Behavioral/Cognitive

The Music of Silence: Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics

Guilhem Marion,[1] Giovanni M. Di Liberto,[1,2,3] and Shihab A. Shamma[1,4]

[1]Laboratoire des Systèmes Perceptifs, Département d'Étude Cognitive, École Normale Supérieure, PSL, 75005, Paris, France, [2]Trinity Centre for Biomedical Engineering, Trinity College Institute of Neuroscience, Department of Mechanical, Manufacturing and Biomedical Engineering, Trinity College, University of Dublin, D02 PN40, Dublin 2, Ireland, [3]School of Electrical and Electronic Engineering and UCD Centre for Biomedical Engineering, University College Dublin, D04 V1W8, Dublin 4, Ireland, and [4]Institute for Systems Research, Electrical and Computer Engineering, University of Maryland, College Park, MD 20742

Musical imagery is the voluntary internal hearing of music in the mind without the need for physical action or external stimulation. Numerous studies have already revealed brain areas activated during imagery. However, it remains unclear to what extent imagined music responses preserve the detailed temporal dynamics of the acoustic stimulus envelope and, crucially, whether melodic expectations play any role in modulating responses to imagined music, as they prominently do during listening. These modulations are important as they reflect aspects of the human musical experience, such as its acquisition, engagement, and enjoyment. This study explored the nature of these modulations in imagined music based on EEG recordings from 21 professional musicians (6 females and 15 males). Regression analyses were conducted to demonstrate that imagined neural signals can be predicted accurately, similarly to the listening task, and were sufficiently robust to allow for accurate identification of the imagined musical piece from the EEG. In doing so, our results indicate that imagery and listening tasks elicited an overlapping but distinctive topography of neural responses to sound acoustics, which is in line with previous fMRI literature. Melodic expectation, however, evoked very similar frontal spatial activation in both conditions, suggesting that they are supported by the same underlying mechanisms. Finally, neural responses induced by imagery exhibited a specific transformation from the listening condition, which primarily included a relative delay and a polarity inversion of the response. This transformation demonstrates the top-down predictive nature of the expectation mechanisms arising during both listening and imagery.

Key words: **auditory cortex; cortical decoding; melodic expectation; musical imagery; predictive coding**

Significance Statement

It is well known that the human brain is activated during musical imagery: the act of voluntarily hearing music in our mind without external stimulation. It is unclear, however, what the temporal dynamics of this activation are, as well as what musical features are precisely encoded in the neural signals. This study uses an experimental paradigm with high temporal precision to record and analyze the cortical activity during musical imagery. This study reveals that neural signals encode music acoustics and melodic expectations during both listening and imagery. Crucially, it is also found that a simple mapping based on a time-shift and a polarity inversion could robustly describe the relationship between listening and imagery signals.

Received Jan. 25, 2021; revised June 23, 2021; accepted June 28, 2021.

Author contributions: G.M., G.M.D.L., and S.A.S. designed research; G.M. performed research; G.M. and G.M.D.L. contributed unpublished reagents/analytic tools; G.M. analyzed data; G.M. wrote the first draft of the paper; G.M., G.M.D.L., and S.A.S. edited the paper; G.M. wrote the paper.

This work was supported by Advanced ERC Grant NEUME 787836 and Air Force Office of Scientific Research and National Science Foundation grants to S.A.S.; and FrontCog Grant ANR-17-EURE-0017, PSL Idex ANR-10-IDEX-0001-02, and a PhD scholarship from the Research Chair on Beauty Studies PSL–L'Oréal to G.M. We thank the Telluride Neuromorphic Engineering Workshop and Stephen McAdams (McGill University) for facilitating collection of preliminary data; Seung-Goo Kim for helping with analyzing pilot data; and Grégoire Blanc and Valérian Fraisse for participating in the pilot experiments.

The authors declare no competing financial interests.

Correspondence should be addressed to Guilhem Marion at guilhem.marion@ens.fr.

https://doi.org/10.1523/JNEUROSCI.0183-21.2021

Introduction

Musical imagery is the voluntary hearing of music internally without the need for physical action or acoustic stimulation. This ability is important in music creation (Godoy and Jorgensen, 2012), from composition and improvisation to mental practice (Bastepe-Gray et al., 2020). One notable example is Robert Schumann's piano method, in which students are asked to reach the point of "hearing music from the page." But what are the neural underpinnings of such musical imagery?

Previous fMRI studies have found shared areas of cortical activation for imagery and listening tasks, but also nonoverlapping ones (for review, see Zatorre and Halpern, 2005). The

shared activation was measured across several areas of the human cortex (Hubbard, 2013), specifically in the auditory belt areas (Zatorre et al., 1996; Halpern et al., 2004; Kraemer et al., 2005; Herholz et al., 2012), the association cortex (Halpern and Zatorre, 1999; Kraemer et al., 2005), the PFC (Halpern and Zatorre, 1999; Herholz et al., 2012; Lima et al., 2015), and Wernicke's area (Zhang et al., 2017). Musical imagery also seems to engage motor areas (e.g., Halpern and Zatorre, 1999; Halpern, 2001; Herholz et al., 2012; Zhang et al., 2017), showing spatial activation patterns that are correlated with those measured during music production (Meister et al., 2004; Miller et al., 2010). Interestingly, there is only limited evidence for activation during musical imagery in primary auditory cortex (e.g., Griffiths, 1999; Yoo et al., 2001; Halpern et al., 2004; Bunzeck et al., 2005; Bastepe-Gray et al., 2020), although this region is strongly activated during musical listening.

Although these previous studies provided detailed insights into which areas are active during musical imagery, the nature and functional role of such activation remain uncertain. One reason lies in the difficulty of studying the temporal dynamics of the underlying neural responses and processes with the relatively slow fMRI measurements. A recent study using broadly distributed electrocorticography (ECoG) recordings has indicated that music listening and imagery activated shared cortical regions, but with a latency of a reversed sequential order between the auditory and motor areas (Ding et al., 2019). Beyond this, little is known about the nature of cortical signals induced by music imagery, especially with regards to their temporal dynamics and the characteristics it might share with listening responses.

Part of the mystery of musical imagination stems from the fact that music is an elaborate symbolic system conveyed via complex acoustic signals, whose appreciation involves several hierarchical levels of processing. The foundations of such hierarchy depend on the processing of fundamental perceptual attributes, such as pitch, loudness, timber, and space, which are extracted and represented at or before the primary auditory cortex (Koelsch and Siebel, 2005; Janata, 2015). Higher-order rules of grammar and engagement are then presumably implemented in secondary auditory areas and other associative regions (Zatorre and Salimpoor, 2013; Cheung et al., 2019; Di Liberto et al., 2020a). These musical rules are related to how listeners interact and anticipate musical streams, in what is usually referred to as melodic expectations. Experimentally, such expectations are assumed to play a critical role in musical listening in relation to auditory memory (Agres et al., 2018) and musical pleasure (Zatorre and Salimpoor, 2013; Gold et al., 2019), and to interact with the reward system (Blood and Zatorre, 2001; Salimpoor et al., 2011; Cheung et al., 2019). However, it is unknown whether these melodic expectations play any role during musical imagery, where they could be related to the ability to recall, create, and become emotionally engaged with the music generated within our own mind.

Melodic expectations can be quantified using statistical models trained on a musical corpus that summarizes the musical material to which listeners have been exposed (Pearce, 2005; Abdallah and Plumbley, 2009; Gillick et al., 2010; Rohrmeier, 2011), thus capturing listeners' perceptual judgments, musical reactions, and expectations (Krumhansl et al., 1999, 2000; Pearce, 2018). In our experiments, the musical corpus was a large repertoire of Western music with which our participants were familiar. Using these models of melodic structure, our experimental results suggest that imagery of naturalistic melodies (Bach chorals) elicits cortical responses to the imagined notes, exhibiting temporal dynamics and expectation modulations that are comparable to the neural responses recorded during music listening. We also find that the neural signal recorded in the imagery condition could be used to robustly identify the imagined melody with a single-trial classifier. A companion study (Di Liberto et al., 2021) expands on these results to demonstrate that the ubiquitous short pauses and silent intervals in ongoing music elicit responses and melodic expectations remarkably similar to those seen during imagery. Furthermore, with the absence of simultaneous stimulus-driven (bottom-up) responses during silence, these two studies are able to attain direct evidence of the top-down predictive signals and processes critically involved in building musical expectations and culture.

Materials and Methods

Participants and data acquisition

Twenty-one professional musicians or in training to become professional musicians (6 female; age: mean = 25 years, SD = 5 years) participated in the EEG experiment. The sample size was consistent with a previous related study from our team (Di Liberto et al., 2020a). Each participant reported no history of hearing impairment or neurologic disorder, provided written informed consent, and was paid for their participation. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the CERES committee of Paris Descartes University (CERES 2013-11). The experiment was conducted in a single session for each participant. EEG data were recorded from 64 electrode positions, digitized at 2048 Hz using a BioSemi Active Two system as well as three extra electrodes placed on participants skin to record the activity of muscles of potential cofound (tongue, masseter, forearm fingers extensor). Audio stimuli were presented at a sampling rate of 44,100 Hz using a Genelec 8010 10w speaker and Python code for the presentation. Testing was conducted at École Normale Supérieure, in a dark soundproof room. Participants were asked to read the music scores fixed at the center of the desk during both imagery and listening conditions; however, they were instructed to minimize motor activities during the whole experiment. An SM58 microphone was placed in the booth to record participant sounds and make sure that they were not singing, tapping, or producing sounds during the experiment. The experimenter listened to those sounds online. Before the experiment, all participants took the Advanced Measures of Music Audiation (AMMA) online using the official website (www.giamusicassessment.com).

A tactile metronome (Peterson Body Beat Vibe Clip) playing 100 bpm bars (each 2.4 s) was placed on the left ankle of the participants to provide them with a sensory cue to synchronize their imagination. The start of each trial (listening and imagery) was signaled by a short vibration on the vibro-tactile metronome device followed by a 4 beat countdown. Notes closer than 500 ms from a metronome vibration were excluded to avoid potential contamination from the tactile stimulus. Experimental assessment showed that the metronome precision was within 5 ms; it thus did not impact our experiment. A constant lag was determined experimentally during the pilot experiments to compensate for perceptual auditory-tactile delays; the latency of 35 ms was determined and applied on all participants.

EEG experimental protocol

All participants were chosen to be very well-trained musicians and were all professionals or students at Conservatoire National Supérieur de Musique in Paris. They were given the musical score of the four stimuli in a 1 page score and could practice on the piano for ~35 min. The experimenter checked their practice and verified that there were no mistakes in the execution. After practice, participants were asked to sing the four pieces in the booth with the tactile metronome; the sound was recorded to check their accuracy offline.

The experiment consisted of a single session with 88 trials. For each condition (listening and imagery), each of the four melodies was repeated 11 times. Trial order was shuffled both in terms of musical pieces and conditions. In the listening condition, participants were asked

to passively listen to the stimuli while reading the musical score. For the imagery condition, they were asked to imagine the melody in sync with the tactile metronome as precisely as they could. At the end of every four trials, a break was possible; participants were able to wait as long as they wanted before they continued with the experiment. A sheet of paper was available in the experimental booth, where participants were instructed to report trials where their imagination did not end with the metronome vibration, and therefore were performing the imagery task with incorrect synchronization. No participants reported any mistakes in that sense.

Synchronizing participants' imagination with stimuli is a challenging problem. Previous studies used the so-called "filling in" paradigm where participants are asked to fill an artificial blank introduced in the musical pieces using imagery (Kraemer et al., 2005; Cervantes Constantino and Simon, 2017; Ding et al., 2019), which was not optimal for our experiment as it does not allow for imagery of long stimuli. Other studies displayed visual cues in karaoke-like fashion (Herholz et al., 2012) or used dynamic pianoroll visuals of the stimuli (Zhang et al., 2017). However, several studies have shown that, given the task of synchronizing movements with a discretely timed metronome (e.g., tapping a finger), humans have a striking advantage with auditory metronomes over visual ones (Jäncke et al., 2000; Repp and Penel, 2004; Repp, 2005). In addition, a recent study showed that such an advantage is conserved with tactile metronomes (Ammirante et al., 2016). We assumed that a tactile metronome was less likely to contaminate imagery responses than an auditory metronome because of the different sensory modality. Therefore, we decided to use a tactile metronome even if some studies suggest that it can induce auditory responses (Ammirante et al., 2016).

Stimuli

Four melodies from the corpus of Bach chorals were selected for this study (BWV 349, BWV 291, BWV 354, BWV 271). All chorals use similar compositional principles: the composer takes a well-known melody from a Lutheran hymn (cantus firmus) and harmonizes three lower parts (alto, tenor, and bass) accompanying the initial melody on soprano; these cantus firmi were usually written during the Renaissance era. Our analysis only uses monophonic melodies; we therefore only use these cantus firmi as stimuli for our experiment, and original keys were kept. The chosen melodies follow the same grammatical structures and show very similar melodic and rhythmic patterns. Participants were asked to listen to and imagine these stimuli at 100 bpm (~30 s each). The audio versions were synthesized using Fender Rhodes simulation software (Neo-Soul Keys). The onset times and pitch values of the notes were extracted from the midi files that were precisely aligned with the audio versions presented during the experiment (see Fig. 1).

Tools

Information dynamics of music (IDyOM). IDyOM is a statistical model of musical expectation based on variable-order Markov chains (Pearce, 2005). This model allows for the quantitative estimation of the expectedness of a musical note, which have been shown to be physiologically valid by a number of studies (Omigie et al., 2012, 2019; Egermann et al., 2013; Song et al., 2016; Agres et al., 2018; Di Liberto et al., 2020a). First, the model has been shown to correctly identify melodic expectation patterns in a consistent way with a musicological analysis (Meyer, 1973) of Schubert's *Octet for strings and winds* made by Leonard Meyer in 1973 (Pearce, 2018). The model also showed correlated expectation values with ones estimated from a behavioral experiment (Manzara et al., 1992). IDyOM was able to account for ~63% of the variance in the mean uncertainty estimates reported by the original authors (Pearce, 2005). Finally, a recent study (Di Liberto et al., 2020a) showed that amplitude modulations in EEG and ECoG responses to monophonic music are correlated with the expectation values computed with IDyOM.

The IDyOM model is composed of two modules: a long-term model, which is pretrained on a musical corpus (which did not include the stimuli presented in this experiment) to capture style-specific global patterns; and a short-term model, which is trained on the preceding proximal context in the current piece to estimate expectedness based on local melodic sequences. Both modules use the same underlying method: Markov chains of different orders (n-grams as states) that can describe melodic patterns at various time scales. All the Markov chains are then aggregated into one model by merging all the probability distributions (Pearce, 2005). In our analysis, we use the IDyOMpy model (https://github.com/GuiMarion/IDyOM), which is an implementation of IDyOM where the Markov chains are combined through a weighting based on the entropy of the distributions from each order. The model was trained using note duration as well as note pitch. The joint distribution was then used to compute the unexpectedness (surprise) of events, which was quantified by means of the Information Content value as follows:

$$IC(x) = -log(P(X_t = x))$$

Multivariate temporal response function (mTRF). We used the mTRF toolbox (downloadable at: https://github.com/mickcrosse/mTRF-Toolbox) (Crosse et al., 2016) to estimate the TRFs describing the linear mapping of melodic features (onsets, expectation) into the EEG signal. This mapping was estimated for individual electrodes and was based on a convolutional kernel w, including various time latencies between the music and the EEG signal as follows:

$$\forall t, r(t,k) = (s * w_k)(t) + \varepsilon(t,k)$$

with t the time indices, k the electrodes, and ε the residual response (unexplained noise).

The optimization problem is to find the vector w that minimizes this residual response ε using ordinary least squares method over the vector w while considering a certain degree of regularization to prevent overfitting by assuming a level of temporal smoothness (ridge regularization). The optimal regularization parameter was identified at the individual subject level with an exhaustive search within the interval $[10^{-6}, 10]$ with a logarithmic step. The time-lag window $[-300, 900$ ms] was used to fit the TRF models. The main figures report weights for the reduced window [−100, 500 ms], where the responses and effects of interest were hypothesized to emerge. This framework has been shown to be effective in assessing the EEG encoding of both low-level auditory features and higher-order auditory expectations (Lalor and Foxe, 2010; Di Liberto et al., 2015; O'Sullivan et al., 2015; Broderick et al., 2018; Daube et al., 2019).

Data preprocessing

EEG data were analyzed offline using MATLAB software. Signals were digitally filtered using Butterworth zero-phase filters (low- and high-pass filters of both order 3 and implemented with the function filtfilt) and downsampled to 64 Hz. The main analysis was conducted on data filtered between 0.1 and 30 Hz. Results were also reproduced with the high-pass cutoff frequencies 0.01 and 1 Hz (see Fig. 5). Data were then rereferenced to the average of all 64 channels. EEG channels with a variance exceeding 3 times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation.

Data analysis

Previous studies showed that EEG responses to continuous melodies encode both the acoustic envelope (Di Liberto et al., 2020b) and melodic expectations (Omigie et al., 2013; Di Liberto et al., 2020a). The main aim of our study was to investigate whether that encoding is conserved during musical imagery. To this end, we assessed the encoding of these features in the EEG signals by means of TRF forward modeling predictions.

The EEG signal was grouped in 88 trials (44 per condition). Each trial was associated with stimulus vectors representing acoustic onsets and melodic expectation:

Onsets vector. One-dimensional vector where the note onsets were marked by an impulse with value 1. All other time point were assigned to zero.

Expectation vector. One-dimensional vector where the note onsets were marked by an impulse with value corresponding to the expectation value assigned to that note by IDyOM.

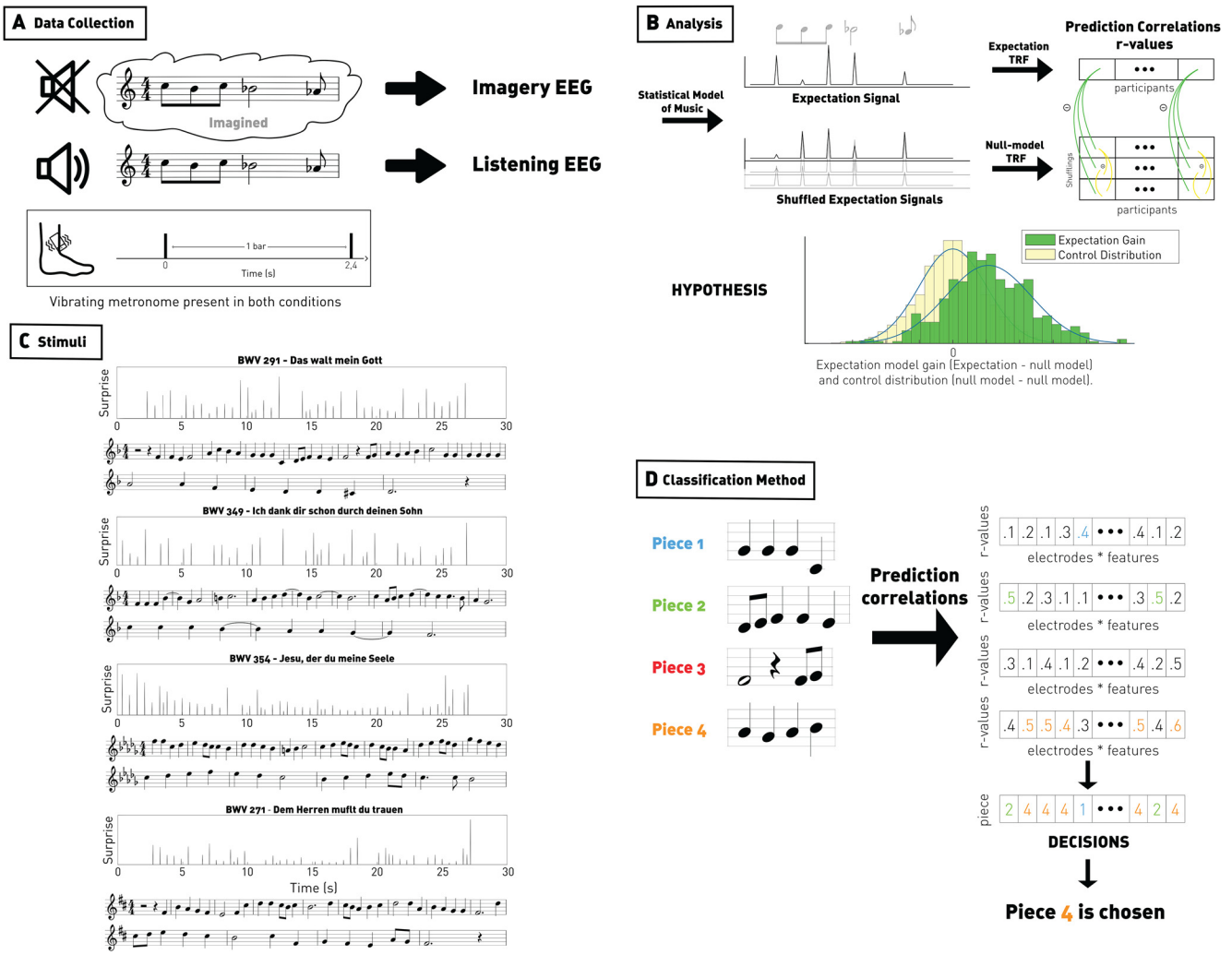

Figure 1. Method figure. ***A***, EEG signal was recorded from participants who listened to and imagined four monophonic Bach melodies. The musical bars were indicated using a vibrotactile metronome. ***B***, Top left panels, Onset vector amplitude modulated according to a statistical model of musical expectations. Null model distributions were derived by shuffling the expectation values while preserving the note onsets. Top right, Forward TRFs were estimated between the melody vectors and the EEG signal. EEG prediction correlations were derived based on the stimulus vectors and subtracted by the ones for the shuffled vectors, providing (Expectation gain; green), reflecting the EEG encoding of melodic expectations. A control distribution was derived by subtracting EEG prediction correlations between pairs of shuffled vectors (yellow). Bottom, We hypothesized a positive shift in expectation gain (green distribution) relative to the control distribution (yellow distribution). ***C***, Stimuli. Musical scores and expectation vectors for each of the four Bach choral stimuli. Melodies were presented at 100 bpm (~30 s each). The expectation signal was computed for each of the melodies using IDyOM. The information content value of each note (the negative log likelihood) was used to modulate the note onset values. Forward TRF models were then fit between the resulting vectors and the EEG signal. ***D***, Classification method. We trained a TRF model with leave-one-out cross-validation and used this model to predict, from the four candidate pieces, the target EEG. We therefore have *nb_electrodes* × *nb_features* prediction correlations. For each of these estimators, we assess which piece maximizes the correlation, and the final decision is the piece that occurs the most across electrodes and features.

Onsets and expectation analyses

Forward TRFs were fit and used to predict independently each channel of the EEG signal from the onsets and the expectation signal using leave-one-trial-out cross-validation. The correlation between the EEG signals and its prediction was computed for each channel separately, resulting in scalp topographies used to assess the spatial activation. This signal (correlation of the feature of the signal of interest with each electrode) accounts for where the signal is computed and not where the amplitude is the strongest, as opposed to ERP topographic maps. Significance of the EEG prediction correlations was assessed by comparing the results with the ones for a null model where parameters of interest were shuffled in our stimuli:

Onsets analysis. We shuffled the order of the trials, ensuring that the resulting shuffling does not produce matching stimulus-EEG pairs.

Expectation analysis. We shuffled the expectation values while preserving the onset times. This produced vectors with correct onset information but meaningless expectation values.

We ran 20 permutations for each analysis. Those distributions were used to assess significance both at the individual-subject and group levels. The group level significance was computed from the correlation gain distribution with respect to the null model (expectation models – null models or onset models – null models). We subtracted the null model prediction correlations to the expectation/onsets model prediction correlations by keeping the participant order. Therefore, we got a distribution of 420 values (21 participants × 20 shuffling = 420). A control distribution was constructed by computing the difference between the null model and other repetitions of itself (here 21 participants × 20 shuffling × 19 different shuffling = 7980). This distribution accounts for the variance of the prediction correlation with a mean of 0. We tested whether the correlation gain was above the control distribution using a Wilcoxon sum rank test. Effect sizes were computed using the common language effect size between the expectation/onset distribution gain and the control distribution. The common language effect size was computed from the *U* statistic computed by the Wilcoxon sum rank test. The common language effect size is defined as follows:

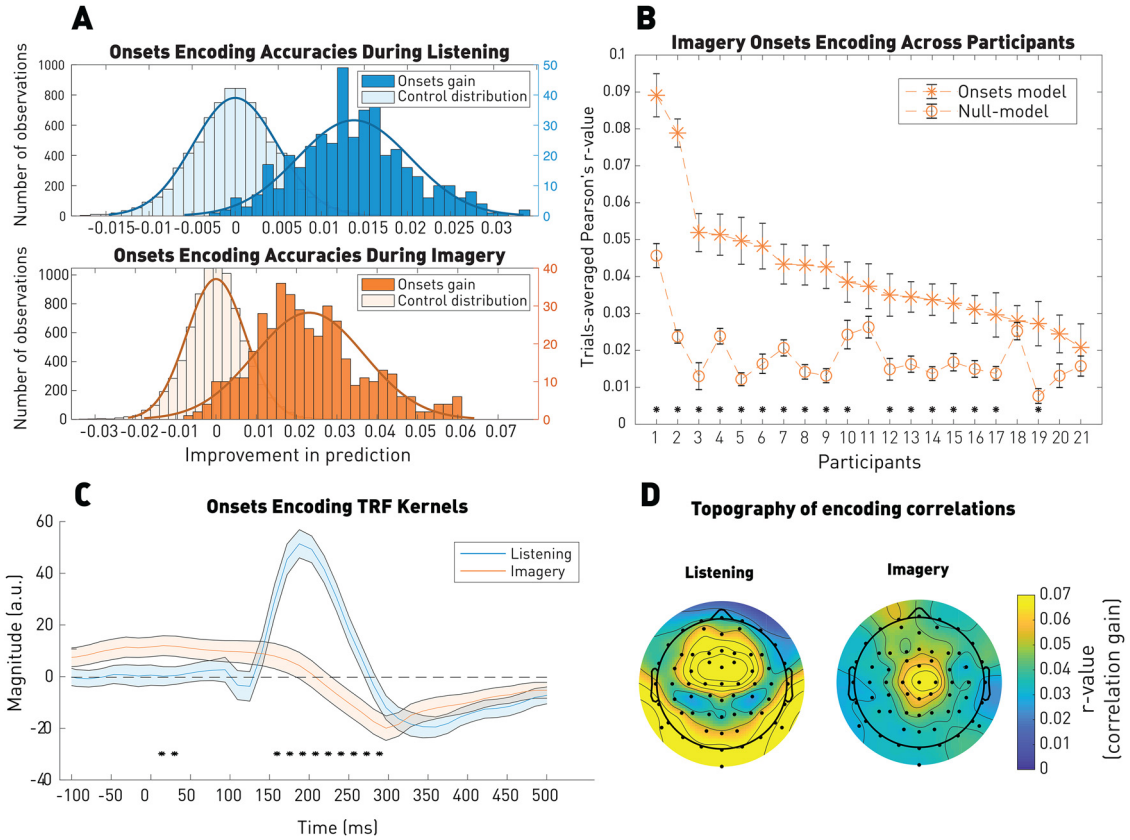

Figure 2. Robust EEG encoding of note onsets during imagery. ***A***, EEG prediction correlations for the listening (top) and imagery (bottom). EEG prediction correlations were significantly above the control distribution in both conditions. Distributions illustrate the note onsets correlation gain, adjusted relative to the null model, as well as the control distribution. As for all the next figures, the left *y* axis corresponds to the number of observations of the control distribution and the right *y* axis corresponds to the ones of the model of interest (here onsets gain). ***B***, EEG prediction correlations for the imagery condition for individual participants. Error bars indicate the SE across the 44 trials. $*p < 0.05$. ***C***, TRF kernels on Cz. Shaded areas represent the SE across participants ($N = 21$). Significance between the two kernels computed by a permutation test: $*p < 0.05$. ***D***, Topography of the EEG predictions gain (onset model − null model). A significant ($p < 0.05$) correlation of $r = 0.3$ was measured between the topographies of the EEG prediction values for the two conditions (Pearson's correlation).

$$f = \frac{U}{n_1 \cdot n_2}$$

with n_1 and n_2, respectively, the sizes of the two distributions (expectation gain and control distribution). As U indicates the number of pairs chosen in the two distributions that satisfy the hypothesis $((i,j)|D1_i > D2_j)$, the common language effect size f therefore indicates the normalized number of pairs that satisfy the hypothesis ($100 \times f$ % of the pairs satisfy the hypothesis).

Cross-conditions analysis

We assessed the consistency between imagery and listening responses by means of a cross-condition TRF approach. Specifically, TRF models were trained on one condition (e.g., listening) and evaluated on the other (e.g., imagery). The resulting EEG prediction correlations were examined to determine whether the two conditions elicited consistent EEG signals. Furthermore, we investigated whether simple transformations (polarity and latency shift) could explain possible differences between the two conditions. First, we applied a simple polarity inversion by multiplying the TRF kernels by −1. Second, we estimated a linear convolution mapping between the averaged listening responses and the averaged imagery signals (and vice versa) for $n - 1$ participants. The learned mapping was then used to transform the listening response into imagery signal (and vice versa) in the left-out participant. The mTRF method was then used to fit subject-specific models on that left-out subject and to predict EEG signals based on the music onsets vectors. The resulting EEG prediction correlations indicate whether the cross-condition mapping is consistent across participants.

Short-term and long-term models

An additional analysis was conducted to assess the relative contribution of the short- and long-term modules of IDyOM to the EEG encoding of melodic expectations. To do so, melodic expectation vectors were derived using the short- and long-term models separately. First, short-term model expectations were used to fit TRF models and predict the EEG. Then we used multivariate regression to predict the EEG when considering the two expectation vectors simultaneously (short-term and long-term). In this multivariate case, the null model was derived by shuffling the values of the long-term expectation vector only. As such, this approach could assess whether the long-term model expectations explain EEG variance that is not captured by the short-term expectations.

Decoding the identity of imagined songs

Classification was performed to decode the identity of a song from a single EEG trial. We devised a classification method using vote-boosting based on the prediction correlations computed from a forward TRF model trained on the left-out trials. Specifically, prediction correlations were calculated for each of the four pieces using, separately, the onsets and the expectation vectors. This procedure produced 128 EEG prediction signals (64 electrodes × 2 features = 128) for each piece. We then computed the correlation between the target EEG data and each predicted EEG signal estimators, leading to 128 correlation values for each of the four pieces. For each estimator, the piece with the highest correlation was chosen, providing one vote for that particular choice. The piece with most votes when considering all estimators was selected as the result of the classification. The methodology is illustrated in Figure 1.

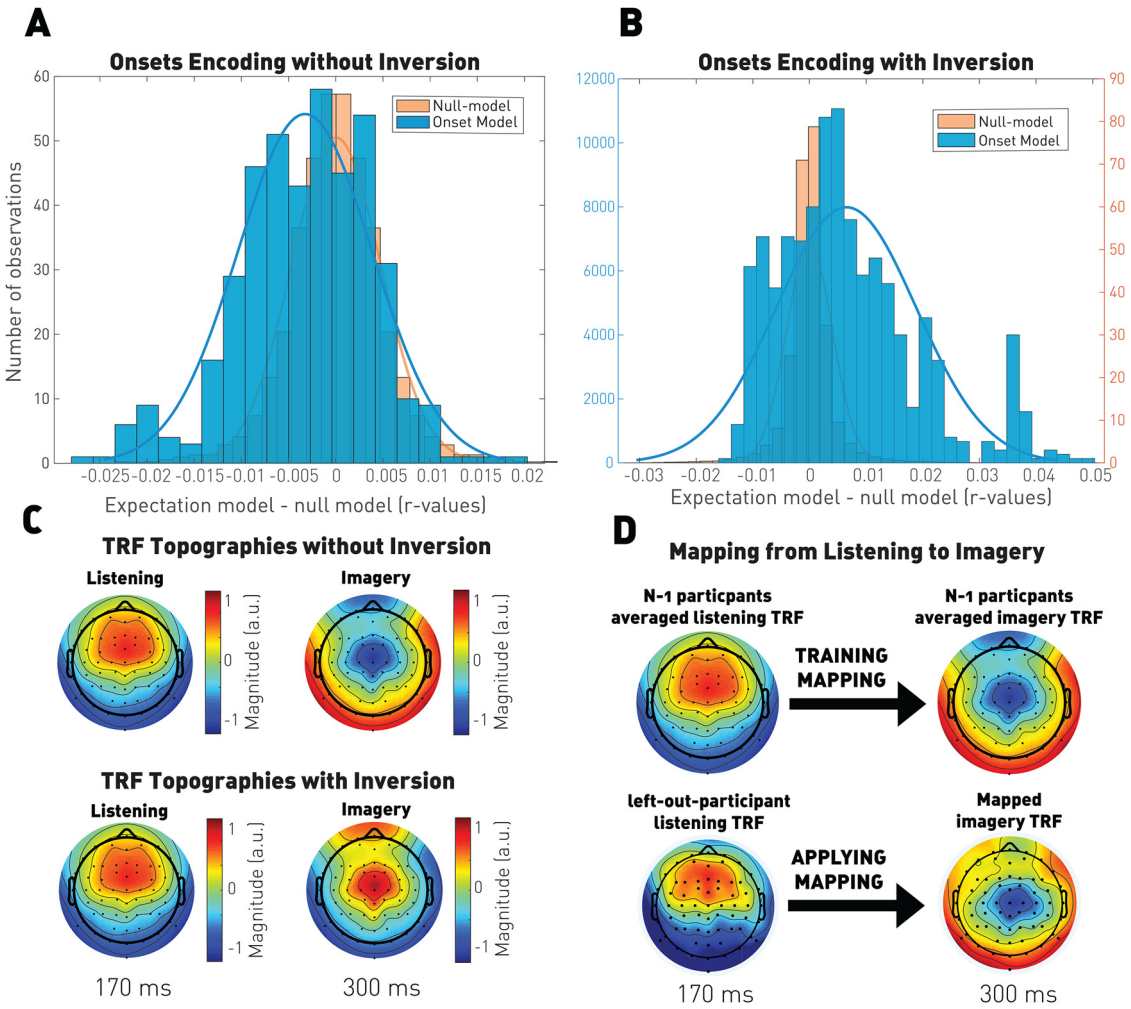

Figure 3. Cross-conditions analysis. TRF models fit on one condition were evaluated on the other one to determine the consistency between conditions. ***A***, Distribution of the difference between the onsets model and the null model prediction of the listening condition based on raw TRF kernels trained on the imagery condition. Significance was computed using a Wilcoxon rank sum test to assess that the distributions are above the control distribution. ***B***, Distribution of the difference between the onsets model and the null model prediction of the listening condition based on inverted TRF kernels trained on the imagery condition. Significance was computed using a Wilcoxon rank sum test to assess that the distributions are above the control distribution ($p = 10^{-46}$). ***C***, TRF kernel topographies. The TRF kernels are normalized and extracted at the time where their global field power was maximum to extract the latency where their responses were the most salient (170 ms for listening and 300 ms for imagery). We can observe a time-shifted inverted polarity of the responses that have been assessed in ***B***. We measured a significant ($p = 10^{-23}$) correlation of $r = 0.9$ between the listening and the imagery-inverted topographic maps. ***D***, A linear convolution mapping between the listening and imagery responses was learned, applied to individual listening responses, and resulted in significant predictions of the imagery EEG using the onsets ($p = 10^{-49}$).

Results

We recorded EEG signals (64-channel recording system) from 21 professional musicians as they imagined and listened to four monophonic Bach chorals (Fig. 1). In both conditions, participants wore a vibrotactile metronome on their left ankle, which allowed for precise synchronization during the imagery task (see Materials and Methods). We first investigated the responses to the notes by regressing the EEG signals with a stimulus vector representing the note onsets at least 500 ms away from the metronome beats. Then, the melodic expectation for each note was estimated using a statistical model of musical structure (IDyOM) (Pearce, 2005) trained on a large corpus of Western melodies, supposed to mimic the musical culture of the listeners participating in this study (Pearce, 2018). We constructed the expectation signal as a sparse vector where time onsets of notes were modulated by the expectation value computed by the statistical model of music. As cortical EEG recordings during music listening have been shown previously to encode this expectation signal (Di Liberto et al., 2020a), our analysis aimed to test the same hypothesis on the imagery condition and to compare the temporal activation between both conditions. The music stimuli, EEG data, and analysis codes are fully available on request to the corresponding author. Main data are accessible through Dryad (https://doi.org/10.5061/dryad.dbrv15f0j).

Onsets encoding

TRFs describing the linear transformation of note onsets to an EEG signal (0.1-30 Hz) were estimated for both conditions using lagged linear regression (mTRF-Toolbox) (Crosse et al., 2016). EEG prediction correlations were derived on left-out portions of the data with cross-validation. The procedure was then repeated after the labels referring to the stimulus order were randomly shuffled (null model; EEG_i was regressed with $stim_j$).

Figure 2 shows that the note onset vector could predict the EEG signal better than chance in both conditions, demonstrating the robust encoding of note onsets in the low-frequency EEG signal (Wilcoxon rank sum test between onsets gain and control distributions; listening: $p = 8.4 \times 10^{-220}$ common language effect

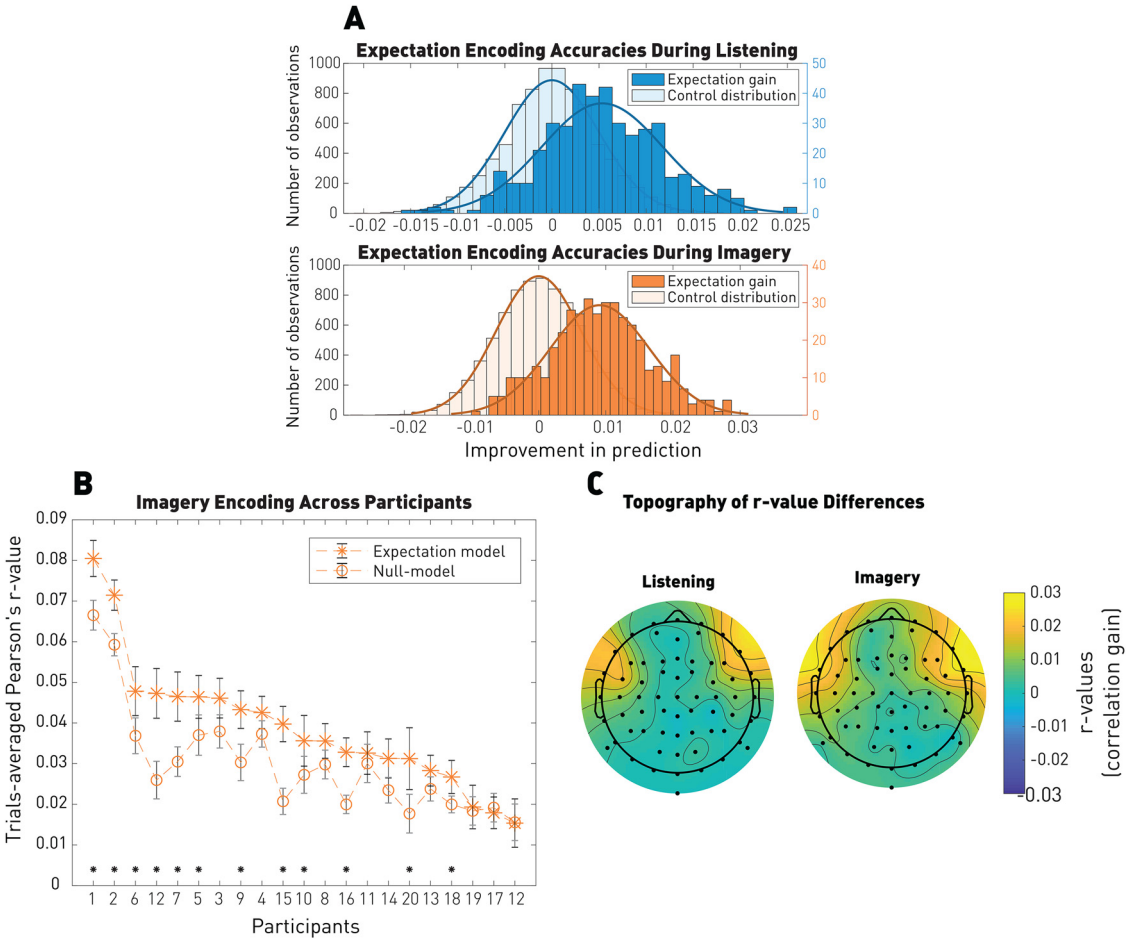

Figure 4. Robust EEG encoding of the expectation signal. ***A***, EEG prediction correlations for the listening and imagery conditions using the expectation TRFs. EEG prediction correlations were significantly above chance in both conditions. ***B***, EEG prediction correlations at the individual participant level for the imagery condition. Error bars indicate the SE across trials. $*p < 0.05$. ***C***, Topographies of the EEG predictions gain (expectation model − null model). Pearson's correlation between conditions: $r = 0.9$.

size $f = 0.98$; imagery: $p = 2.7 \times 10^{-209}$ common language effect size $f = 0.97$; see Materials and Methods). The note onset encoding was significant at the individual participant level (17 of 21, $p < 0.05$, FDR-corrected p values extracted from the null model distributions) and was most accurately encoded on central scalp areas, as previously shown in response to auditory experiments (Di Liberto et al., 2020a,b; Van Canneyt et al., 2020). A significant ($p = 0.02$) correlation of $r = 0.3$ was measured between the topographies of the EEG prediction values for the two conditions (Pearson's correlation).

Cross-condition analysis

In line with previous fMRI studies showing partly overlapping neural activation for auditory listening and imagery, we anticipated that a certain degree of similarity exists between the TRFs measured for the two tasks. Indeed, the TRF weights in Figure 2*C* provided us with a qualitative indication of whether the cortical dynamics for listening and imagery are different. Nevertheless, further quantitative assessment was conducted to determine the precise nature of the similarities between the two conditions and the consistency of such similarities across participants. One dominant difference between the two conditions is a time-shifted inverted polarity of the TRF dynamics. This effect of condition was quantitatively assessed by the cross-condition TRF analysis that follows (Fig. 3).

First, we used the imagery TRF kernels to predict the listening EEG signal, and vice versa, the listening TRF kernels to predict the imagery EEG signal. As expected, these analyses did not produce EEG predictions that were significantly larger than the null distribution (listening → imagery: $p = 0.83$; imagery → listening: $p = 10^{-19}$, with null model > onsets-model), confirming that listening and imagery responses are different. Next, we predicted listening EEG responses from the imagery TRF kernels after a polarity inversion, leading to significant EEG predictions ($p = 10^{-46}$; Fig. 3), indicating that listening and imagery signals are inversely correlated. However, inverting the listening EEG responses did not lead to an adequate prediction of the EEG in the imagery condition ($p = 0.14$). Such an asymmetry in cross-conditions predictions most likely stems from the large difference in the amplitude (and hence the SNR) between the two types of signals. Furthermore, it is also evident from Figure 3 that using only a simple polarity inversion is likely to be a suboptimal description of the mapping between the TRFs in the two conditions. Therefore, we implemented a further refinement in characterizing the relationship between the two TRFs, which included a linear mapping with a convolutional kernel as we describe next. In principle, the identification of such a reliable mapping would usher new ways to decode imagined melodies without the need for training imagery EEG data.

A linear mapping with a convolutional kernel was computed between the averaged listening responses and the averaged imagery responses for $n - 1$ participants. We then applied the learned cross-condition mapping to estimate the imagery EEG

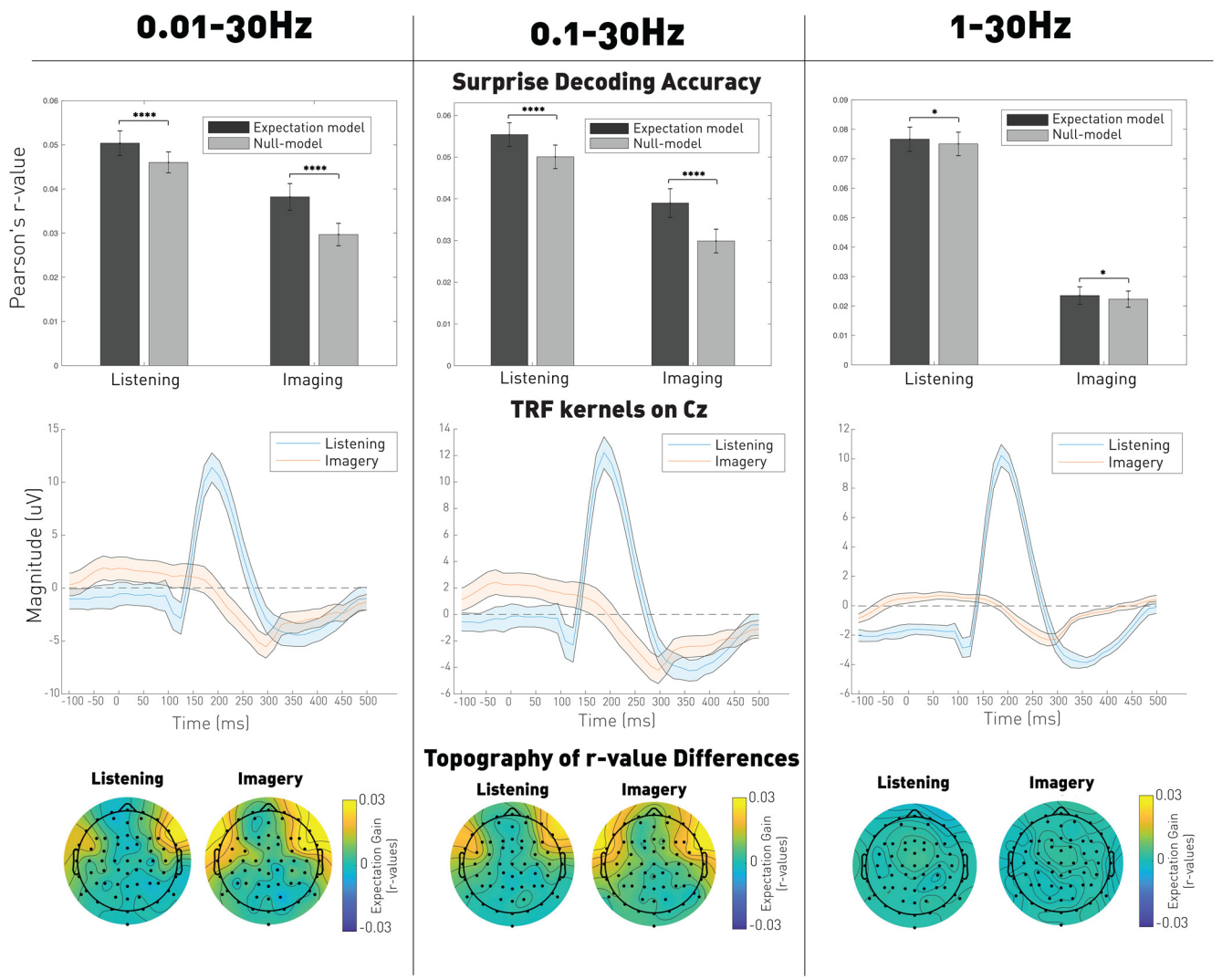

Figure 5. EEG encoding of the expectation signal by frequency bands (0.01-30, 0.1-30, and 1-30 Hz). Top, Averaged prediction correlations for both the expectation model and null models. Significance was computed using a Wilcoxon signed rank test paired by participants and averaged by trials and shuffling: ***$p < 0.001$; *$p < 0.05$. Middle, TRF kernels reflecting the average neural response on Cz. Shaded error bars indicate the SE across participants. Bottom, Topography of the prediction correlations gain (expectation model – null model) over the electrodes.

signal of the left-out participant based on their listening responses and the note onset vectors. This approach led to significant predictions ($p = 10^{-49}$) of the imagery EEG, confirming a reliable relationship between the listening and imagery responses (Fig. 3). However, the EEG prediction correlations derived with this methodology were not larger than the ones from the cross-participants analysis ($p = 0.12$), where we directly used the averaged TRF kernels from $n - 1$ participants on the left-out participant (see Fig. 9). Using more complex nonlinear transformations between the two TRF kernels may lead to better performances and then allow the computation of the imagery TRF kernels directly from the listening ones without having to measure imagery responses.

Encoding of melodic expectations

TRF models were computed to relate melodic expectations to the EEG signal. Expectation vectors were determined by modulating note onset vectors according to the expectation values derived with the statistical model of melodic structure IDyOM (Pearce, 2005). Null models were computed by shuffling the expectation values in the stimulus vectors while preserving the note onset information. A null distribution of EEG prediction correlations was then computed by running the TRF analysis on 20 shuffled versions of the expectation vectors per participant. The correlation gains achieved by using the expectation model (expectation – null model) were compared with the control distribution of "gains" determined based on the null models ($nullmodel_i - nullmodel_j$; see Fig. 1).

Figure 4 shows that EEG prediction correlations were larger for the expectation signal than the null model in both the listening and imagery conditions (Wilcoxon rank sum test; listening: $p = 4.2 \times 10^{-66}$, common language effect size $f = 0.77$; imagery: $p = 3.4 \times 10^{-111}$, common language effect size $f = 0.85$), with significance at the individual level for 12 of 21 participants ($p < 0.05$, FDR-corrected p values extracted from the null model distributions). We did not expect to observe within-subjects significance for all participants as each model was trained on one condition and therefore half of the data.

The shapes of the TRF kernels shown in Figure 5 were qualitatively similar to those depicted in Figure 2 when regressing the onset signal. Interestingly, the effect of expectations (correlation gain) emerged on EEG channels that had little or no sensitivity to the unmodulated onsets, thus possibly reflecting different

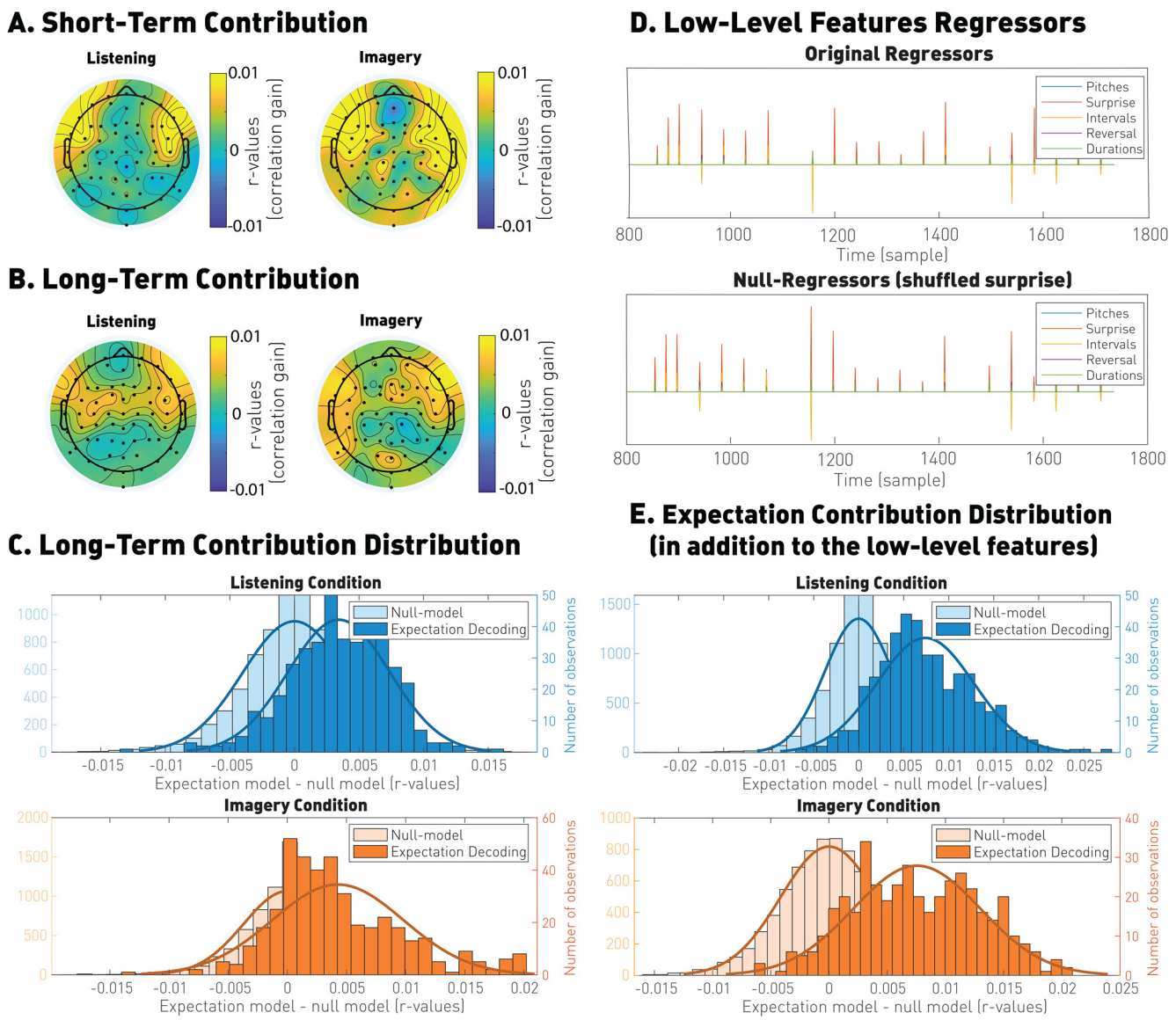

Figure 6. Comparison of the short- and long-term expectation and low-level features. ***A***, Unique correlation contribution for short-term expectations. These values were calculated as the EEG prediction correlations with TRF models based on both long- and short-term expectations, minus the EEG correlations after shuffling the short-term expectation values. ***B***, Unique correlation contribution for long-term expectations. Correlation contribution of the long-term expectation model minus the EEG prediction correlations after shuffling the long-term expectation values. ***C***, Unique correlation contribution of the long-term model, showing that long-term expectations explain EEG variance that is not captured by long-term expectations. ***D***, TRF models were fit by combining low-level features (pitch, duration from the previous note, interval, reversal in pitch direction) were combined with the expectation vector. The null model was derived by combining the same low-level features with a scrambled expectation vector. ***E***, The result of the TRF analysis shows that the expectation signal explains EEG variance that was not captured by the low-level features.

cortical generators for the EEG encoding of acoustics and expectations. Indeed, the expectation gain emerged primarily in frontal scalp areas, which were previously linked with auditory expectations (Opitz et al., 2002; Tillmann et al., 2003; Schönwiesner et al., 2007). Furthermore, the effect of expectation (correlation gain) had similar topographical distributions for the listening and imagery conditions (Pearson's correlation: $r = 0.9$). This also suggests that the expectation signal is the same in both cases and originates from the same source. Figure 5 indicates that low frequencies (<1 Hz) are important for expectation responses. However, even the analysis of the 1-30 Hz band displays significant encoding of expectation. Finally, the topographic distributions are similar for each frequency bands, although somewhat weaker for 1-30 Hz.

Using the methodology above, we measured and compared the impact of the IDyOM short-term model, which relies on music patterns within a piece only, and long-term model, which relies on music statistics derived from a large corpus of music not including the present piece (see Tools). First, we found that short-term expectations contribute significantly to the prediction of the EEG signals (listening: $p = 1.1 \times 10^{-107}$, common language effect size: $f = 0.84$; imagery: $p = 6.9 \times 10^{-134}$, common language effect size: $f = 0.88$), indicating that neural signals encode statistics based on the proximal melodic context. To examine and demonstrate that the long-term model is distinguishable and augments the expectation because of short time-scale expectation features, we compared the expectations generated by a combined short-term + long-term model to one based on expectations from short-term + scrambled long-term processes (null model). The resulting distributions shown in Figure 6 show a positive shift for the genuine models compared with the shuffled ones

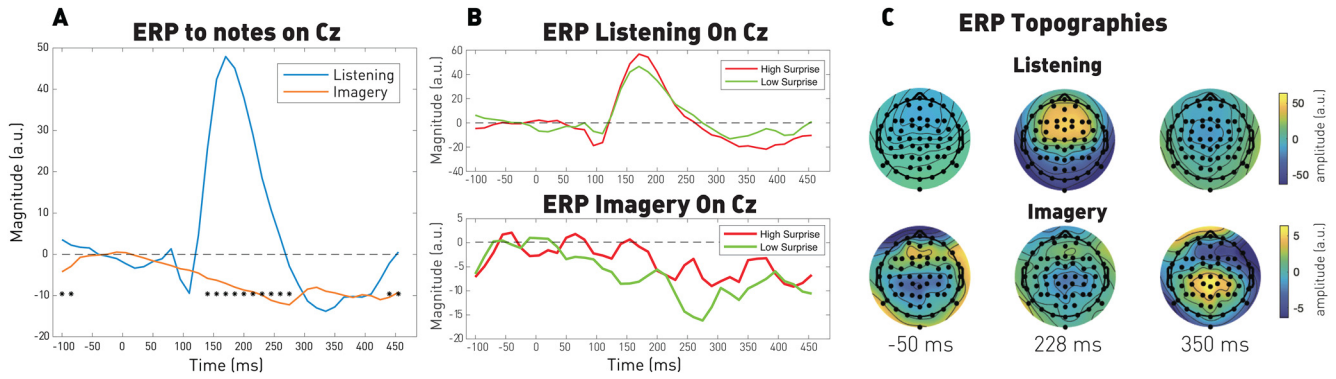

Figure 7. ERP analysis of listened and imagined notes. ***A***, Averaged responses for all notes. Significance between listening and imagery responses was computed using a permutation test from the values distributed by participants ($p < 0.05$). ***B***, Averaged responses for the 20% less and most expected notes in both listening (top) and imagery (bottom) conditions. ***C***, Participant-averaged topographic distributions from the ERP of all notes at least 500 ms away from the metronome.

(listening: $p = 9.5 \times 10^{-64}$, common language effect size: $f = 0.77$; imagery: $p = 1.2 \times 10^{-63}$, common language effect size: $f = 0.77$). We then used the same analysis approach on the short-term expectations, showing that the short-term model captures information not explained by the long-term model (listening: $p = 3.7 \times 10^{-41}$, common language effect size: $f = 0.72$; imagery: $p = 1.2 \times 10^{-77}$; common language effect size $f = 0.79$). The topographical distributions of such contributions resemble those seen with the full expectation signal (the expectation values built by combining long- and short-term statistics; see Fig. 4; short-term contribution: listening: $r = 0.67$, imagery: $r = 0.53$; long-term contribution: listening: $r = 0.75$, imagery: $r = 0.70$). Furthermore, similar topographic patterns were measured for the contributions of short- and long-term models (listening: $r = 0.64$; imagery: $r = 0.47$), suggesting that the neural activity explained by long- and short-term expectations originates is similar or overlapping brain areas.

Finally, we also examined the extent to which the correlation contribution because of the expectation signal is specifically related to the low-level features of the music signal (pitch, intervals, reversal in pitch direction, and duration). To do so, we compared the distribution of the correlations when regressing all these low-level features and the expectation signal on one side, compared with the distribution of the correlations computed when scrambling only the expectation vector (null model). The difference between the two distributions shown in Figure 6 indicated that the expectation signal indeed contributed information beyond that because of the low-level features (listening: $p = 3.8 \times 10^{-155}$; common language effect size: $f = 0.9$; imagery: $p = 7.9 \times 10^{-138}$; common language effect size: $f = 0.89$). All these comparisons led us to conclude that the long-term model, learned through exposure to a large corpus of music, is operable during both the listening and imagery conditions and in addition to the low-level musical features.

ERP analysis

We conducted an ERP analysis by computing the average neural response in a window of [−100 ms, 500 ms] around the note onsets at least 500 ms away from the metronome beats. The average power in the window of [−50 ms, 0 ms] was subtracted as a baseline. Significance between listening and imagery responses was computed using a permutation test from the values distributed by participants, and topographic distributions were computed by plotting the response power over the scalp at specific time latencies. Finally, we also computed averaged responses for the 20% most expected and 20% less expected notes.

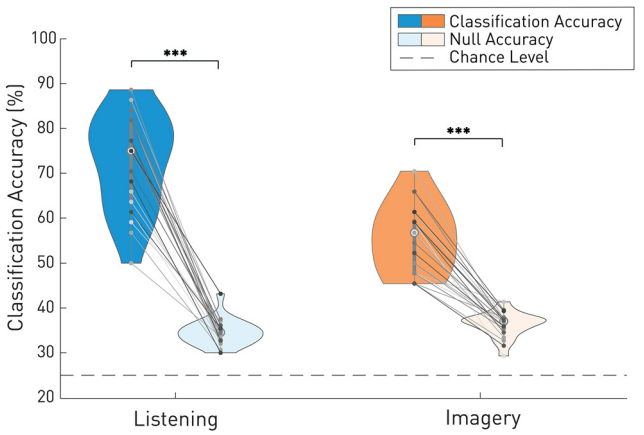

Figure 8. Piece classification accuracy. EEG predictions for note onsets and melodic expectations were combined to determine which song was being listened to or imagined. The data are shown for each participant and indicate overall significance. The null model was calculated from labels-shuffled data.

Figure 7*A* shows that imagined notes elicit negative responses that are similar to the TRF kernels observed in Figure 4. In addition, notes in both listening and imagery conditions elicited stronger responses on the Cz electrodes for notes related to low expectation (high surprise) as shown in Figure 7*B*. This trend, even if not significant here, is consistent with the TRF analysis and in line with the literature (Omigie et al., 2013; Di Liberto et al., 2020a). Finally, the topographic distribution of the ERPs in the two conditions is illustrated in Figure 7*C*, highlighting the relative delay and inverted polarity of the imagery relative to listened responses.

Decoding imagined song identity from the EEG

We tested whether the EEG encoding of note onset and melodic expectation was sufficiently robust to reliably classify the song identity on single trials. To do so, EEG recordings were predicted using the TRF by regressing all four musical stimuli separately. The stimulus leading to highest EEG prediction correlation was then selected for each trial (for more details, see Materials and Methods section). A null model was computed by shuffling the songs order to estimate the classification chance level.

Figure 8 shows significant classification accuracies, following the same trend, for each individual participant. Significance was computed using a Wilcoxon signed rank test paired by participants (listening: $p < 10^{-7}$, common language effect size $f = 1.0$;

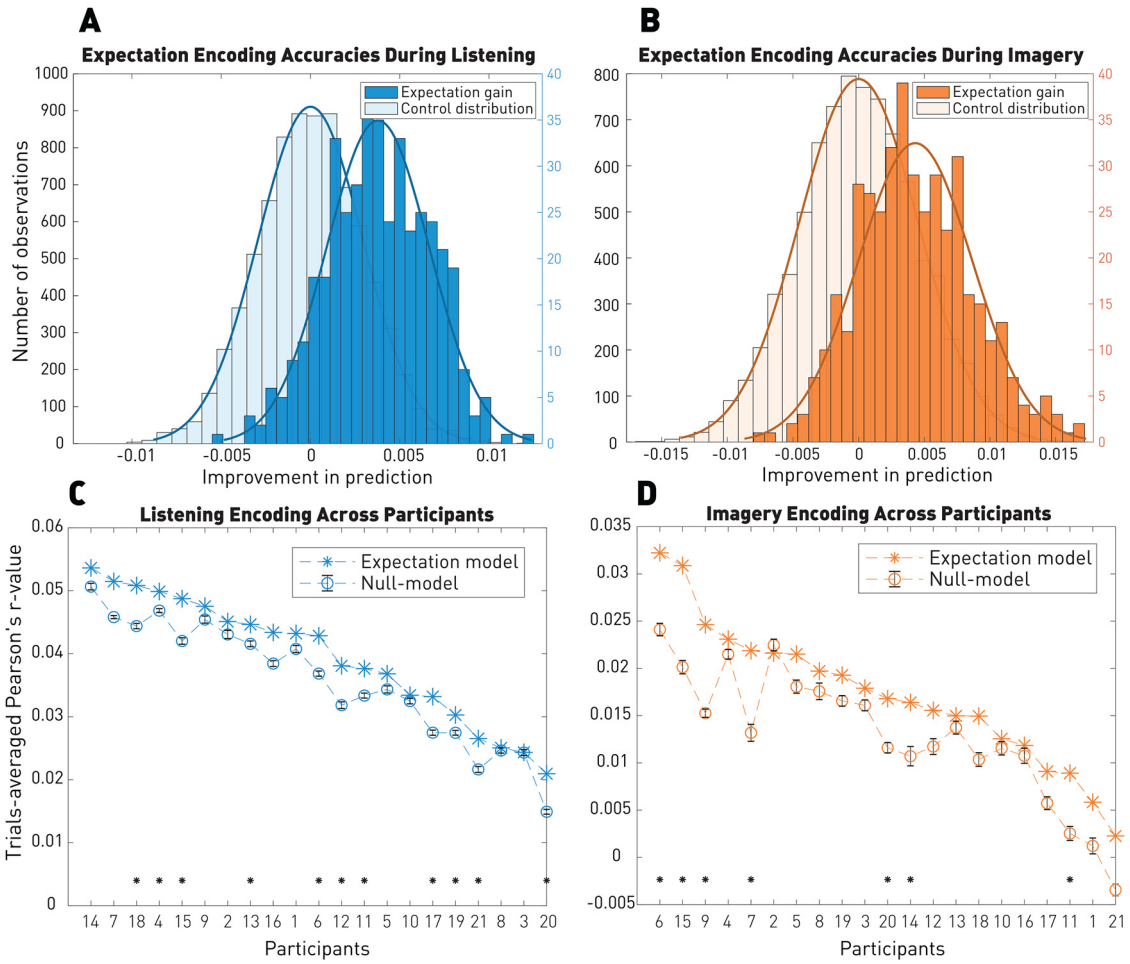

Figure 9. Cross-participants analysis. TRF models were fit by combining EEG data from all participants but one and evaluated on the left-out participant. ***A***, Distribution of expectation EEG prediction correlation gains (expectation – null model) during listening were significant when models were trained on different participants than the one of the evaluation. ***B***, Distribution of the expectation gain during imagery. The gain is conserved with models trained on different participants than the one of the evaluation. ***C***, ***D***, Individual EEG prediction correlations for the listening (***C***) and imagery (***D***) conditions. Error bars for null models indicate the SE across shuffles. Significance within participant: $*p < 0.05$.

imagery: $p < 10^{-7}$, common language effect size $f = 1.0$). Statistical significance was determined based on the null model performance rather than the theoretical chance level, which instead assumes infinite data points (Combrisson and Jerbi, 2015).

Cross-participants analysis

In order to assess the variability in the neural responses across individuals, we used a leave-one-participant-out cross-validation technique. Specifically, average TRF models were trained on all participants but one, which was instead used for evaluation. The goal was to test whether the neural signals of individual participants were sufficiently consistent and synchronized between participants to allow for significant EEG predictions.

Figure 9 shows that the cross-participants analysis allowed for significant encoding of expectation. Significance was computed using a Wilcoxon rank sum test between expectation gain and control distributions (listening: $p = 1.3 \times 10^{-108}$, common language effect size $f = 0.85$; imagery: $p = 8.8 \times 10^{-68}$, common language effect size $f = 0.78$). Results were also significant on 11 of 21 individual participants for listening and 7 of 21 participants for imagery. Significance ($p < 0.05$) was assessed by comparing the probability of the observed expectation prediction correlation with the null model distribution.

This analysis indicates that cortical responses were consistent between participants in both listening and imagery conditions, meaning that models can be trained and evaluated on different participants and that expectation encoding is shared between individuals within a same sociocultural environment (here professional classical musicians).

Comparison with behavioral audiation measures

The literature is rich in behavioral measures of audiation capabilities (Gerhardstein, 2002; Gelding et al., 2015; Halpern, 2015). We specified our analysis on one of these measures: the Gordon's AMMA designed by Edwin Gordon in 1989 to tackle audiation capabilities in musicians to tailor musical training and checked whether this test was correlated with the between-participant variability observed in our data.

Figure 10 shows that the onsets gain computed as the improvement of the onsets model with respects to its respective null model (labels shuffled) does not significantly correlate with the AMMA test. This finding suggests that the audiation capability as defined and measured by Gordon is something that is not reflected by the neural encoding of acoustics during imagery. A similar analysis based on the expectation gain instead of the acoustic gain has been conducted and resulted in similar results.

Discussion

Neural responses recorded with EEG during musical imagery exhibited detailed temporal dynamics that reflected the effects of

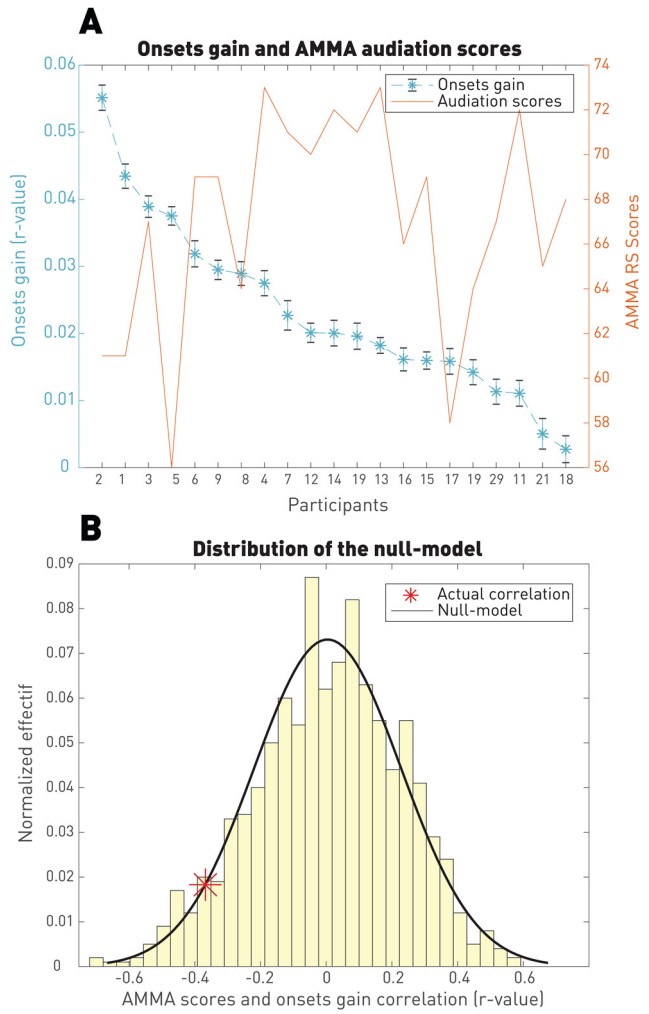

Figure 10. Correlation of the onset-model gain with the AMMA test. ***A***, Raw signals are shown in different axis. The Pearson's correlation computed on these two signals is $r = -0.36$. ***B***, This correlation is not significant as it resulted in $p > 0.05$ when looking at the null distribution built by shuffling the order of participants. We therefore conclude that the AMMA does not reflect the onsets gain.

melodic expectations, and a TRF that is delayed and with an inverted polarity relative to that of responses exhibited during listening. The responses shared substantial characteristics across individual participants, and were also strong and detailed enough to be robustly and specifically associated with the musical pieces that the participants listened to or imagined.

This study demonstrates, for the first time, that melodic expectation mechanisms are as faithfully encoded during imagery as during musical listening. EEG responses to music (and other signals such as speech) segments are typically modulated by the probability of hearing that sound within the ongoing sequence: the less probable (unexpected) it is, the stronger is the EEG expectation response (Di Liberto et al., 2020a). Therefore, the finding that imagined music is modulated similarly to listened music hints at the nature and role of musical expectation in setting the grammatical markers of our perception. Thus, as in speech, expectation mechanisms are used to parse the musical phrases and extract grammatical features to be used later in other purposes. This idea has already been discussed, and several studies have shown that musical expectations are used as primary features in other cognitive processes from memory (Agres et al., 2018) to musical pleasure (Gold et al., 2019). For instance, thwarted or fulfilled expectations have been shown to modulate activity in brain regions related to the reward system (Cheung et al., 2019), specifically to emotional pleasure (Blood and Zatorre, 2001; Zatorre and Salimpoor, 2013) and dopamine release (Salimpoor et al., 2011). Therefore, it is likely that imagery induces the same emotions and pleasure felt during musical listening because melodic expectations are encoded similarly in both cases. This explains why musical imagery is a versatile place for music creation and plays a significant role in music education.

When Robert Schumann asked his students to arrive at the point of "hearing music from the page," he suggested that there exists individual variability in the vividness of imagery, which can be shaped and improved by practice. This ability can be assessed via behavioral measures (Gerhardstein, 2002; Gelding et al., 2015; Halpern, 2015), and has also been shown to correlate with neural activity in fMRI (Halpern, 2015). Indeed, it may also reflect language deficits as seen in children with specific language impairment who often exhibit significantly lower scores in behavioral musical imagery tests, suggesting shared neurodevelopmental deficits (Heaton et al., 2018). Curiously, we did not find a significant correlation between the strength of the neural encoding of music and the participants' audiation scores from the widely used Gordon's AMMA test (Fig. 10). This can partially be explained by the weak SNR of the EEG signal, as well as by complex aptitudes that are not captured by the AMMA test. Therefore, we still lack an adequate demonstration of a link between our participants' ability to imagine and behavioral measures that can better indicate the cognitive underpinnings of the vividness of their imagery. By extension, the same lack of evidence applies to language deficits and their potential remediation through musical training.

From a system's perspective, auditory imagery responses can be thought of as "predictive" responses, induced by top-down processes that normally model how an incoming stimulus is perceived in the brain, or the perceptual equivalent of the efference copy, often triggered by the motor system (Ventura et al., 2009). This analogy has inspired numerous studies of auditory imagery in motor contexts as in covert speech, suggesting that imagined responses can be of a predictive motor nature (Tian and Poeppel, 2010, 2012, 2013; Whitford et al., 2017; Ding et al., 2019). In musical imagery, rhythm in particular has been closely linked to the activity of the supplementary motor areas and pre-supplementary motor areas (Halpern and Zatorre, 1999; Halpern, 2001; Meister et al., 2004; Zatorre and Halpern, 2005; Herholz et al., 2012; Lima et al., 2015, 2016; Gelding et al., 2019; Bastepe-Gray et al., 2020), whereas notational audiation (Brodsky et al., 2008) (musical imagery driven by reading music scores) and listening (Pruitt et al., 2019) have been shown to generate covert excitation of the vocal folds with a neural signature similar to that observed during musical imagery (Zatorre et al., 1996). This motor-imagery link also runs in reverse, as demonstrated by an ECoG study that reveals strong auditory responses induced by silent playing of a keyboard (Martin et al., 2018). In conclusion, it is evident that imagery may well be facilitated by the intimate links that exist between motor and sensory areas that are normally coactivated in task performance, for example, vocal-tract and speech production (Shamma et al., 2020), fingers and piano playing, and vision and reading. This also makes it difficult experimentally to disentangle the two sources of activity (Zatorre et al., 2007) since auditory imagery may partially be affected by motor components (Halpern and Zatorre, 1999).

Regardless of their origins, imagery responses should be fully considered as top-down predictive signals, with the most striking

evidence in our data being their inverted polarity relative to the listening responses. Such an inversion facilitates the comparison between bottom-up sensory activation and its top-down prediction by generating the "error" signal, long postulated in predictive coding theories to be the critical information that is propagated deep into the brain (Rao and Ballard, 1999; Koster-Hale and Saxe, 2013). This key observation is explored in detail in the companion study (Di Liberto et al., 2021), which analyzed the EEG responses evoked during the pauses or short silences that are naturally interspersed within a musical score. These responses are analogous to imagery responses in that both lack direct stimuli to evoke them. The combined findings in the present work and the companion study provide a common framework that remarkably and seamlessly links listened and imagined music perception, and more broadly, sensory responses and their prediction in the brain.

References

Abdallah S, Plumbley M (2009) Information dynamics: patterns of expectation and surprise in the perception of music. Connect Sci 21:89–117.

Agres K, Abdallah S, Pearce M (2018) Information-theoretic properties of auditory sequences dynamically influence expectation and memory. Cogn Sci 42:43–76.

Ammirante P, Patel AD, Russo FA (2016) Synchronizing to auditory and tactile metronomes: a test of the auditory-motor enhancement hypothesis. Psychon Bull Rev 23:1882–1890.

Bastepe-Gray SE, Acer N, Gumus KZ, Gray JF, Degirmencioglu L (2020) Not all imagery is created equal: a functional magnetic resonance imaging study of internally driven and symbol driven musical performance imagery. J Chem Neuroanat 104:101748.

Blood AJ, Zatorre RJ (2001) Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. Proc Natl Acad Sci USA 98:11818–11823.

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28:803–809.e3.

Brodsky W, Kessler Y, Rubinstein BS, Ginsborg J, Henik A (2008) The mental representation of music notation: notational audiation. J Exp Psychol Hum Percept Perform 34:427–445.

Bunzeck N, Wuestenberg T, Lutz K, Heinze HJ, Jancke L (2005) Scanning silence: mental imagery of complex sounds. Neuroimage 26:1119–1127.

Cervantes Constantino F, Simon JZ (2017) Dynamic cortical representations of perceptual filling-in for missing acoustic rhythm. Sci Rep 7:17536.

Cheung VK, Harrison PM, Meyer L, Pearce MT, Haynes JD, Koelsch S (2019) Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. Curr Biol 29:4084–4092.e4.

Combrisson E, Jerbi K (2015) Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 250:126–136.

Crosse MJ, Di Liberto GM, Lalor EC (2016) The Multivariate Temporal Response Function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604.

Daube C, Ince RA, Gross J (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr Biol 29:1924–1937.e9.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.

Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, de Cheveigné A, Shamma S, Mesgarani N (2020a) Cortical encoding of melodic expectations in human temporal cortex. Elife 9:e51784.

Di Liberto GM, Pelofi C, Shamma S, de Cheveigné A (2020b) Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. Acoust Sci Tech 41:361–364.

Di Liberto GM, Marion G, Shamma S (2021) The music of silence reveals neural auditory predictions. J Neurosci. Advance online publication. Retrieved 2 August 2021. doi:10.1523/JNEUROSCI.0184-21.2021.

Ding Y, Zhang Y, Zhou W, Ling Z, Huang J, Hong B, Wang X (2019) Neural correlates of music listening and recall in the human brain. J Neurosci 39:8112–8123.

Egermann H, Pearce MT, Wiggins GA, McAdams S (2013) Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. Cogn Affect Behav Neurosci 13:533–553.

Gelding RW, Thompson WF, Johnson BW (2015) The pitch imagery arrow task: effects of musical training, vividness, and mental control. PLoS One 10:e0121809.

Gelding RW, Thompson WF, Johnson BW (2019) Musical imagery depends upon coordination of auditory and sensorimotor brain activity. Sci Rep 9:16823.

Gerhardstein RC (2002) The historical roots and development of audiation: a process for musical understanding. In: Musical understanding: perspectives in theory and practice (Hanley B, Goolsby TW, eds). Canada: Canadian Music Educators' Association.

Gillick J, Tang K, Keller R (2010) Learning jazz grammars. Computer Music J 34:56–66.

Godoy R, Jorgensen H (2012) Musical imagery. New York: Routledge Research in Music.

Gold BP, Pearce MT, Mas-Herrero E, Dagher A, Zatorre RJ (2019) Predictability and uncertainty in the pleasure of music: a reward for learning? J Neurosci 39:9397–9409.

Griffiths TD (1999) Human complex sound analysis. Clin Sci (Lond) 96:231–234.

Halpern AR (2001) Cerebral substrates of musical imagery. Ann NY Acad Sci 930:179–192.

Halpern AR (2015) Differences in auditory imagery self-report predict neural and behavioral outcomes. Psychomusicology 25:37–47.

Halpern AR, Zatorre RJ (1999) When that tune runs through your head: a PET investigation of auditory imagery for familiar melodies. Cereb Cortex 9:697–704.

Halpern AR, Zatorre RJ, Bouffard M, Johnson JA (2004) Behavioral and neural correlates of perceived and imagined musical timbre. Neuropsychologia 42:1281–1292.

Heaton P, Tsang WF, Jakubowski K, Mullensiefen D, Allen R (2018) Discriminating autism and language impairment and specific language impairment through acuity of musical imagery. Res Dev Disabil 80:52–63.

Herholz SC, Halpern AR, Zatorre RJ (2012) Neuronal correlates of perception, imagery, and memory for familiar tunes. J Cogn Neurosci 24:1382–1397.

Hubbard T (2013) Auditory aspects of auditory imagery. In: Multisensory imagery. New York: Springer.

Janata P (2015) Neural basis of music perception. In: The human auditory system, Vol. 129: Handbook of clinical neurology (Aminoff MJ, Boller F, Swaab DF, eds), pp 187–205. Amsterdam: Elsevier.

Jäncke L, Loose R, Lutz K, Specht K, Shah N (2000) Cortical activations during paced finger-tapping applying visual and auditory pacing stimuli. Brain Res Cogn Brain Res 10:51–66.

Koelsch S, Siebel WA (2005) Towards a neural basis of music perception. Trends Cogn Sci 9:578–584.

Koster-Hale J, Saxe R (2013) Theory of mind: a neural prediction problem. Neuron 79:836–848.

Kraemer DJ, Macrae CN, Green AE, Kelley WM (2005) Sound of silence activates auditory cortex. Nature 434:158–158.

Krumhansl C, Louhivuori J, Toiviainen P, Järvinen T, Eerola T (1999) Melodic expectation in Finnish spiritual folk hymns: convergence of statistical, behavioral, and computational approaches. Music Perception 17:151–195.

Krumhansl C, Toivanen P, Eerola T, Toiviainen P, Järvinen T, Louhivuori J (2000) Cross-cultural music cognition: cognitive methodology applied to North Sami yoiks. Cognition 76:13–58.

Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31:189–193.

Lima CF, Lavan N, Evans S, Agnew Z, Halpern AR, Shanmugalingam P, Meekings S, Boebinger D, Ostarek M, McGettigan C, Warren JE, Scott SK (2015) Feel the noise: relating individual differences in auditory imagery to the structure and function of sensorimotor systems. Cereb Cortex 25:4638–4650.

Lima CF, Krishnan S, Scott SK (2016) Roles of supplementary motor areas in auditory processing and auditory imagery. Trends Neurosci 39:527–542.

Manzara LC, Witten IH, James M (1992) On the entropy of music: an experiment with Bach chorale melodies. Leonardo Music J 2:81–88.

Martin S, Mikutta C, Leonard MK, Hungate D, Koelsch S, Shamma S, Chang EF, Millán JR, Knight RT, Pasley BN (2018) Neural encoding of auditory features during music perception and imagery. Cereb Cortex 28:4222–4233.

Meister I, Krings T, Foltys H, Boroojerdi B, Müller M, Töpper R, Thron A (2004) Playing piano in the mind: an fMRI study on music imagery and performance in pianists. Brain Res Cogn Brain Res 19:219–228.

Meyer L (1973) Explaining music: essays and explorations. Berkeley, CA: University of California.

Miller KJ, Schalk G, Fetz EE, den Nijs M, Ojemann JG, Rao RP (2010) Cortical activity during motor execution, motor imagery, and imagery-based online feedback. Proc Natl Acad Sci USA 107:4430–4435.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706.

Omigie D, Pearce M, Stewart L (2012) Tracking of pitch probabilities in congenital amusia. Neuropsychologia 50:1483–1493.

Omigie D, Pearce MT, Williamson VJ, Stewart L (2013) Electrophysiological correlates of melodic processing in congenital amusia. Neuropsychologia 51:1749–1762.

Omigie D, Pearce MT, Lehongre K, Hasboun D, Navarro V, Adam C, Samson S (2019) Intracranial recordings and computational modelling of music reveal the time-course of prediction error signaling in frontal and temporal cortices. J Cogn Neurosci 31:855–873.

Opitz B, Rinne T, Mecklinger A, von Cramon D, Schröger E (2002) Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. Neuroimage 15:167–174.

Pearce MT (2005) The construction and evaluation of statistical models of melodic structure in music perception and composition. PhD dissertation.

Pearce MT (2018) Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. Ann NY Acad Sci 1423:378–395.

Pruitt T, Halpern A, Pfordresher P (2019) Covert singing in anticipatory auditory imagery. Psychophysiology 56:e13297.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Repp BH (2005) Sensorimotor synchronization: a review of the tapping literature. Psychon Bull Rev 12:969–992.

Repp BH, Penel A (2004) Rhythmic movement is attracted more strongly to auditory than to visual rhythms. Psychol Res 68:252–270.

Rohrmeier M (2011) Towards a generative syntax of tonal harmony. J Math Music 5:35–53.

Salimpoor VN, Benovoy M, Larcher K, Dagher A, Zatorre RJ (2011) Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nat Neurosci 14:257–262.

Schönwiesner M, Novitski N, Pakarinen S, Carlson S, Tervaniemi M, Näätänen R (2007) Heschl's gyrus, posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes. J Neurophysiol 97:2075–2082.

Shamma S, Patel P, Mukherjee S, Marion G, Khalighinejad B, Han C, Herrero J, Bickel S, Mehta A, Mesgarani N (2020) Learning speech production and perception through sensorimotor interactions. Cereb Cortex Commun 2:tgaa091.

Song Y, Dixon S, Pearce MT, Halpern AR (2016) Perceived and induced emotion responses to popular music: categorical and dimensional models. Music Perception 33:472–492.

Tian X, Poeppel D (2010) Mental imagery of speech and movement implicates the dynamics of internal forward models. Front Psychol 1:166.

Tian X, Poeppel D (2012) Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. Front Hum Neurosci 6:314.

Tian X, Poeppel D (2013) The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. J Cogn Neurosci 25:1020–1036.

Tillmann B, Janata P, Bharucha JJ (2003) Activation of the inferior frontal cortex in musical priming. Brain Res Cogn Brain Res 16:145–161.

Van Canneyt J, Wouters J, Francart T (2020) Enhanced neural tracking of the fundamental frequency of the voice. bioRxiv 2020.10.28.359034.

Ventura MI, Nagarajan SS, Houde JF (2009) Speech target modulates speaking induced suppression in auditory cortex. BMC Neurosci 10:58.

Whitford TJ, Jack BN, Pearson D, Griffiths O, Luque D, Harris AW, Spencer KM, Le Pelley ME (2017) Neurophysiological evidence of efference copies to inner speech. Elife 6:e28197.

Yoo SS, Lee CU, Choi BG (2001) Human brain mapping of auditory imagery: event-related functional MRI study. Neuroreport 12:3045–3049.

Zatorre RJ, Halpern AR (2005) Mental concerts: musical imagery and auditory cortex. Neuron 47:9–12.

Zatorre RJ, Salimpoor VN (2013) From perception to pleasure: music and its neural substrates. Proc Natl Acad Sci USA 110 Suppl 2:10430–10437.

Zatorre RJ, Halpern AR, Perry DW, Meyer E, Evans AC (1996) Hearing in the mind's ear: a pet investigation of musical imagery and perception. J Cogn Neurosci 8:29–46.

Zatorre RJ, Chen JL, Penhune VB (2007) When the brain plays music: auditory–motor interactions in music perception and production. Nat Rev Neurosci 8:547–558.

Zhang Y, Chen G, Wen H, Lu KH, Liu Z (2017) Musical imagery involves Wernicke's area in bilateral and anti-correlated network interactions in musicians. Sci Rep 7:17066.