weight (a scalar) of component k in voxel i. If we concatenate the responses of many voxels into a sound x voxel data matrix, D, this approximation corresponds to matrix factorization:

$$D \approx RW$$

(Eq. 2)

where R is a sound x component response matrix and W is a component by voxel weight matrix. The component responses and weights were jointly inferred in our prior studies by maximizing the non-Gaussianity of the weights across voxels, akin to standard algorithms for independent components analysis (ICA). The voxel decomposition method is described in detail in our previous paper (Norman-Haignere et al., 2015) and code is available online (https://github.com/snormanhaignere/nonparametric-ica).

Deriving components from scratch requires a large amount of data because both the component responses (R) and weights (W) need to be inferred, which requires using statistics computed across thousands of voxels (requiring multiple participants). However, if the component responses are known (e.g. from a prior study), the voxel weights can be inferred in a single new participant using linear regression (separately for each voxel). In addition, once the component voxel weights are known, the component responses to a new set of sounds, measured in the participants for whom the weights are known, can also be easily inferred using linear regression (separately for each sound). This latter procedure is analogous to identifying a set of voxels that respond to a particular 'localizer' contrast, and then measuring their response to a new set of stimuli.

Stimuli

To reduce the amount of scan time needed to infer the component weights, we chose a subset of 30 sounds from our original 165-sound set that was best able to identify the six components. These sounds were selected by greedily discarding sounds so as to minimize the expected variance of the inferred component weights. Assuming isotropic voxel noise, the expected variance of the weights, inferred by ordinary least-squares (OLS) regression, is proportional to (Dale, 1999):

$$\mathrm{trace}((R^T R)^{-1})$$

(Eq. 3)

where R is the sound x component response matrix. Intuitively, the optimization chooses sounds that have high response variance for each component (e.g., lots of music and non-music sounds for the music-selective component), and where that response variance is relatively uncorrelated across components. Each column/component of R was z-scored prior to performing our greedy optimization (which amounts to iteratively discarding rows of R to minimize the equation above).

Stimulus presentation and scanning procedure

The data for the component localizer came from previous studies on the same participants (n = 2 from Norman-Haignere et al., 2015; n = 7 from Boebinger et al., 2021; and from n = 6 new participants scanned for this study). As a result, there were three versions of the component localizer, differing in the minor details of stimulus presentation and scanning parameters (see **Table 1**), but which otherwise were very similar. All three versions contained the same set of 30 2-second natural sounds. In all versions, stimuli were presented during scanning in a "mini-block design," in which each 2-second sound was repeated multiple times in a row. For versions 1 and 2, stimuli were repeated 3 times in a row, and 5 times for version 3. During scanning, stimuli were presented over MR-compatible earphones (Sensimetrics S14) at 75 dB SPL. Each stimulus was presented in silence, with a single fMRI volume collected between each repetition (i.e. "sparse scanning"; Hall et al., 1999). To encourage participants to pay attention to the sounds, either the second or third repetition in each "mini-block" was 8dB quieter (presented at 67 dB SPL), and participants were instructed to press a button when they heard this quieter sound. Data pre-processing, denoising, and initial GLM analyses were the same as in the main experiment (see "*fMRI data acquisition and analysis*" below).

Despite the fact that participants were run on different versions of the component localizer experiment, we found that the components from Norman-Haignere et al. (2015) explain similar amounts of variance for each group of participants (n = 2 from Norman-Haignere et al. 2015, median R^2 = 0.78, SD = 0.04; n = 7 from Boebinger et al. 2021, median R^2 = 0.80, SD = 0.04; n = 6 new participants, median R^2 = 0.83, SD = 0.04; small effect of group, ϵ^2 = 0.03).

Component weight estimation

To estimate each participant's component weights (W), we multiplied each participant's data matrix of voxel responses to the 30 natural sound stimuli (D_{30}; sound x voxel) by the pseudoinverse of the known component response matrix for those same 30 sounds (R_{30}; sound x component):

$$W = ({R_{30}}^T R_{30})^{-1} {R_{30}}^T D_{30}$$

(Eq. 4)

Multiplying by the pseudoinverse is equivalent to performing ordinary least-squares regression. For visualization purposes (e.g. **Figures 2B**, **3A**, and **4A**), component weights were averaged across participants in standardized anatomical coordinates (Freesurfer's FsAverage template), and plotted with a color scale that spans the central 95% of the weight distribution for each component.

Localizer	Version 1	Version 2	Version 3
MRI scanner	Trio	Prisma	Trio
Head coil	32-channel	32-channel	32-channel
TR	3.59	3.4	3.4
TA	1.19	1.02	1
TE	30ms	33ms	30ms
Flip angle	90	90	90
Number of slices	45 (whole-brain)	48 (whole-brain)	15 (partial coverage)
Slice thickness	2.8 mm	3 mm	4 mm
Slice gap	10%	10%	10%
Matrix	96 x 96	96 x 96	96 x 96
Voxel size	2 x 2 x 2.8 mm	2.1 x 2.1 x 3 mm	2.1 x 2.1 x 4 mm
Simultaneous multi-slice acceleration (SMS)	3	4	none
Number of stimuli	30	192	165
Number of stimulus rep. in each "mini-block"	3	3	5
Number of stimulus blocks per run	30	24	15
Run duration	6.5 min	5.5 min	5.4 min
Number of runs	10	48	33
Number of stimulus repetitions	10	6	3
Number of participants who completed this version	6	7	2

Table 1. Scan parameters for the three versions of the component localizer scan sessions.

Component response estimation

The goal of the current study was to then use participants' inferred component weights to estimate the component responses to each of the stimulus conditions in the main experiment. To do this, we multiplied participants' data matrix of voxel responses to the experimental stimuli (D_{exp}) by the pseudoinverse of the estimated component weights:

$$R_{exp} = D_{exp} W^T (W W^T)^{-1}$$

(Eq. 5)

Unless otherwise noted in the text, we included all voxels in a large anatomical mask region encompassing bilateral superior temporal and posterior parietal cortex (shown as colored regions in **Figures 2B**, **3A**, and **4A**). For analyses comparing different cortical regions (e.g. right and left hemispheres), we estimated the component response for each cortical region separately, using the voxel responses and estimated component weights from the corresponding subset of voxels.

Main experiment

Stimuli

The stimuli for the main experiment consisted of 10-second clips of intact and scrambled versions of synthetic music, as well as several non-music control conditions. Specifically, we created versions of each music clip in which we manipulated (1) the note pitches, (2) the note onset times, and (3) both the note pitches and onset times. So that we could examine the effect of manipulating temporal structure in the absence of pitch structure, we also included a set of synthetic drum stimuli and scrambled versions of each of those stimuli in which we manipulated the note onset times.

To construct these stimuli, we first selected a set of 20 music clips from the stimulus set used in a previous study by our group (Fedorenko et al., 2012). These clips originated from polyphonic musical pieces from a variety of genres common in modern Western musical traditions, each with complex tonal and rhythmic content, and were rendered in MIDI (musical instrument digital interface) format. Each piece contained multiple instrumental tracks, which were all converted to a piano timbre for use in the experiment. The keys of the clips were assigned by drawing (without replacement) from a pool of all 12 major and 12 minor keys (preserving the mode, i.e. "major" vs. "minor"), and the key of each clip was then changed by adding/subtracting a constant (corresponding to the difference in semitones between the original and desired keys) to the note pitches. As a result, the 14 major clips included in the stimulus set spanned all 12 major keys (with C# and G# major occurring twice), and the 6 minor clips were all in different keys.

We then created three additional versions of each music clip (20 unique clips x [1 intact + 3 scrambled versions] = a total of 80 MIDI stimuli), in which the musical structure was manipulated by scrambling either the pitches and/or timing of the notes (**Figure 1**). MIDI represents music as a matrix of notes, with different columns specifying the onset, duration, pitch, and instrumentation of the notes, which makes it possible to manipulate these features independently and to scramble across-note structure without altering acoustic properties of individual notes. In all cases, within-chord structure was preserved by performing the scrambling operation on clusters of notes with a similar onset time (defined for our purposes as within 0.1 beats of each other), rather than on individual notes (**Figure 1A** & **B**). This is a more conservative scrambling procedure than was used in our previous study using related methods (Fedorenko et al., 2012). For the pitch-scrambled music clips, each note cluster was shifted either up or down by 0-2 semitones (sampled from a discrete, uniform distribution), which disrupted both melodic and across-chord tonal

structure (**Figure 1C** & **D**). For time-scrambled music clips, the inter-onset intervals between note clusters were jittered by adding an amount drawn from a uniform distribution spanning ±0.25 beats (the tempi of the MIDI stimuli ranged from 72 to 220 bpm, so 0.25 beats corresponded to between 68-208ms depending on the stimulus example), but with the constraint that no inter-onset interval could be shorter than 0.1 beats. To ensure that the total duration of the time-scrambled clip was the same as the original version, we divided the vector of jittered inter-onset intervals by its mean and then multiplied it by the mean of the original vector of inter-onset intervals. Next, the jittered inter-onset intervals were randomly reordered, and the note onset times of the scrambled piece were derived from these reordered inter-onset intervals (via cumulative summation). This process effectively eliminated any isochronous beat or metrical hierarchy (**Figure 1E**). For the music clips that were both pitch- and time-scrambled, they first went through the time-scrambling procedure and then the pitch-scrambling procedure. This set of 80 MIDI stimuli was supplemented by a set of 20 MIDI drum stimuli, chosen to represent a variety of genres, and versions of these that were time-scrambled following the procedure described above (**Figure 1F**), for a total of 40 MIDI drum stimuli. Each MIDI file was converted into audio using AppleScript and QuickTime Player 7 (Ellis, 2014).

To complement these 120 MIDI music stimuli, we selected 16 additional non-music stimuli to serve as controls for assessing selectivity. Note that we use “selectivity” not to imply a binary distinction between responses to “music” and “non-music” stimuli, but instead to allow for a graded pattern of responses in which the response magnitude to the preferred stimulus category is greater than to all other tested categories (Downing et al., 2007; Fedorenko et al., 2010; Leaver & Rauschecker, 2010; Op De Beeck et al., 2008). Specifically, we chose 4 English speech excerpts (taken from audiobook recordings and radio broadcasts), 4 foreign speech excerpts (German, Hindi, Italian, Russian), 4 animal vocalizations (chimpanzees, cow, sheep, wolf), and 4 environmental sounds (footsteps, ping-pong ball, power saw, copy machine). All stimuli were resampled to 20 kHz, trimmed to be 10 seconds long, windowed with a 25 ms half-Hanning-window ramp, and normalized to have the same RMS.

Because the MIDI scrambling manipulations were intended to be subtle and to preserve some surface aspects of musical structure (e.g. consonance/dissonance amongst simultaneous notes within a chord), an online experiment (via Amazon’s Mechanical Turk) was used to ensure that the manipulation produced detectable changes in musical structure. In this experiment, participants listened to the same 120 MIDI stimuli used in the full experiment (n = 46; 16 self-identified female, 27 self-identified male, 0 self-identified non-binary, and 3 did not answer; mean age = 42.9 years, SD = 11.9 years; an additional 7 participants completed the experiment but were excluded due to performance of <60% in every condition). Stimuli were presented in pairs, with each trial containing an intact and scrambled version of the same MIDI stimulus. Participants were instructed that one of the stimuli was “corrupted by having the melody and/or the note timing disrupted,” and were told to choose which of the two music clips had been “corrupted.” Stimuli were blocked by condition, so that each participant completed four blocks of 20 trials each (i.e. intact vs. pitch-scrambled, intact vs. time-scrambled, intact vs. pitch-and-time-scrambled, and intact vs. time-scrambled drums); participants were told which type of “corruption” to listen for in each block. Prior to completing the experiment participants completed a headphone check to help ensure they were wearing headphones or earphones (Woods et al., 2017); we have found that this headphone check is typically sufficient to obtain performance on par with that of experiments conducted in the lab under more controlled listening conditions (McPherson et al., 2020; Traer et al., 2021; Woods & McDermott, 2018).

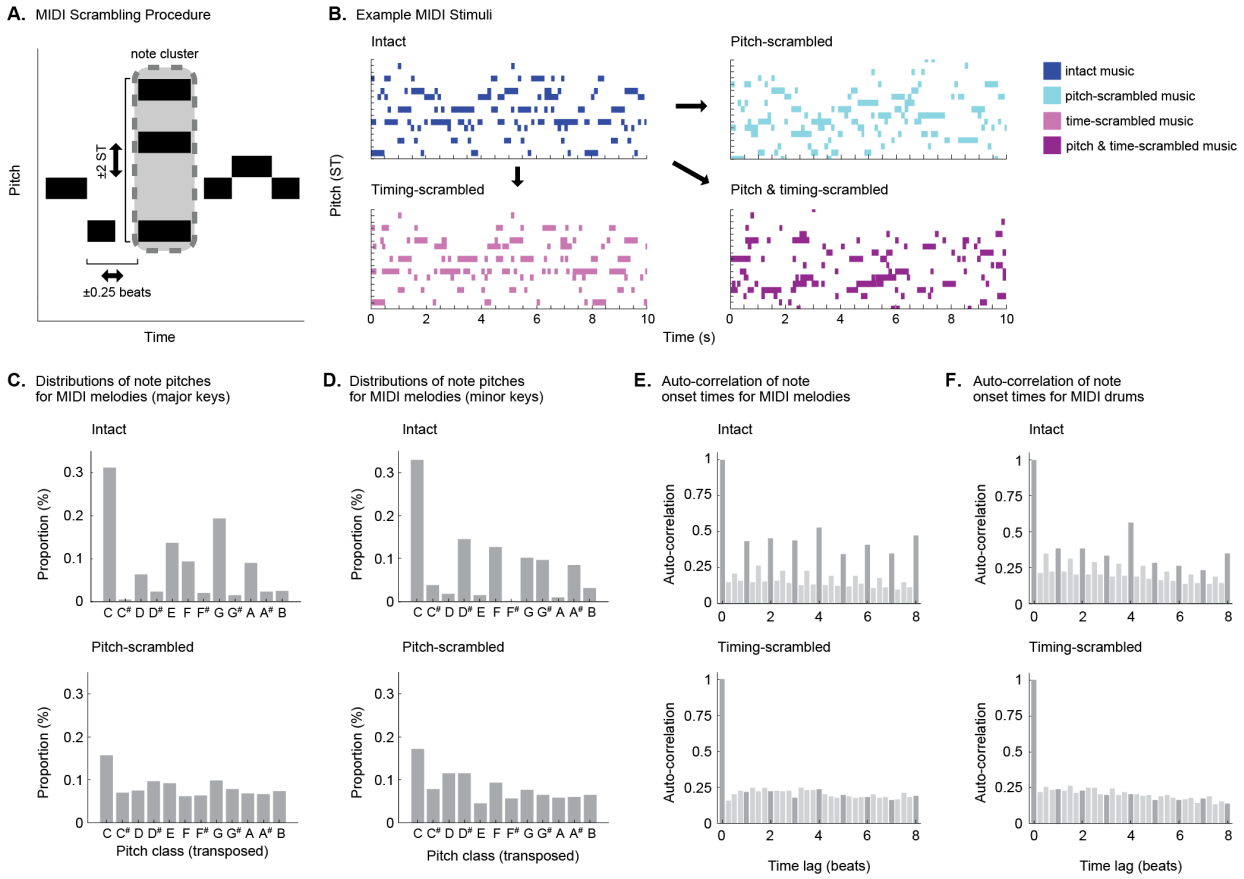

Figure 1. Scrambling procedure for MIDI music stimuli. **A.** Schematic of scrambling procedure. First, notes with the same onset (within 0.1 beats) were grouped together, and scrambled as a unit. For pitch-scrambling, each note cluster was shifted by up to ±2 semitones. For timing-scrambling, the inter-onset intervals (IOIs) between note clusters were jittered by adding an amount drawn from a uniform distribution spanning ±0.25 beats (with the constraint that no IOI could be shorter than 0.1 beats). The jittered IOIs were normalized to ensure the total duration of the timing-scrambled music clip was the same as the original intact version and were randomly reordered. For the music clips that were both pitch- and timing-scrambled, they first went through the timing-scrambling procedure and then the pitch-scrambling procedure. **B.** Intact and scrambled versions of an example MIDI stimulus. **C.** Histograms showing the distribution of note pitch classes averaged across the 14 MIDI clips in major keys, in both their intact (top) and pitch-scrambled (bottom) versions. For the purpose of this analysis, all clips were transposed to C major. Pitch class distributions were weighted by note durations, as computed using the *"pcdist1"* function in the MATLAB MIDI Toolbox (Eerola & Toiviainen, 2004). For the intact music clips (top), certain pitch classes are more common than others, reflecting the tonal hierarchy typical of major keys in Western music. The pitch-scrambling procedure (bottom) disrupts melodic and tonal structure, leveling out the distribution of pitch classes to a certain extent. **D.** Same as **C**, but for the 6 MIDI clips in minor keys. **E.** Histograms showing the auto-correlation of note onset times in units of beats averaged across all 20 intact (top) MIDI stimuli and their timing-scrambled versions (bottom). Onset times were weighted by note durations, as computed using the *"onsetacorr"* function in the MATLAB MIDI Toolbox (Eerola & Toiviainen, 2004). For the intact music clips (top), peaks can be seen on every beat (darker gray bars) that reflect the isochronous beat and metrical structure typical of Western music. The timing-scrambling procedure (bottom) disrupts this metrical structure, resulting in no clear peaks in the auto-correlation function at multiples of the beat. **F.** Same as **E**, but for the 20 MIDI drum stimuli.

Performance was well above chance in all four blocks, indicating that the scrambling manipulation was clearly detectable in all conditions (pitch-scrambling: median = 87.5%, interquartile range or IQR = 20%; time-scrambling: median = 100%, IQR = 5%; pitch-and-time-scrambling: median = 100%, IQR = 5%; time-scrambling drums: median = 100%, IQR = 5%; one-sample Wilcoxon signed rank tests comparing percent correct to chance performance of 50%; all p's < 1e-08, two-tailed).

Stimulus presentation and scanning procedure

During the scanning session for the main experiment, auditory stimuli were presented over MR-compatible earphones (Sensimetrics S14) at 80 dB SPL. The 10-second stimuli were presented back-to-back with a 500ms inter-stimulus interval. To encourage participants to pay attention to the stimuli, each stimulus either ramped up (by 6 dB) or down (by 9 dB) in level over 1 second starting at the 5s point, and the participant indicated whether each stimulus got "louder" or "quieter" via button press. The increment and decrement values were chosen to approximately equate the subjective salience of the level change. Participants' average performance on this task was 97.1% correct (SD = 2%), and performance never fell below 85% correct for any participant on any run.

The scanning session for the main experiment consisted of twelve 7.3-minute runs (all participants completed 12 runs, except for one participant who only completed 8 runs), with each run consisting of thirty-four 10-second stimuli and six 10-second silent periods during which no sound was presented. These silent blocks were the same duration as the stimuli, and were distributed randomly throughout each run, providing a baseline. The thirty-four stimuli presented in each run consisted of five stimuli from of each of the MIDI music conditions ("intact," "pitch-scrambled," "time-scrambled," "pitch-and-time-scrambled," "intact drums," and "time-scrambled drums"), and one stimulus each from the four non-music control conditions ("English speech," "foreign speech," "animal sounds," "environmental sounds"). To maximize the temporal interval between a given MIDI music stimulus and its scrambled versions, only one version of a given MIDI stimulus (i.e. intact, pitch-scrambled, time-scrambled, pitch-and-time-scrambled) was presented per run for all but three participants (the remaining three participants were run on an earlier version of the experiment that was identical except without the constraint maximizing the temporal interval between intact and scrambled versions of a given MIDI stimulus, and with an extra null block in each run). The full set of 136 stimuli was repeated every four runs, meaning that each individual stimulus was presented a total of three times during the course of the 12 experimental runs (twice for the participant who only completed 8 runs). The entire scanning session lasted approximately 2 hours.

fMRI data acquisition and analysis

MRI data were collected at the Athinoula A. Martinos Imaging Center of the McGovern Institute for Brain Research at MIT, on a 3T Siemens Trio with a 32-channel head coil. Because the sound stimuli were long (10 seconds), we used continuous instead of sparse scanning (TR = 2.1 sec, TE = 30ms, 90-degree flip angle, 3 discarded initial acquisitions). Each functional acquisition consisted of 46 roughly axial slices (oriented parallel to the anterior-posterior commissure line) covering the whole brain, with voxel size 2 x 2 x 2.8mm (96 x 96 matrix, 10% slice gap). A simultaneous multi-slice (SMS) acceleration factor of 2 was used in order to minimize acquisition time. To localize functional activity, a high-resolution anatomical T1-weighted MPRAGE image was obtained for each participant (1 mm isotropic voxels).

Preprocessing and data analysis were performed using FSL software and custom Matlab scripts. Functional volumes were motion-corrected, slice-time-corrected, skull-stripped, linearly detrended, and aligned to each participant's anatomical image (using FLIRT and BBRegister; (Greve & Fischl, 2009; Jenkinson & Smith, 2001). Motion correction and functional-to-anatomical registration was done separately for each run. Each participant's preprocessed data were then resampled to the cortical surface reconstruction computed by FreeSurfer (Dale et al., 1999), registered to the FsAverage template brain, and then smoothed on the surface using a 3mm FWHM kernel to improve the signal-to-noise ratio (SNR) of the data by removing local, uncorrelated noise.

GLM-denoise (Kay et al., 2013) was used to further improve SNR and then estimate the response to the stimuli using a general linear model (GLM). For the main analyses between conditions, separate boxcar regressor was used for each of the 10 experimental conditions, which were convolved with a hemodynamic response function (HRF) that specifies the shape of the BOLD response. GLM-denoise assumes a common HRF that is shared across all conditions and voxels, the shape of which is estimated from the data using an iterative linear fitting procedure (Kay et al., 2013). For analyses between individual stimuli within conditions (e.g., comparing responses to intact vs. scrambled versions of the same music clips), separate boxcar regressors were used for each of the 136 individual stimuli.

To remove large-scale correlated noise across voxels, GLM-denoise also includes noise regressors based on principal components analysis (PCA). These PCA-based regressors were derived from the time series of a "noise pool" consisting of voxels with <0% of variance explained by the task regressors (R^2 was measured using leave-one-run-out cross-validation, which is negative when the null hypothesis that all beta weights equal 0 outperforms the model on left-out data). The optimal number of noise regressors was determined by systematically varying the number of principal components included in the model, and estimating the R^2 of each model using leave-one-run-out cross-validation.

GLM-denoise then performs a final fit of the model, and estimates of beta weight variability are obtained by bootstrapping across runs 100 times. The final model estimates are taken as the median across bootstrap samples, and these beta weights were converted to units of percent BOLD signal change by dividing by the mean signal intensity in each voxel. GLM-denoise was applied to the data from all runs for each participant. The resulting beta weights were downsampled to a 2mm isotropic grid on the Freesurfer-flattened cortical surface. To accommodate the partial-brain coverage for two of the participants (see "version 3" in **Table 1**), we limited all analyses to voxels within a large anatomical mask region encompassing bilateral superior temporal and posterior parietal cortex (shown as colored regions in **Figures 2B** and **3A**). All subsequent analyses were conducted in this downsampled 2D surface space, and for ease of description we refer to the elements as "voxels" throughout the paper.

Anterior and posterior auditory cortex anatomical regions of interest (ROIs)

To quantify the effect of scrambling musical structure separately for different subregions of auditory cortex, we divided voxels into anterior and posterior anatomical ROIs. We divided auditory cortex in half by drawing a straight line on the downsampled 2D cortical surface, approximately halfway along Heschl's gyrus (HG) and roughly perpendicular to STG (see white dashed line in **Figure 2B**). While this is a coarse subdivision, it is motivated by the fact that our previous work has found that the music-selective component weights are clustered into "hotspots," one in anterior STG and the other in posterior STG. Thus, it seemed natural to test whether these two clusters might represent different aspects of musical structure.

Statistics

In order to quantify the effect of scrambling on neural responses, we ran 2 x 2 repeated measures ANOVAs on the four melodic MIDI conditions (excluding the drum conditions) with factors "pitch-scrambling" and "time-scrambling". In all cases, we used the Shapiro-Wilk test to verify that the standardized residuals for every combination of factors were normally distributed, and the results of this test were non-significant ($p > 0.05$) unless otherwise noted in the text. When the assumption of normality was not met, the significance of the F-statistic was evaluated using approximate permutation tests. To do this, we randomized the assignment of the data points for

each participant across the conditions being tested 10,000 times, recalculated the F-statistic for each permuted sample, and then compared the observed F-statistic to this null distribution to determine significance. Note that there was no need to check for violations of sphericity, because sphericity necessarily holds for repeated-measures factors with only two levels. For both parametric and non-parametric ANOVAs, effect sizes were quantified using partial eta-squared (η_p^2). ANOVAs were conducted across participants (n = 15).

To compare responses across two conditions, such as intact vs. scrambled drums, two-sided Wilcoxon signed-rank tests were used and effect sizes were calculated as:

$$r = Z/\sqrt{N}$$

(Eq. 6)

where Z is the z-score output of the Wilcoxon signed rank test, and N is the number of pairs (i.e. the number of participants).

When comparing responses to individual stimuli within different conditions, analyses were run across individual stimuli. Wilcoxon signed-rank tests were used when comparisons were between stimuli within two of the four melodic MIDI conditions (i.e., intact vs. scrambled versions of the same music clips), whereas Wilcoxon rank sum tests were used when stimuli were independent (i.e., comparisons between melodic vs. drum stimuli, or any of the music stimuli vs. the non-music control stimuli).

For all figures, we used "within-subject" scatter plots and error bars (Loftus & Masson, 1994), which remove between-subject variance by subtracting each participant's mean response across conditions. The global mean across participants is then added to this de-meaned data before plotting and computing the standard error (SEM).

RESULTS

Music-selective neural populations are sensitive to the patterning of musical notes

We sought to test whether music-selective neural populations in human non-primary auditory cortex are driven by the patterning of notes in pitch or time. We therefore measured responses to MIDI music and drum stimuli that had been scrambled in pitch or time from 15 participants. Because music selectivity is weak when measured using standard voxel-wise fMRI analyses, we utilized techniques from our prior work (Norman-Haignere et al., 2015) to isolate music-selective responses from other, potentially spatially overlapping brain responses. We used our component localizer data to estimate a set of voxel weights (**Figure 2B**) for each participant that best approximated these previously inferred component response profiles. The weights for the music component showed at least two distinct clusters in anterior and posterior non-primary auditory cortex, replicating prior studies. We then used these weights to infer the component responses to the stimuli in the main experiment (**Figure 2C & D**) (by multiplying the pseudoinverse of the weights with the participant's voxel responses from the main experiment). As expected, the component replicated its selectivity for music, responding most strongly to intact music, and minimally to the non-music control stimuli (**Figure 2C & D**). Further, the response of the music component to drum stimuli was significantly higher than the non-music sounds (Wilcoxon signed-rank test, Z = 3.29, p = 8.9e-04, r = 0.85) but significantly lower (Z = 3.41, p = 6.55e-04, r = 0.88) than the pitched music stimuli, replicating previous findings (Boebinger et al., 2021).

Our first question was whether the music-selective component's response was primarily driven by selectivity for note-level structure, such that the response would be similar for intact and note-scrambled music, or whether the component would show evidence of sensitivity to note pattern structure in either pitch or time, and thus respond more weakly to scrambled stimuli that lack this structure. We thus began by comparing the response to MIDI music with both pitch and temporal structure intact, compared with the response to MIDI music that had both pitch and temporal structure scrambled. Strikingly, we found for all 20 of the MIDI clips tested, the response to the intact version was stronger than the response to the fully scrambled version (**Figure 2C**), an effect that was highly significant (Z = 3.92, p = 8.86e-05, r = 0.88). On average the magnitude of the response to pitch and time scrambled stimuli was approximately 22% lower than that for fully intact music, suggesting that note pattern structure makes a non-trivial contribution to this component's selectivity for music. At the same time, we found that the response to pitch and time-scrambled music stimuli was still substantially higher than that for other non-music stimuli. Indeed, the response for every single pitch and time scrambled stimulus was higher than the response for every other non-music stimulus tested, an effect that was highly significant (Wilcoxon rank sum test, Z = 5.08, p = 3.82e-07, r = 1.20), and suggests sensitivity to note level structure as well as note pattern structure. Consistent with this interpretation, the 22% drop in the response of the music-selective component for pitch- and time-scrambled was less than that observed in prior studies for manipulations that disrupt note structure but preserve other low-level acoustic properties (Norman-Haignere et al., 2015; Norman-Haignere & McDermott, 2018).

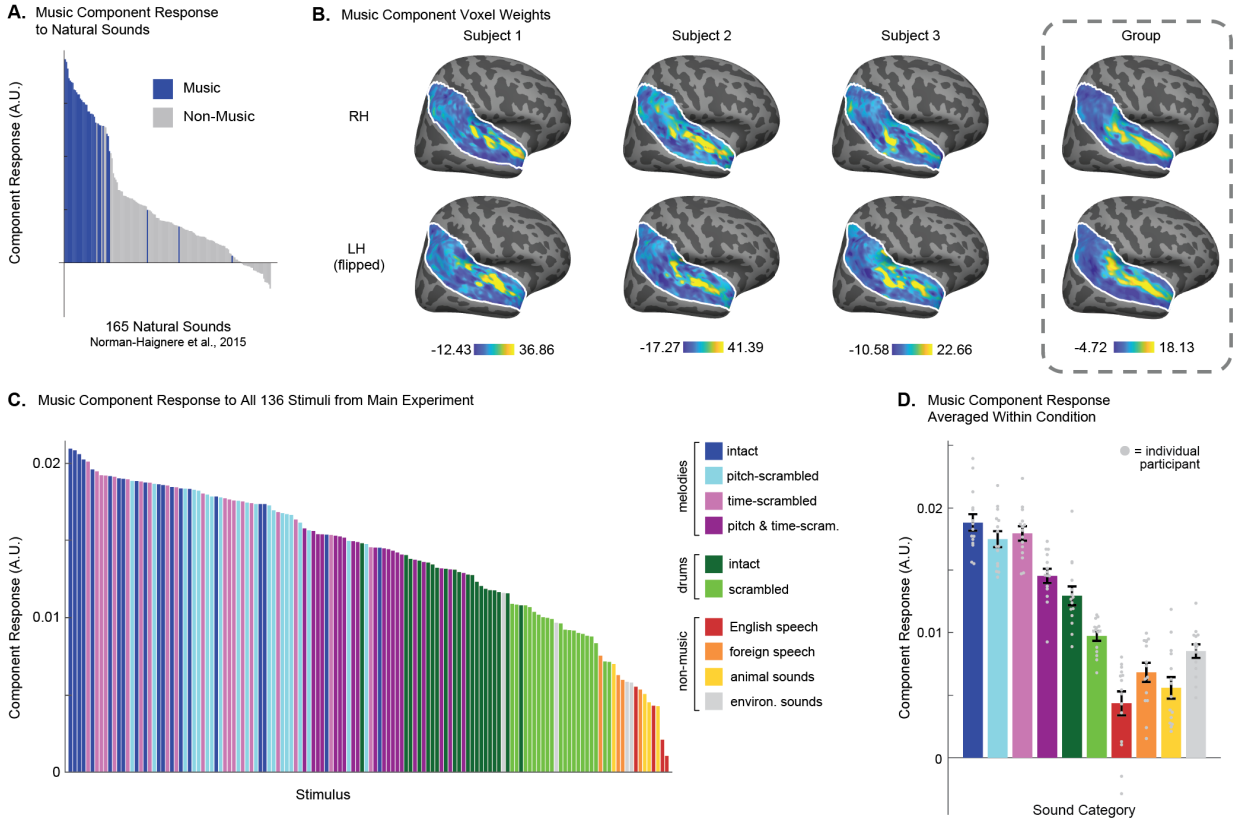

Figure 2. Music component response to MIDI music and non-music control stimuli. **A.** Music component response profile across all 165 natural sounds from Norman-Haignere et al. (2015) study. Sounds categorized as both "instrumental music" and "vocal music" (as determined by raters on Amazon Mechanical Turk) are shaded blue to highlight the degree of music selectivity. **B.** Spatial distribution of music component voxel weights for participants in the current study. Shown for three example participants transformed to standardized anatomical coordinates (FreeSurfer's FsAverage template), and averaged across participants (n = 15). Color scale spans the central 95% of the weight

distribution. **C.** Music component response to each of the 136 individual stimuli in the main experiment, averaged across participants. **D.** Music component response to the stimulus conditions in the main experiment, averaged across participants. Each gray dot corresponds to one participant's component response to a given condition. Scatter plots are within-subject, such that differences between participants are removed by subtracting each participant's mean response across conditions and then adding the global mean across participants. Error bars indicate ±1 within-subject SEM.

Music-selective neural populations are sensitive to note patterns in both pitch and time

Does this sensitivity to the patterning of notes reflect sensitivity to pitch patterns, time patterns, or both? Our cleanest manipulation of temporal structure was our drum stimuli, since these stimuli lacked any pitch structure. Even though on average the response to drum stimuli was lower than the response to pitched music, we still observed a very clear effect of scrambling. Indeed, the response to intact drum stimuli was higher than the response to drum-scrambled stimuli for all 20 of the stimuli tested, leading to a highly significant difference (Wilcoxon signed-rank test, $Z = 3.92$, $p = 8.86e\text{-}05$, $r = 0.88$). On average the response to scrambled drum stimuli was 26% lower than the response to intact drum stimuli, suggesting a substantial contribution of temporal pattern structure to the response of the music-selective component. On the other hand, our most direct manipulation of pitch structure is the difference between time-scrambled and pitch-and-time scrambled conditions, both of which lack coherent rhythmic structure. This difference is highly significant, with a stronger average response to all 20 time-scrambled stimuli than their pitch-and-time-scrambled counterparts (Wilcoxon signed-rank test, $Z = 3.92$, $p = 8.86e\text{-}05$, $r = 0.88$).

Overall, the response of the music-selective component showed clear evidence for sensitivity to pitch pattern structure. We observe clear main effects of both pitch and time scrambling for the MIDI music stimuli (pitch: $F(1, 14) = 39.50$, $p = 2.01e\text{-}05$, $\eta_p^2 = 0.74$; time: $F(1, 14) = 15.20$, $p = 0.002$, $\eta_p^2 = 0.52$), as well as a significant interaction between pitch and temporal structure. The interaction was driven by the fact that scrambling in one domain (pitch or time) was larger if the other domain was scrambled as well ($F(1, 14) = 26.91$, $p < 1.38e\text{-}04$, $\eta_p^2 = 0.66$). This suggests that the presence of either pitch structure or temporal structure is sufficient to produce a comparatively strong response in the music component, an interpretation that is also consistent with our drum scrambling results since in this case there was no consistent pitch structure.

Sensitivity to temporal and pitch patterns is widely distributed across anterior and posterior music-selective neural populations in both the left and right hemispheres

We next sought to test whether there might be differences in sensitivity for pitch pattern and temporal pattern structure between the two hemispheres, motivated by prior hypotheses about hemispheric organization (Albouy et al., 2020; Zatorre & Belin, 2001). To address this question, we repeated our entire analysis using component weights and voxel responses from each hemisphere separately (**Figure 3A**). We found that the response pattern was strikingly similar between the two hemispheres for all of the conditions tested (**Figure 3B**). There were significant main effects of pitch and temporal scrambling for the MIDI music conditions in both hemispheres (left hemisphere, pitch-scrambling: $F(1, 14) = 29.35$, $p = 9.1e\text{-}05$, $\eta_p^2 = 0.68$, time-scrambling: $F(1, 14) = 18.16$, $p = 7.90e\text{-}04$, $\eta_p^2 = 0.56$; right hemisphere, pitch-scrambling: $F(1, 14) = 40.22$, $p = 1.8e\text{-}05$, $\eta_p^2 = 0.74$, time-scrambling: $F(1, 14) = 13.91$, $p = 0.002$, $\eta_p^2 = 0.50$), as well as significant effects of temporal scrambling for drum stimuli in both hemispheres (Wilcoxon signed-rank test, both hemispheres: $Z = 3.92$, $p = 8.86e\text{-}05$, $r = 0.88$). A 3-way ANOVA for the MIDI stimuli that crossed hemisphere (left vs. right) with pitch and time scrambling did not reveal any significant interactions involving the "hemisphere" factor (2-way and 3-way interaction p's > 0.10) (**Figure 3B**). This finding demonstrates that sensitivity to pitch and temporal pattern structure is strongly present in both the left and right hemispheres.

Music-selective neural populations are concentrated in at least two widely spaced clusters in anterior and posterior non-primary auditory cortex. Do these two clusters show differential sensitivity to temporal pattern and pitch pattern information? To answer this question, we conducted an analysis in which we divided auditory cortex along the middle of Heschl's gyrus, roughly perpendicular to STG (see white dashed line in **Figure 3A**), and measured the music component response in anterior and posterior auditory cortex separately (**Figure 3C**). The division was chosen such that individual participants' anterior and posterior clusters were maximally separated into distinct regions. For the MIDI conditions with pitched notes, we observed clear effects of pitch and temporal scrambling and their interaction in both anterior and posterior regions (anterior ROI, pitch-scrambling: $F(1, 14) = 22.12$, $p = 3.4e\text{-}04$, $\eta_p^2 = 0.61$, time-scrambling: $F(1, 14) = 17.22$, $p = 9.8e\text{-}04$, $\eta_p^2 = 0.55$, interaction: $F(1, 14) = 13.58$, $p = 0.002$, $\eta_p^2 = 0.49$; posterior ROI, pitch-scrambling: $F(1, 14) = 53.28$, $p = 3.9e\text{-}06$, $\eta_p^2 = 0.79$, time-scrambling: $F(1, 14) = 11.97$, $p = 0.004$, $\eta_p^2 = 0.46$, interaction: $F(1, 14) = 28.57$, $p = 1.0e\text{-}04$, $\eta_p^2 = 0.67$). Moreover, we did not observe any main effect of region in a 3-way ANOVA crossing region with the two types of scrambling ($F(1, 14) = 0.64$, $p = 0.43$, $\eta_p^2 = 0.04$), nor any significant interactions involving region (2-way and 3-way interaction p's ≥ 0.05). For the drumming conditions, we observed a significant effect of scrambling in both the posterior and anterior regions (anterior ROI: $Z = 3.88$, $p = 1.03e\text{-}04$, $r = 0.87$; posterior ROI: $Z = 3.92$, $p = 8.86e\text{-}05$, $r = 0.88$). However, interestingly, we did find a significant interaction between region and scrambling in a 2-way ANOVA, driven by a larger effect of time-scrambling drum stimuli in the posterior vs. anterior music-selective component ($F(1, 14) = 13.53$, $p = 0.002$, $\eta_p^2 = 0.49$; significance of F-statistic determined using a non-parametric permutation test, see "*Statistics*" section of Methods for more details). Posterior regions also responded more strongly overall to drums, as evidenced by a stronger response to intact drum stimuli compared to intact music in posterior compared to anterior regions ($F(1, 14) = 33.44$, $p = 4.7e\text{-}05$, $\eta_p^2 = 0.70$). These results show that music-selective neural populations in both anterior and posterior auditory cortex are sensitive to the patterning of notes in both pitch and time, but suggest that posterior auditory cortical regions may play a prominent role in representing temporal musical structure, at least in the absence of pitch structure.

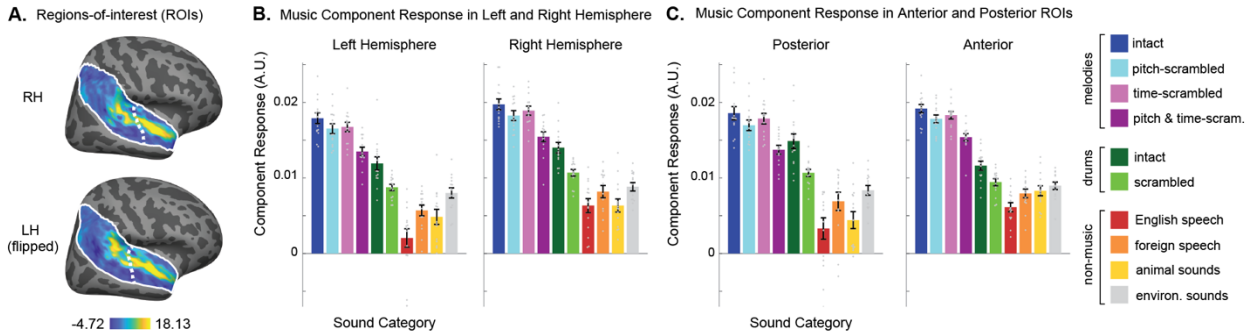

Figure 3. A. Spatial distribution of music component voxel weights, averaged across participants in standardized anatomical coordinates (FreeSurfer's FsAverage template). Color scale spans the central 95% of the weight distribution. The white dashed line shows the dividing line between anterior and posterior ROIs. This line was defined by drawing a straight line on the downsampled 2D cortical surface to maximize the separation of individual participants' clusters of music selectivity. **B.** Music component response to the stimulus conditions in the main experiment, plotted separately for left and right hemispheres. Each gray dot corresponds to one participant's component response to a given condition. Scatter plots are within-subject, such that differences between participants are removed by subtracting each participant's mean response across conditions and then adding the global mean across participants. Error bars indicate ±1 within-subject SEM. **C.** Same as **B**, but plotted separately for the anterior and posterior ROIs shown in **A**.

Sensitivity to musical note patterns is concentrated in music-selective neural populations

Is sensitivity to note pattern structure concentrated in music-selective neural populations or present in other nearby neural populations that also respond strongly to music? In our previous study, we found that voxel responses to a wide variety of natural sounds could be approximated by a set of six canonical response profiles, or "components", one of which was music-selective. Four other components were found to reflect acoustic properties of the sound set (e.g. frequency, spectrotemporal modulation) and were concentrated in and around primary auditory cortex (PAC), consistent with prior results (Da Costa et al., 2011; Herdener et al., 2013; Hullett et al., 2016; Humphries et al., 2010; Norman-Haignere & McDermott, 2018; Santoro et al., 2014; Schönwiesner & Zatorre, 2009). Another component responded selectively to speech (Component 5) and clustered lateral to primary auditory cortex in the middle superior temporal gyrus.

We repeated our analyses for these other five canonical response components (**Figure 4A**). Some of these components showed substantial spatial overlap with music-selective neural populations and produced strong responses to many of the music stimuli tested. For example, Component 4, which shows sensitivity to pitch (Boebinger et al., 2021; Norman-Haignere et al., 2015), had weights that overlapped anterior music-selective neural populations and responded strongly to the melodic pitched MIDI music as well as speech and animal sounds, but only weakly to MIDI drums. Despite its strong response to music, we did not observe a significant effect of scrambling musical structure in the response of this pitch-sensitive component (repeated-measures ANOVA with factors "pitch-scrambling" and "time-scrambling" produced no significant main effects or interaction; all p's ≥ 0.05, see **Table 2**). Thus, even though the spatial distribution of the pitch-sensitive and music-selective components substantially overlap, their response properties dissociate, highlighting the ability of voxel decomposition to isolate neural subpopulations within voxels that show different response characteristics.

We observed weak but significant effects of scrambling in some of the other remaining non-music components (see **Table 2**), however the effects were much smaller than that in the music component in all cases (see **Table 3** for full ANOVA results; all 2-way and 3-way interactions involving "component" were significant: p's < 0.01). We note that scrambling manipulations never perfectly match lower-level acoustic structure and our component decomposition methods likely do not perfectly isolate music selectivity. It is thus not surprising that some non-music selective components show small effects of scrambling, and since component responses are highly reliable, small differences are able to reach statistical significance. Notably, the music-selective component was also the only component that showed evidence for a selective response to note-level structure. Specifically, the music component was the only component where the response to pitch and rhythm scrambled MIDI music was higher than the response to every non-music sound. Collectively, these results suggest that sensitivity to both note level and note pattern structure is concentrated in music-selective neural populations of anterior and posterior human non-primary auditory cortex, bilaterally.

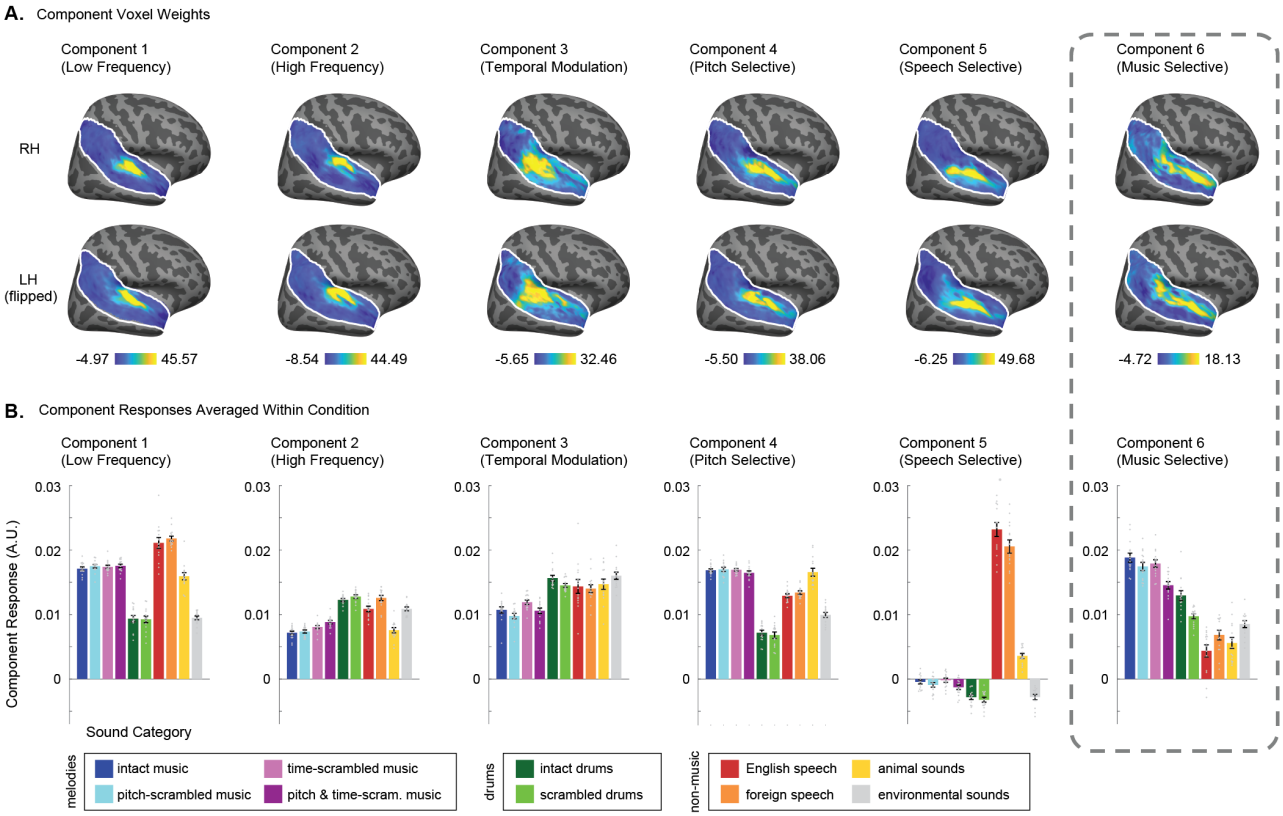

Figure 4. Response of all components to MIDI music and non-music control stimuli**. A.** Spatial distribution of component voxel weights for participants in the current study, inferred using the response components from Norman-Haignere et al. (2015). Voxel weights are averaged across participants in standardized anatomical coordinates. Color scale spans the central 95% of the weight distribution for each component. Music component voxel weights are the same as depicted in **Figures 2B** and **3A**. **B.** Component responses to the stimulus conditions in the main experiment, averaged across participants. Each gray dot corresponds to one participant's component response to a given condition. Scatter plots are within-subject, such that differences between participants are removed by subtracting each participant's mean response across conditions and then adding the global mean across participants. Error bars indicate ±1 within-subject SEM. Music component response is the same as depicted in **Figure 2D**.

Comp.	Pitch-scrambling main effect			Timing-scrambling main effect			Pitch x timing interaction		
	F-value	p-value	η_p^2	F-value	p-value	η_p^2	F-value	p-value	η_p^2
Low freq.	1.93	0.19	0.12	0.66	0.43	0.05	0.69	0.42	0.05
High freq.	12.88**	0.003	0.50	53.96***	3.6e-06	0.79	6.85	0.02	0.133
Temp. mod.	19.75***	6.0e-04‡	0.59	20.42**	0.008‡	0.42	2.06	0.17‡	0.13
Pitch	0.85	0.37	0.06	1.67	0.22	0.11	4.88	0.05	0.26
Speech	45.53***	9.36e-06	0.76	0.18	0.68	0.01	32.11***	5.8e-05	0.70

Table 2. ANOVA results for all auditory cortical response components. ‡ = Shapiro-Wilk test indicated that the standardized residuals for one or more combinations of factors was non-normal ($p < 0.05$), so the significance of the F-statistic was evaluated using a non-parametric permutation test randomizing the assignment of the data points across the relevant conditions 10,000 times (see "*Statistics*" section of Methods). * = Significant at $p < 0.05$, ** = significant at $p < 0.01$, *** = significant at $p < 0.001$.

		Low freq.	High freq.	Temp. mod.	Pitch	Speech
Pitch-scrambling main effect	F-value	19.40***	32.10***	43.68***	31.12***	49.06***
	p-value	6.0e-04	5.8e-05	<1.0e-04‡	6.8e-05	6.2e-06
	η_p^2	0.58	0.70	0.76	0.69	0.78
Timing-scrambling main effect	F-value	7.54*	2.99	1.63	17.63**	13.42**
	p-value	0.02	0.11	0.23‡	8.9e-04	0.003
	η_p^2	0.35	0.18	0.10	0.56	0.50
Component main effect	F-value	0.03	32.91***	17.56***	0.12	277.87***
	p-value	0.87	5.1e-05	9.0e-04‡	0.74	1.3e-10
	η_p^2	0.002	0.70	0.56	0.008	0.95
Pitch x timing interaction	F-value	20.25***	10.67**	20.77***	21.99***	44.28***
	p-value	5.0e-04	0.006	<1.0e-04‡	3.5e-04	1.1e-05
	η_p^2	0.59	0.43	0.60	0.61	0.76
Pitch x comp. interaction	F-value	50.74***	38.71***	11.35***	34.66***	23.37***
	p-value	5.1e-06	2.2e-05	8.0e-04‡	4.0e-05	2.6e-04
	η_p^2	0.78	0.73	0.45	0.71	0.63
Timing x comp. interaction	F-value	27.89***	27.12***	53.00***	9.83**	15.63**
	p-value	1.2e-04	1.3e-04	<1.0e-04‡	0.007	0.001
	η_p^2	0.67	0.66	0.79	0.41	0.53
Pitch x timing x comp. interaction	F-value	17.41***	34.16***	16.49***	14.10**	11.34**
	p-value	9.4e-04	4.3e-05	4.0e-04‡	0.002	0.005
	η_p^2	0.55	0.71	0.54	0.50	0.48

Table 3. Results for 3-way repeated-measures ANOVAs comparing the music component to each of the other auditory cortical response components. ‡ = Shapiro-Wilk test indicated that the standardized residuals for one or more combinations of factors was non-normal ($p < 0.05$), so the significance of the F-statistic was evaluated using a non-parametric permutation test randomizing the assignment of the data points across the relevant conditions 10,000 times (see "*Statistics*" section of Methods). * = Significant at $p < 0.05$, ** = significant at $p < 0.01$, *** = significant at $p < 0.001$.

DISCUSSION

Our study demonstrates that non-primary neural populations just beyond primary auditory cortex show selectivity for how musical notes are patterned across both pitch and time. These effects are concentrated in small, music-selective neural populations and weak or absent in all other auditory populations identified by our analysis, including nearby neural populations that substantially overlap music-selective populations. Yet, within music-selective populations, we find that selectivity for pitch and temporal pattern structure is broadly distributed, present in both and anterior and posterior music-selective populations and in both the left and right hemisphere. On the other hand, scrambled music still produced a substantially stronger response than all other non-music sounds tested (speech, animal calls, and environmental sounds) consistent with sensitivity to note-level structure. These results suggest that neural representations of music-

specific structure at both the single-note and note-pattern level are concentrated in focal, music-specific neural populations of anterior and posterior non-primary auditory cortex, bilaterally.

Selectivity for musical note and note pattern structure in the auditory cortex

A prominent feature of music in nearly all cultures is the presence of musical "notes" (Savage et al., 2015). Individual notes are characterized by spectral profiles that evoke a strong sense of pitch, relatively flat pitch contours, and an amplitude envelope that follows an attack-sustain-decay pattern (Agus et al., 2012; Lindblom & Sundberg, 2007; McAdams et al., 1995; Ogg et al., 2017; Ozaki et al., 2024; Savage et al., 2015). Listeners are highly sensitive to these acoustic properties. For example, spoken speech with more stable pitch contours can be heard as song-like (Sankaran et al., 2024), particularly when repeated (Tierney et al., 2013, 2018). The patterning of notes in music also shows exquisite structure. For example, in most cultures' music, notes tend to be drawn from scales that follow a tonal hierarchy, and to be combined to form sequences that follow a rising/falling melodic contour and tend to have an isochronous beat (Savage et al., 2015). In Western music, this takes the form of major and minor scales with strict tonal hierarchies centered around a "tonic" note. Perceptual studies using "probe tones" and out-of-key "sour notes" have shown that listeners have internalized this tonal structure, consistently rating certain pitches as more fitting within an established melodic or harmonic context (e.g., Krumhansl & Kessler, 1982). Rhythmically, metrical hierarchies in Western music tend to be organized into simple meters like 3/4 and 4/4. These meters are reflected in mental representations of rhythms that favor simpler integer ratios than those evident for listeners from other cultures whose music contains more complex metrical hierarchies (Jacoby et al., 2024; Jacoby & McDermott, 2017). Both behavioral tapping studies and neural electroencephalography (EEG) studies have shown that even untrained listeners are exquisitely sensitive to deviations from their culture's typical musical rhythms (Bigand & Poulin-Charronnat, 2006; Rohrmeier et al., 2011). Studies with infants indicate that enculturation into one's musical culture occurs early in development (Hannon & Trainor, 2007).

How are note-level and note-pattern structure represented in the brain? Prior research suggests human auditory cortex integrates information in natural sounds within a time-limited "integration window", beyond which stimuli have little effect on the neural response (Norman-Haignere, Long, et al., 2022). Primary neural populations integrate information between ~50 and 150 ms and non-primary neural populations selective for music and speech show integration windows spanning approximately 200 to 500 ms, suggestive of a short-term representation of note or syllabic structure. Consistent with these findings, selectivity for music, speech, and singing emerges rapidly in the auditory cortex within a few hundred milliseconds after sound onset. Prior work has also found that manipulating spoken pitch contours to be more note-like enhances cortical responses in regions of the auditory cortex that also show selectivity for music (Sankaran et al., 2024; Tierney et al., 2013), consistent with the hypothesis that note-level structure is an important driver of music selectivity in the auditory cortex. These findings raise the possibility musical pattern level structure might be predominantly represented beyond the auditory cortex, for example, in regions of the frontal cortex that integrate across much longer timescales. This idea is consistent with prior studies that have reported enhanced responses to highly surprising notes or chords in frontoparietal regions of the ventral attention network (Levitin & Menon, 2003, 2005; Maess et al., 2001; Tillmann et al., 2003). On the other hand, recent work has found evidence for enhanced responses to surprising notes in natural music in the auditory cortex (without violations), some of which show evidence of being music-specific (Di Liberto et al., 2020; Sankaran et al., 2024). For example, a recent intracranial study found that notes with more surprising pitches produced larger responses in the auditory cortex, an effect that was not explainable by generic acoustic (e.g., spectrogram) and pitch features (e.g., contour), and correlated with an electrode's overall

selectivity for music vs. speech (Sankaran et al., 2024). However, note surprisal represents just one aspect of musical note pattern structure and only explains a small fraction of the neural response to natural music.

Our study suggests that both note and note pattern structure contribute substantially to auditory cortical representations of music, and that these representations are localized to small patches of music selective neural populations. Music without any consistent note pattern structure, but with intact notes, produced substantially lower responses than music with consistent note pattern structure. At the same time, maximally scrambled music still produced a larger response than other natural sounds such as speech and environmental music, suggestive of a response to note-level structure. While scrambled music differs acoustically from speech and other environmental sounds, prior work using more controlled manipulations suggests that this difference cannot be explained by generic forms of spectrotemporal tuning that predict responses in primary auditory cortex. For example, prior studies have found much stronger responses to natural music than spectrotemporally matched music in non-primary auditory cortex (Norman-Haignere & McDermott, 2018), and the magnitude of this difference is substantially larger than that what we observed in the current study when comparing intact vs. scrambled music. This suggests that at least part of the scrambling effect is explained by sensitivity to note-level structure, rather than exclusively reflecting acoustic differences between conditions. Moreover, the greater response to natural music than modulation-matched stimuli is not seen in the non-primary auditory cortex of ferrets, suggesting that it is at least somewhat specific to the human auditory cortex (Landemard et al., 2021).

What types of musical pattern structure might underlie the selectivity we observe here? Our scrambling procedure was careful to largely preserve the acoustics of individual notes. For example, notes within the same chord were shifted together in pitch and time, unlike a prior study that used scrambling (Fedorenko et al., 2012), which reduces the likelihood that simple acoustic confounds such as differences in spectrotemporal modulation can explain our results. At the same time, our scrambling manipulations were intended to remove as much note pattern structure as possible. Our pitch scrambling manipulations removed scale structure that characterizes tonal music, as well as statistics of melodic contour and voice leading properties that characterize typical music. Our temporal scrambling manipulation removed both periodicity and higher-order metrical structure, and thus both could potentially influence music selectivity.

Our results, therefore, suggest that music-selective neural populations are strongly influenced by both note-level and note pattern structure. How can this finding be reconciled with prior work suggesting short timescale representations and integration windows for music? Many prior studies have used shorter musical stimuli (e.g., 2-seconds) than those tested here (e.g., 10 seconds), and sensitivity to musical structure may require a long time to build-up. Note pattern structure might also partially reflect short-term pattern statistics on the scale of a note or two. For example, there might exist neural sensitivity for pitch intervals between nearby notes that are more common in tonal contexts (Krumhansl & Kessler, 1982; Mehr et al., 2019; Temperley, 2008, 2014), or temporal intervals that exhibit short-term temporal periodicity (Jacoby & McDermott, 2017; Mehr et al., 2019; Savage et al., 2015).

Methods for investigating sensitivity to musical structure

How should one investigate neural sensitivity for musical structure? In this study, we used a scrambling manipulation, inspired in part by prior work on visual and speech processing. The logic of this approach is that if there are neural populations that respond to the presence of statistical structure that characterizes natural stimuli (e.g., music, speech, language, faces), they will

respond more weakly when this structure is altered. For example, the fusiform face area (FFA) responds more strongly to images of intact faces than to faces in which the eyes, nose, and mouth have been spatially rearranged (Liu et al., 2010), revealing a sensitivity to typical facial structure. Analogously, studies that scrambled words within sentences have shown that responses in language regions in the frontal and temporal lobes reflect the combination of words into meaningful clauses and sentences (Mollica et al., 2020; Pallier et al., 2011), and studies scrambling foreign speech have found regions of the superior temporal gyrus that respond more to intact speech structure (Overath et al., 2015).

A few fMRI studies have applied similar scrambling techniques to music stimuli (Abrams et al., 2011; Farbood et al., 2015; Fedorenko et al., 2012), though these studies have not been able to isolate music-selective neural populations and used very different scrambling methods than those employed here. One particularly relevant prior study reported greater responses to intact compared with scrambled music in voxels of the temporal and frontal lobes (Fedorenko et al., 2012). However, these voxels were scattered throughout auditory cortex and also produced strong responses to visually presented sentences, suggesting that they may not be specifically tuned to music per se, unlike the clusters of music-selective voxels found here. Prior work suggests that music selectivity is weak in individual voxels due to overlap between music-selective and non-selective neural populations, and component methods, like those used here, are thus needed to isolate music-selective neural populations with fMRI. The scrambling effects observed in Fedorenko et al. (2012) may partially reflect differences in lower-level acoustics because the scrambling manipulation was coarser than that used here. For example, the pitches of individual notes were randomized, which substantially disrupted the harmonicity of simultaneous notes. Two other studies used naturalistic music and applied their scrambling methods to the audio files themselves (Abrams et al., 2011; Farbood et al., 2015; Norman-Haignere et al., 2015), rather than using synthetic MIDI music, which can result in acoustic artifacts and stimuli that sound highly unnatural. We note that no scrambling method perfectly preserves all lower-level acoustics, including those used in the present study, and our component methods are also likely imperfect and thus do not completely isolate music selectivity in a single component. Thus, it is not surprising that we occasionally observed small differences between our various scrambling conditions in the other non-music-selective components. However, these effects were much smaller than that observed in our music-selective component, even in components that are spatially proximate and overlap substantially with music-selective neural populations.

Another approach for examining musical structure is to examine neural responses to sounds that violate this structure. These effects have often been interpreted in terms of theories of prediction error and predictive coding, in which neurons respond to and encode events that are unlikely given the prior context (e.g., an out-of-key note or chord) (Heilbron & Chait, 2017; Koelsch et al., 2019). These violations often result in strong responses in regions of frontal and parietal cortex associated with the ventral attention network (Levitin & Menon, 2003, 2005; Maess et al., 2001; Tillmann et al., 2003), but these regions also produce strong responses to violations in non-musical contexts suggesting that these effects are not specific to music per se (Doeller et al., 2003; Schubotz & von Cramon, 2002).

Violation studies have traditionally measured responses to individual events that are highly unnatural given the context. An alternative approach that has grown in popularity is to compute multiple different features from a natural stimulus such as music and then measure the unique variance that each feature explains using an encoding model (Di Liberto et al., 2020; Sankaran et al., 2024). However, natural stimuli such as music are complex and multidimensional, and the mapping between sound features and the neural response is often highly nonlinear. As a

consequence, individual features (e.g., note surprisal) often only explain a tiny fraction of the neural response to a complex stimulus such as music (Sankaran et al., 2024). Scrambling methods and encoding models are thus complementary. Scrambling methods provide a way to test sensitivity to a large class of structural properties without the need for strong and potentially inaccurate assumptions about how each of these features are encoded. The results from this work can then motivate more detailed encoding model studies investigating how each of those features is represented.

To what extent can the results of our study be explained in terms of prediction error and musical surprisal? The naïve hypothesis would be that scrambling musical structure would result in stronger prediction-error responses because it violates listeners' expectations (Di Liberto et al., 2020; Koelsch et al., 2019; Sankaran et al., 2024; Vuust et al., 2022), which is inconsistent with our data. However, it is possible that removing all note pattern structure causes the predictions in the auditory cortex to weaken, suppressing prediction error responses that would be present for natural music (Quiroga-Martinez et al., 2019). This hypothesis predicts that small, local perturbations of the stimulus may result in enhanced neural responses while more global, longer perturbations suppress responses, leading to lower responses to scrambled stimuli over long timescales, a prediction that could be tested in future work.

A challenge with attempting to distinguish between representations of pitch and time in music is that melody and rhythm are tightly intertwined in music. Drum stimuli are useful in this respect because they effectively remove any influence from musical pitch, cleanly isolating effects of temporal pattern processing. However, isolating effects of pitch patterning is more challenging because the constituent pitches of a melody necessarily unfold over time. As a consequence, the perception of melody and harmony was almost certainly influenced by time scrambling and the perception of rhythm was almost certainly influenced by pitch scrambling. Given this, our most direct measure of pitch pattern processing is arguably the difference between time-scrambled and pitch-and-time-scrambled conditions, since in both cases, there was no consistent rhythm for pitch scrambling to disrupt.

The relatively small pairwise differences we observed in the response to the intact vs. pitch- or time-scrambled conditions (**Figure 2C**) could be due to the subtlety of our scrambling manipulation. While we did verify that listeners were readily able to detect both pitch- and time-scrambling in our stimulus set (see the behavioral experiments described in the *"Main Experiment: Stimuli"* section of the Methods), the scrambled stimuli are still clearly identifiable as "music." Further, previous studies have shown that only a minimal amount of information is needed for listeners to be able to fully instantiate complex tonal (Cuddy & Badertscher, 1987; Dowling, 1978; Oram et al., 1995; Smith & Schmuckler, 2004) and metrical structure (Brochard et al., 2003; Desain & Honing, 2003; Palmer & Krumhansl, 1990; Povel & Okkerman, 1981). Listeners might therefore be able to draw on their vast amount of experience with Western tonal music in order to form expectations for what the scrambled stimuli "should" sound like. This perceptual "unscrambling" might have reduced the effects that we observed. While it could have been the case that the music component would respond strongly only when both pitch and time structure are present, it is also possible that the presence of either pitch or time structure would be sufficient to elicit a strong neural response, as we observe here. This latter hypothesis is analogous to prior studies in vision that show fairly strong responses in face-selective brain regions when any face structure is present (e.g. Arcurio et al., 2012; Tong et al., 2000).

Functional organization of auditory cortex

A striking feature of music selectivity is the presence of at least two distinct widely-spaced clusters in anterior and posterior auditory cortex. Prior research suggests that there are prominent differences in the representations of sound between anterior and posterior non-primary auditory cortex. For example, anterior auditory cortex has been found to contain neural populations that respond to tones with resolved harmonics that have a strong sense of pitch (Norman-Haignere et al., 2013). The most anterior portion of these tone-selective regions also shows strong sensitivity to the presence of pitch changes (Norman-Haignere et al., 2013; Patterson et al., 2002), which may play a role in melody perception. Consistent with these findings, more anterior regions show tuning for fine spectral modulations characteristic of tonal sounds, while more posterior regions show tuning for rapid temporal modulations characteristic of impulsive sounds (e.g., drums) (Hamilton et al., 2018; Hullett et al., 2016; Norman-Haignere et al., 2015; Santoro et al., 2014; Schönwiesner & Zatorre, 2009). Spectrotemporal modulation is distinct from sensitivity to the pitch and temporal pattern structure, as demonstrated by the fact that our scrambling manipulations largely preserve the spectrotemporal modulation structure of the tested sounds while disrupting the perception of harmony, melody, and rhythm. Nonetheless, it seemed a priori highly plausible that anterior regions of the auditory cortex might show greater sensitivity to pitch pattern structure and posterior regions might show greater sensitivity to temporal pattern structure.

Our results demonstrate that sensitivity to both temporal and pitch pattern structure is present in both anterior and posterior music-selective neural populations. Anterior music-selective populations, for example, showed stronger responses to intact compared with scrambled drum stimuli that lack pitch structure entirely, and these effects were much stronger than in non-music-selective neural populations of non-primary auditory cortex. In addition, posterior regions showed clear sensitivity to pitch scrambling manipulations, which were much stronger than those in tone-selective neural populations of anterior non-primary auditory cortex. At the same time, we did observe weak trends that are consistent with the hypothesized anterior-posterior division. In particular, posterior music-selective regions showed a stronger overall response to drum stimuli, consistent with prior reports that posterior regions show sensitivity to rapid temporal modulations characteristic of impulsive sounds (Hamilton et al., 2018; Hullett et al., 2016; Norman-Haignere et al., 2015; Santoro et al., 2014; Schönwiesner & Zatorre, 2009). Notably, posterior regions also showed a stronger drum scrambling effect than anterior music-selective populations, consistent with a weak bias for temporal pattern sensitivity. Thus, while our results rule out the possibility of a strong anatomical anterior-posterior division between temporal and pitch pattern selectivity, they leave open the possibility that there might be weak biases aligned with this distinction.

Additionally, prior studies have proposed that there are hemispheric asymmetries, such that the right hemisphere is specialized for pitch-based aspects of music and the left hemisphere for rhythmic processing (Albouy et al., 2020; Zatorre & Belin, 2001). A natural prediction is thus that we might have observed stronger pitch pattern scrambling effects in the right hemisphere and stronger temporal scrambling effects in the left hemisphere. In contrast with this prediction, we find that effects of both types of scrambling are indistinguishable between the two hemispheres. One possible explanation is that hemispheric biases may be more strongly related to differences in spectrotemporal tuning (Boemio et al., 2005; Flinker et al., 2019; Obleser et al., 2008; Zatorre & Belin, 2001), that are unlikely to explain the responses observed here, although studies explicitly measuring spectrotemporal tuning and tone selectivity have typically reported strongly bilateral responses (Moerel et al., 2015; Norman-Haignere et al., 2013, 2015; Santoro et al., 2014).

Collectively, our results suggest that the coding of musical pitch and temporal pattern structure is widely distributed throughout the auditory cortex, in that it is present in the left and right hemisphere and in both anterior and posterior auditory cortex. But at the same time, sensitivity to

musical structure is highly focal in that it is concentrated in relatively small clusters of music selectivity that also show evidence for sensitivity to note-level structure. Indeed, we ran an online experiment in which listeners were asked to rate the musicality of each MIDI music clip, and the variation in average rating across stimuli resembled the response pattern of the music component (**Figure S1**), raising the possibility that music-selective neural populations may contribute substantially to the perception of many aspects of musical structure. As a consequence, understanding their computations and representations may significantly advance our understanding of the neural basis of music perception. At the same time, music perception is almost certainly influenced by many brain systems, including those beyond the auditory cortex that may not be specific to music per se. For example, the perception of metrical structure may be influenced by motor regions (Brett & Grahn, 2007; Fedorenko et al., 2012; Janata et al., 2002; Lee et al., 2011; Matthews et al., 2020; Williams et al., 2021), frontal regions that code expectation may modulate activity in the auditory cortex (Koelsch et al., 2019), and limbic regions likely play an important role in linking musical representations with memory and emotional structures.

Like all studies, our conclusions are limited by the resolution of the methods used to measure neural responses. The voxels measured by functional MRI pool activity from hundreds of thousands of neurons and thus inevitably reflect mixed selectivities from distinct underlying neuronal populations. Component methods can to some degree unmix overlapping neural selectivity, but the results are still limited by the coarse, indirect nature of neuroimaging methods. Compared with fMRI, methods with higher spatial and temporal resolution have been able to infer a larger number of reliable response components that reflect finer-grained patterns of selectivity across sounds, such as a component selective for song that was distinct from response components reflecting selectivity for music and speech more generally (Norman-Haignere, Feather, et al., 2022). High temporal resolution might in fact be required to observe dissociations between processing of pitch and temporal structure in music, as suggested by another intracranial EEG study that found differences in the temporal dynamics of expectations associated with note pitches vs. onset times (Di Liberto et al., 2020), which would not be detectable with fMRI. As the precision of human neural recordings continues to advance (Leonard et al., 2024), it is possible that we may be able to resolve finer sub-divisions within the domain of music, thus furthering our understanding of the neural computations and representations that underlie the human proclivity for music.

REFERENCES

Abrams, D. A., Bhatara, A., Ryali, S., Balaban, E., Levitin, D. J., & Menon, V. (2011). Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. *Cerebral Cortex*, *21*(7), 1507–1518. https://doi.org/10.1093/cercor/bhq198

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*(5), 4124–4133. https://doi.org/10.1121/1.3701865

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 16. https://doi.org/10.1126/science.1252826

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F. A., Armony, J. L., & Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. *Cortex*, *59*, 126–137. https://doi.org/10.1016/j.cortex.2014.07.013

Arcurio, L. R., Gold, J. M., & James, T. W. (2012). The response of face-selective cortex with single face parts and part combinations. *Neuropsychologia*, *50*(10), 2454–2459. https://doi.org/10.1016/j.neuropsychologia.2012.06.016

Bendor, D., & Wang, X. (2005). The Neuronal Representation of Pitch in Primate Auditory Cortex. *Nature*, *436*(7054), 1161–1165.

Bianco, R., Zuk, N. J., Bigand, F., Quarta, E., Grasso, S., Arnese, F., Ravignani, A., Battaglia-Mayer, A., & Novembre, G. (2024). Neural encoding of musical expectations in a non-human primate. *Current Biology*, *34*(2), 444-450.e5. https://doi.org/10.1016/j.cub.2023.12.019

Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*, 100–130.

Bizley, J. K., Walker, K. M. M., King, A. J., & Schnupp, J. W. H. (2010). Neural ensemble codes for stimulus periodicity in auditory cortex. *Journal of Neuroscience*, *30*(14), 5078–5091. https://doi.org/10.1523/JNEUROSCI.5475-09.2010

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*, *125*(6), 2237–2263. https://doi.org/10.1152/jn.00588.2020

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, *8*(3), 389–395. https://doi.org/10.1038/nn1409

Brett, M., & Grahn, J. A. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience*, *19*(5), 893–906.

Brochard, R., Abecasis, D., Potter, D., Ragot, R., & Drake, C. (2003). The “Ticktock” of our Internal Clock: Direct Brain Evidence of Subjective Accents in Isochronous Sequences. *Psychological Science*, *14*(4), 362–366. https://doi.org/10.1111/1467-9280.24441

Cuddy, L. L., & Badertscher, B. (1987). *Recovery of the tonal hierarchy : 41*(6).

Da Costa, S., Van Der Zwaag, W., Marques, J. P., Frackowiak, R. S. J., Clarke, S., & Saenz, M. (2011). Human Primary Auditory Cortex Follows the Shape of Heschl’s Gyrus. *The Journal of Neuroscience*, *31*(40), 14067–14075. https://doi.org/10.1523/JNEUROSCI.2000-11.2011

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*(2–3), 109–114. https://doi.org/10.1002/(SICI)1097-0193(1999)8:2/3<109::AID-HBM7>3.0.CO;2-W

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

DeCharms, R. C., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, *280*(5368), 1439–1443. https://doi.org/10.1126/science.280.5368.1439

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234. https://doi.org/10.1152/jn.2001.85.3.1220

Desain, P., & Honing, H. (2003). The formation of rhythmic categories and metric priming. *Perception*, *32*(3), 341–365. https://doi.org/10.1068/p3370

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., de Cheveigné, A., Shamma, S., & Mesgarani, N. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife*, *148*, 26–74. https://doi.org/10.1101/714634

Doeller, C. F., Opitz, B., Mecklinger, A., Krick, C., Reith, W., & Schröger, E. (2003). Prefrontal cortex involvement in preattentive auditory deviance detection: Neuroimaging and electrophysiological evidence. *NeuroImage*, *20*(2), 1270–1282. https://doi.org/10.1016/S1053-8119(03)00389-6

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*(4), 341–354. https://doi.org/10.1037/0033-295X.85.4.341

Downing, P. E., Wiggett, A. J., & Peelen, M. V. (2007). Functional magnetic resonance imaging investigation of overlapping lateral occipitotemporal activations using multi-voxel pattern analysis. *Journal of Neuroscience*, *27*(1), 226–233. https://doi.org/10.1523/JNEUROSCI.3619-06.2007

Eerola, T., & Toiviainen, P. (2004). *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä.

Ellis, D. (2014). *Midi2aiff.scpt*. https://github.com/dpwe/alignmidi/blob/master/midi2aiff.scpt

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., & Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in Neuroscience*, *9*(APR), 1–13. https://doi.org/10.3389/fnins.2015.00157

Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., McDermott, J. H., Norman-Haignere, S. V., & Kanwisher, N. G. (2012). Sensitivity to musical structure in the human brain. *Journal of Neurophysiology*, *108*(12), 3289–3300. https://doi.org/10.1152/jn.00209.2012

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0548-z

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, *28*(12), 1860–1871. https://doi.org/10.1016/j.cub.2018.04.033

Hannon, E. E., & Trainor, L. J. (2007). Music acquisition: Effects of enculturation and formal training on development. *Trends in Cognitive Sciences*, *11*(11), 466–472. https://doi.org/10.1016/j.tics.2007.08.008

Heilbron, M., & Chait, M. (2017). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2017.07.061

Herdener, M., Esposito, F., Scheffler, K., Schneider, P., Logothetis, N. K., Uludag, K., & Kayser, C. (2013). Spatial representations of temporal and spectral sound cues in human auditory cortex. *Cortex*, *49*(10), 2822–2833. https://doi.org/10.1016/j.cortex.2013.04.003

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience*, *36*(6). https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, *50*, 1202–1211. https://doi.org/10.1016/j.neuroimage.2010.01.046

Jacoby, N., & McDermott, J. H. (2017). Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Current Biology*, *27*(3), 359–370. https://doi.org/10.1016/j.cub.2016.12.031

Jacoby, N., Polak, R., Grahn, J. A., Cameron, D. J., Lee, K. M., Godoy, R., Undurraga, E. A., Huanca, T., Thalwitzer, T., Doumbia, N., Goldberg, D., Margulis, E. H., Wong, P. C. M., Jure, L., Rocamora, M., Fujii, S., Savage, P. E., Ajimi, J., Konno, R., … McDermott, J. H. (2024). Commonality and variation in mental representations of music revealed by a cross-cultural comparison of rhythm priors in 15 countries. *Nature Human Behaviour*, *8*(5), 846–877. https://doi.org/10.1038/s41562-023-01800-9

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The cortical topography of tonal structures underlying western music. *Science*, *298*(5601), 2167–2170. https://doi.org/10.1126/science.1076262

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience*, *7*(7 DEC), 1–15. https://doi.org/10.3389/fnins.2013.00247

Kilgard, M. P., & Merzenich, M. M. (2002). Order-sensitive plasticity in adult primary auditory cortex. *Proceedings of the National Academy of Sciences*, *99*(5), 3205–3209. https://doi.org/10.1073/pnas.261705198

Koelsch, S., Fritz, T., Schulze, K., Alsop, D., & Schlaug, G. (2005). Adults and children processing music: An fMRI study. *NeuroImage*, *25*(4), 1068–1076. https://doi.org/10.1016/j.neuroimage.2004.12.050

Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive Processes and the Peculiar Case of Music. *Trends in Cognitive Sciences*, *23*(1), 63–77. https://doi.org/10.1016/j.tics.2018.10.006

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368. https://doi.org/10.1037/0033-295X.89.4.334

Landemard, A., Bimbard, C., Demené, C., Shamma, S., Norman-Haignere, S., & Boubenec, Y. (2021). Distinct higher-order representations of natural sounds in human and ferret auditory cortex. *eLife*, *10*, e65566. https://doi.org/10.7554/eLife.65566

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *The Journal of Neuroscience*, *30*(22), 7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Lee, Y. S., Janata, P., Frost, C., Hanke, M., & Granger, R. (2011). Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. *NeuroImage*, *57*(1), 293–300. https://doi.org/10.1016/j.neuroimage.2011.02.006

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., Mesgarani, N., Welkenhuysen, M., Dutta, B., & Chang, E. F. (2024). Large-scale single-neuron speech

sound encoding across the depth of human cortex. *Nature*, *626*(7999), 593–602. https://doi.org/10.1038/s41586-023-06839-2

Levitin, D. J., & Menon, V. (2003). Musical structure is processed in “language” areas of the brain: A possible role for Brodmann Area 47 in temporal coherence. *NeuroImage*, *20*(4), 2142–2152. https://doi.org/10.1016/j.neuroimage.2003.08.016

Levitin, D. J., & Menon, V. (2005). The Neural Locus of Temporal Structure and Expectancies in Music: Evidence From Functional Neuroimaging At 3 Tesla. *Music Perception*, *22*(3), 563–575.

Lindblom, B., & Sundberg, J. (2007). The Human Voice in Speech and Singing. In *Spring Handbook of Acoustics* (pp. 669–712). Springer Handbooks.

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of Face Parts and Face Configurations: An fMRI Study. *Journal of Cognitive Neuroscience*, *22*(1), 203–211. https://doi.org/10.1162/jocn.2009.21203

Loftus, R., & Masson, J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490.

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca’s area: An MEG study. *Nature Neuroscience*, *4*(5), 540–545. https://doi.org/10.1038/87502

Matthews, T. E., Witek, M. A. G., Lund, T., Vuust, P., & Penhune, V. B. (2020). The sensation of groove engages motor and reward networks. *NeuroImage*, *214*(March), 116768. https://doi.org/10.1016/j.neuroimage.2020.116768

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*(3), 177–192. https://doi.org/10.1007/BF00419633

McPherson, M. J., Dolan, S. E., Durango, A., Ossandon, T., Valdés, J., Undurraga, E. A., Jacoby, N., Godoy, R. A., & McDermott, J. H. (2020). Perceptual fusion of musical notes by native Amazonians suggests universal representations of musical intervals. *Nature Communications*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-16448-6

Mehr, S. (2024). *Principles of music perception*. OSF. https://doi.org/10.31234/osf.io/bdpjc

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Egner, A., Jacoby, N., Hopkins, E. J., Howard, M., Donnell, T. J. O., Pinker, S., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(970).

Merchant, H., Pérez, O., Zarco, W., & Gámez, J. (2013). Interval Tuning in the Primate Medial Premotor Cortex as a General Timing Mechanism. *Journal of Neuroscience*, *33*(21), 9082–9096. https://doi.org/10.1523/JNEUROSCI.5513-12.2013

Miller, L. M., Escabí, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, *87*(1), 516–527. https://doi.org/10.1152/jn.00395.2001

Moerel, M., De Martino, F., Santoro, R., Yacoub, E., & Formisano, E. (2015). Representation of pitch chroma by multi-peak spectral tuning in human auditory cortex. *NeuroImage*, *106*, 161–169. https://doi.org/10.1016/j.neuroimage.2014.11.044

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., Kean, H., Qian, P., & Fedorenko, E. (2020). Composition is the Core Driver of the Language-selective Network. *Neurobiology of Language*, *1*(1), 104–134. https://doi.org/10.1162/nol_a_00005

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology*, *32*(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Norman-Haignere, S. V., Kanwisher, N. G., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, *33*(50), 19451–19469. https://doi.org/10.1523/JNEUROSCI.2880-13.2013

Norman-Haignere, S. V., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, *88*(6), 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*(3), 455–469. https://doi.org/10.1038/s41562-021-01261-y

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biology*, *16*(12). https://doi.org/10.1371/journal.pbio.2005127

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *Journal of Neuroscience*, *28*(32), 8116–8123. https://doi.org/10.1523/JNEUROSCI.1290-08.2008

Ogg, M., Slevc, L. R., & Idsardi, W. J. (2017). The time course of sound category identification: Insights from acoustic features. *The Journal of the Acoustical Society of America*, *142*(6), 3459–3473. https://doi.org/10.1121/1.5014057

Op De Beeck, H. P., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: Maps, modules and dimensions. *Nature Reviews Neuroscience*, *9*(2), 123–135. https://doi.org/10.1038/nrn2314

Oram, N., Cuddy, L. L., & Oram, N. (1995). Responsiveness of Western adults to pitch-distributional information in melodic sequences. *Psychological Research*, *57*(2), 103–118. https://doi.org/10.1007/BF00447080

Overath, T., Mcdermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911. https://doi.org/10.1038/nn.4021

Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J. M., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., Hegde, S., Rocamora, M., Thorne, R., Nweke, F., Sadaphal, D. P., Sadaphal, P. M., Hadavi, S., … Savage, P. E. (2024). Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech: A Registered Report. *Science Advances*, *10*(20). https://doi.org/10.1126/sciadv.adm9797

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Palmer, C., & Krumhansl, C. L. (1990). Mental Representations for Musical Meter. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 728–741. https://doi.org/10.1037/0096-1523.16.4.728

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–776. https://doi.org/10.1016/S0896-6273(02)01060-7

Peretz, I., Vuvan, D., Laine Lagrois, M.-É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Phil. Trans. R. Soc. B*, *370*. https://doi.org/10.1098/rstb.2014.0090

Povel, D. J., & Okkerman, H. (1981). Accents in equitone sequences. *Perception & Psychophysics*, *30*(6), 565–572. https://doi.org/10.3758/BF03202011

Quiroga-Martinez, D. R., Hansen, N. C., Højlund, A., Pearce, M. T., Brattico, E., & Vuust, P. (2019). Reduced prediction error responses in high-as compared to low-uncertainty musical contexts. *Cortex*, *120*, 181–200. https://doi.org/10.1016/j.cortex.2019.06.010

Rohrmeier, M., Rebuschat, P., & Cross, I. (2011). Incidental and online learning of melodic structure. *Consciousness and Cognition*, *20*(2), 214–222. https://doi.org/10.1016/j.concog.2010.07.004

Sankaran, N., Leonard, M. K., Theunissen, F., & Chang, E. F. (2024). Encoding of melody in the human auditory cortex. *Science Advances*, *10*(7). https://doi.org/10.1126/sciadv.adk0010

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, *10*(1). https://doi.org/10.1371/journal.pcbi.1003412

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences*, *106*(34), 14611-. https://doi.org/10.1073/pnas.131591898

Schubotz, R. I., & von Cramon, D. Y. (2002). Predicting Perceptual Events Activates Corresponding Motor Schemes in Lateral Premotor Cortex: An fMRI Study. *NeuroImage*, *15*(4), 787–796. https://doi.org/10.1006/nimg.2001.1043

Smith, N. A., & Schmuckler, M. A. (2004). The Perception of Tonal Structure Through the Differentiation and Organization of Pitches. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 268–286. https://doi.org/10.1037/0096-1523.30.2.268

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science*, *32*(2), 418–444. https://doi.org/10.1080/03640210701864089

Temperley, D. (2014). Probabilistic Models of Melodic Interval. *Music Perception*, *32*(1), 85–99. https://doi.org/10.1525/mp.2014.32.1.85

Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2013). Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. *Cerebral Cortex*, *23*, 249–254. https://doi.org/10.1093/cercor/bhs003

Tierney, A., Patel, A. D., & Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, *147*(6), 888–904. https://doi.org/10.1037/xge0000455

Tillmann, B., Janata, P., & Bharucha, J. J. (2003). Activation of the inferior frontal cortex in musical priming. *Brain Research. Cognitive Brain Research*, *16*(2), 145–161. https://doi.org/10.1016/s0926-6410(02)00245-8

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response Properties of the Human Fusiform Face Area. *Cognitive Neuropsychology*, *17*(1–3), 257–280. https://doi.org/10.1080/026432900380607

Traer, J., Norman-Haignere, S. V., & McDermott, J. H. (2021). Causal inference in environmental sound recognition. *Cognition*, *214*(May), 104627. https://doi.org/10.1016/j.cognition.2021.104627

Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews Neuroscience*. https://doi.org/10.4324/9781315194738-3

Williams, J. A., Margulis, E. H., Nastase, S. A., Chen, J., Hasson, U., Norman, K. A., & Baldassano, C. (2021). High-order areas and auditory cortex both represent the high-level event structure of music. *bioRxiv*, 2021.01.26.428291.

Wilson, B., Slater, H., Kikuchi, Y., Milne, A. E., Marslen-Wilson, W. D., Smith, K., & Petkov, C. I. (2013). Auditory Artificial Grammar Learning in Macaque and Marmoset Monkeys. *Journal of Neuroscience*, *33*(48), 18825–18835. https://doi.org/10.1523/JNEUROSCI.2414-13.2013

Woods, K. J. P., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(14), E3313–E3322. https://doi.org/10.1073/pnas.1801614115

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379. https://doi.org/10.1038/nn1536

Yaron, A., Hershenhoren, I., & Nelken, I. (2012). Sensitivity to Complex Statistical Regularities in Rat Auditory Cortex. *Neuron*, *76*(3), 603–615. https://doi.org/10.1016/j.neuron.2012.08.025

Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavior. *Journal of Neuroscience*, *34*(12), 4396–4408. https://doi.org/10.1523/JNEUROSCI.2799-13.2014

Yin, P., Mishkin, M., Sutter, M., & Fritz, J. B. (2008). Early stages of melody processing: Stimulus-sequence and task-dependent neuronal activity in monkey auditory cortical fields A1 and R. *Journal of Neurophysiology*, *100*(6), 3009–3029. https://doi.org/10.1152/jn.00828.2007

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*, 946–953.

SUPPLEMENTAL MATERIAL

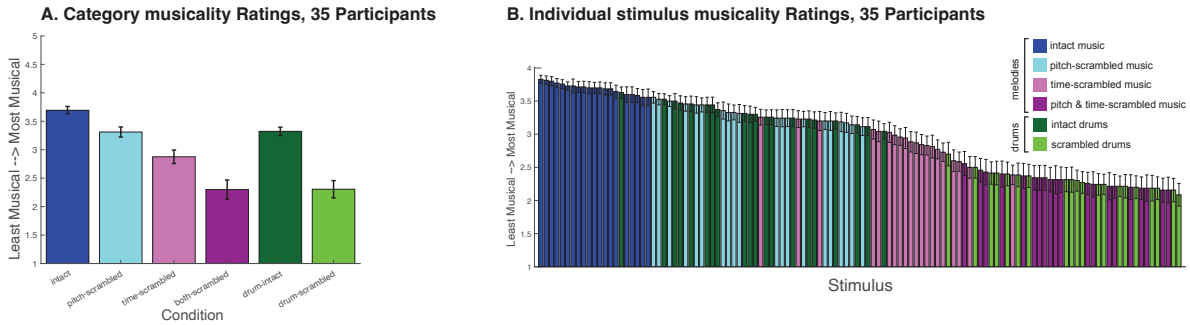

Figure S1. Results of an online behavioral experiment in which participants (N = 35) heard each of the 120 MIDI stimuli from this experiment and were asked "how musical is this clip on a scale from 1 (not at all musical) to 5 (extremely musical)?". **A.** Musicality ratings averaged within condition (error bars are SEM across participants). **B.** Musicality ratings for each individual stimulus (error bars are SEM across participants).