Behavioral/Cognitive

The Music of Silence: Part II: Music Listening Induces Imagery Responses

Giovanni M. Di Liberto,[1,2,3] Guilhem Marion,[1,4] and Shihab A. Shamma[1,5]

[1]Laboratoire des systémes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, 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, Dublin 2, Ireland, [3]School of Electrical and Electronic Engineering and University College Dublin Centre for Biomedical Engineering, University College Dublin, Dublin 4, Ireland, [4]Research Chair on Beauty Studies, Université Paris Sciences et Lettres, 75006 Paris, France, and [5]Institute for Systems Research, Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20740

During music listening, humans routinely acquire the regularities of the acoustic sequences and use them to anticipate and interpret the ongoing melody. Specifically, in line with this predictive framework, it is thought that brain responses during such listening reflect a comparison between the bottom-up sensory responses and top-down prediction signals generated by an internal model that embodies the music exposure and expectations of the listener. To attain a clear view of these predictive responses, previous work has eliminated the sensory inputs by inserting artificial silences (or sound omissions) that leave behind only the corresponding predictions of the thwarted expectations. Here, we demonstrate a new alternate approach in which we decode the predictive electroencephalography (EEG) responses to the silent intervals that are naturally interspersed within the music. We did this as participants (experiment 1, 20 participants, 10 female; experiment 2, 21 participants, 6 female) listened or imagined Bach piano melodies. Prediction signals were quantified and assessed via a computational model of the melodic structure of the music and were shown to exhibit the same response characteristics when measured during listening or imagining. These include an inverted polarity for both silence and imagined responses relative to listening, as well as response magnitude modulations that precisely reflect the expectations of notes and silences in both listening and imagery conditions. These findings therefore provide a unifying view that links results from many previous paradigms, including omission reactions and the expectation modulation of sensory responses, all in the context of naturalistic music listening.

Key words: **EEG; expectation; omission; predictive processing; TRF**

Significance Statement

Music perception depends on our ability to learn and detect melodic structures. It has been suggested that our brain does so by actively predicting upcoming music notes, a process inducing instantaneous neural responses as the music confronts these expectations. Here, we studied this prediction process using EEGs recorded while participants listen to and imagine Bach melodies. Specifically, we examined neural signals during the ubiquitous musical pauses (or silent intervals) in a music stream and analyzed them in contrast to the imagery responses. We find that imagined predictive responses are routinely co-opted during ongoing music listening. These conclusions are revealed by a new paradigm using listening and imagery of naturalistic melodies.

Introduction

Silence is an essential component of our auditory experience, which serves important communicative functions by contributing to expectation, emphasis, and emotional expression. Here, we investigate the neural encoding of silence with electroencephalography (EEG) and music stimuli.

That perception is underpinned by an interplay of sensory input and endogenous neural processes has been a long-standing area for debate (den Ouden et al., 2012; Pouget et al., 2013; Clark, 2016; Heeger, 2017). Prediction theories (Spratling, 2017) suggest that the brain continuously attempts to predict

Received Jan. 22, 2021; revised June 22, 2021; accepted June 24, 2021.

Author contributions: G.M.D.L., G.M., and S.A.S. designed research; G.M.D.L. and G.M. performed research; G.M.D.L. analyzed data; G.M.D.L., S.A.S., and G.M. wrote the paper.

This work was supported by an Advanced European Research Council grant (NEUME, 787836) and Air Force Office of Scientific Research and National Science Foundation grants to S.A.S., FrontCog Grant ANR-17-EURE-0017, and Université Paris Sciences et Lettres Grant ANR-10-IDEX-0001-02. G.M. was supported by a PhD scholarship from Chaire Beauté(s) PSL-L'Oréal.

The authors declare no competing financial interests.

Correspondence should be addressed to Giovanni M. Di Liberto at diliberg@tcd.ie.

https://doi.org/10.1523/JNEUROSCI.0184-21.2021

upcoming sensory inputs, comparing (subtracting) them and hence deriving a prediction error (δ_{sur}) that is used to improve an internal (prediction) model of the world. A large body of research has found prediction effects to be in line with several neurophysiological phenomena, such as the magnitude modulation of sensory responses with the expectation of these responses (Kutas and Hillyard, 1980, 1984; Rabovsky et al., 2018), where larger responses were measured for more unexpected inputs. In auditory neurophysiology, this prediction phenomenon has been extensively investigated using the responses evoked by sound stimuli (Sutton et al., 1965; Friederici et al., 1993; Mars et al., 2008; Kutas and Federmeier, 2011; Strauß et al., 2013; Seer et al., 2016). A less common approach involves studying the predictions in the absence of the acoustic input, that is, during silence, a strategy that potentially unveils neural predictive processing and top-down mechanisms of this processing by decoupling it from the simultaneous bottom-up sensory inputs (Heilbron and Chait, 2018; Walsh et al., 2020).

Vigorous responses to silences have been observed across modalities when a sensory stimulus was strongly expected, for example, corresponding to an omission during the rapid isochronous presentation of tones (Simson et al., 1976; Joutsiniemi and Hari, 1989; Yabe et al., 1997; Chennu et al., 2016). This finding demonstrated that unexpected silences can elicit robust neural responses that do not require a concurrent sensory input. However, silence has a much more pervasive presence in our auditory experience than what can be captured in the stimulus omission scenario, which is limited to silences occurring in place of highly expected stimuli. In fact, silence is a fundamental component of the rhythmic structure of music, which can correspond to a wide range of expectation strengths. The regularities of music prompt our brain to build such expectations, which are accurately estimated by computational models of musical structure (Pearce, 2005), allowing us to assess the precise neural encoding of music expectations. Although such expectations have been shown to be encoded in the neural responses to notes in a melody during listening (Di Liberto et al., 2020), little is known about the neural encoding of internally generated music.

Here, we investigate the role of silence on the neural processing of music with EEGs recorded as participants listened to or performed mental imagery of excerpts from Bach chorales. Endogenous and exogenous components of the neural signal are discerned by studying the comparison between listening and imagery conditions. According to prediction theories, the brain continuously builds predictions of upcoming music notes, with the prediction (P) signal appropriately modulated by the uncertainty of the prediction (Koelsch et al., 2019). When subtracted from the sensory (S) response, it produces a surprise or prediction error signal that is measurable with EEG (δ_{sur} = S−P; Heilbron and Chait, 2018; Grisoni et al., 2019). In this study, we assumed S and −P to contribute to the EEG signal as two distinct additive components, where P mimics S and, conversely, −P has inverse polarity compared with S. Under that assumption, encountering silence when a note is plausible would correspond to a measurable EEG signal reflecting the neural prediction error signal δ_{sur} = −P, which depends solely on the prediction signal P as S = 0, thus presenting the inverse polarity of the otherwise dominant sensory response (Fig. 1; Bendixen et al., 2009; Heilbron and Chait, 2018). For these reasons, we hypothesized robust neural correlates to emerge in correspondence to the silent-events of music, reflecting the prediction error δ_{sur} = −P and with magnitude changing with the expectation strengths.

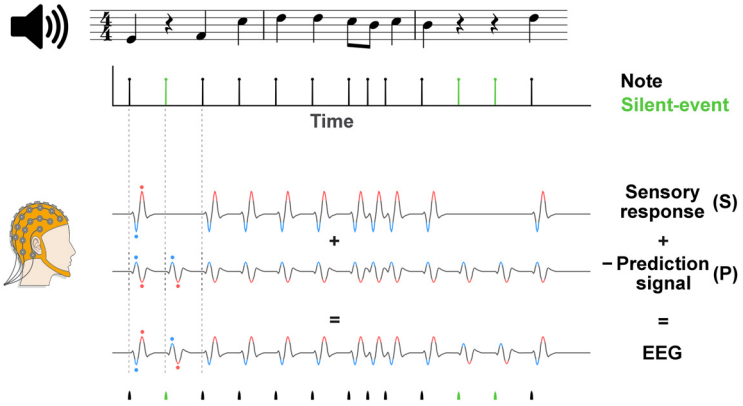

Figure 1. Simplified predictive processing model demonstrating the predictive processing hypothesis for the perception of melodies. The EEG signal recorded during monophonic music listening was hypothesized to reflect the linear combination of an S-evoked response and a neural P signal. In line with the predictive processing framework, we modeled the EEG signal as a combination of the distinct components S and P; specifically as the subtraction S−P or equivalently S + (−P). Having defined P as a signal reflecting the attempt of our brain to predict the sensory stimulus, we posited P to emulate S (with |S|>|P|) and to have larger magnitude with stronger expectations (For simplicity, the expectation strengths are not included in this figure.). As such, the S−P signal would become −P when a prediction is possible but no sensory stimulus is present (S = 0), producing an overall EEG signal with inverse polarity compared with the response to a note. In other words, EEG responses with opposite polarities were expected for events with and without an input sound. Black lines at the top and bottom of the figure indicate notes; green lines indicate silent-events. Neural signals for notes and silent-events were expected to exhibit inverted polarities. Polarities in the neural signals were highlighted in red (positive) and blue (negative) and dots were used for ease of visualisation. After selecting silent-events as the instants where a note was plausible but did not occur (based on IDyOM, see below, Materials and Methods), the existence and precise dynamics of the prediction signal was assessed (1) by comparing the responses to silent-events during melody listening, where P could be measured in isolation as S = 0, (2) by studying the neural processing of music during imagery, where P could be isolated as S = 0 for both notes and silent-events, and (3) by separating S and P with a component analysis method.

The music imagery task allowed us to study the neural encoding of music silence further by investigating endogenous neural components in absence of sensory responses. In an accompanying study (Marion et al., 2021), we have shown robust neural activation corresponding to imagined notes, extending previous work on auditory imagery (Halpern and Zatorre, 1999; Kraemer et al., 2005; Zhang et al., 2017) by demonstrating that cortical signals encode melodic expectation during imagery. Here, in line with prediction theories, we hypothesized that P is the main source of such neural activity as S = 0. As such, we anticipated a prediction signal (δ_{sur} = −P) to emerge in the EEG responses to both imagined notes and silent-events, with inverse polarity relative to a sensory response. Finally, we anticipated the magnitude of the responses to silent-events to reflect the precise expectation strengths of each music event, which were estimated by means of a computational model of melodic structure (Pearce, 2005) as it was demonstrated for music listening (Di Liberto et al., 2020) and imagery (Marion et al., 2021).

Materials and Methods

EEG experiment 1

Data acquisition and experimental paradigm. Twenty healthy subjects (10 female, between 23 and 42 years old, median = 29) participated in the EEG experiment. Ten of them were highly trained musicians with

a degree in music and at least 10 years of experience, whereas the other participants had no musical background. Each subject reported no history of hearing impairment or neurologic disorder, provided written informed consent, and was paid for participating. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the Health Research Ethics Evaluation Board 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 512 Hz using a BioSemi ActiveTwo system. Audio stimuli were presented at a sampling rate of 44,100 Hz using Sennheiser HD 650 headphones and Presentation software (http://www.neurobs.com). Testing was conducted at École Normale Supérieure, in a dark room, and subjects were instructed to maintain visual fixation on a crosshair centered on the screen and to minimize motor activities while music was presented.

Stimuli and procedure. Monophonic musical instrument digital interface (MIDI) versions of 10 music pieces from Bach's monodic instrumental corpus were partitioned into short snippets of ~150 s. The selected melodies were originally extracted from violin [partita Bach Works Catalog (BWV) 1001, presto; BWV 1002, allemande; BWV 1004, allemande and gigue; BWV 1006, loure and gavotte] and flute (partita BWV 1013 allemande, corrente, sarabande, and bourrée angloise) scores and were synthesized by using piano sounds with MuseScore 2 software, each played with a fixed rate (between 47 and 140 bpm). This was done to reduce familiarity for the expert pianist participants while enhancing their neural response by using their preferred instrument timbre (Pantev et al., 2001). Each 150 s piece, corresponding to an EEG trial, was presented three times throughout the experiment, adding up to 30 trials that were presented in a random order. At the end of each trial, participants were asked to report on their familiarity with the piece (from 1 = unknown to 7 = know the piece very well). This rating could take into account both their familiarity with the piece on first occurrence in the experiment as well as the buildup of familiarity across repetitions. Participants reported repeated pieces as more familiar (paired t test on the average familiarity ratings for all participants across repetitions: rep2 > rep1, $p = 6.9 \times 10–6$; rep3 > rep2, $p = 0.003$, Bonferroni correction). No significant difference emerged between musicians and nonmusicians on this account (two-sample t test, $p = 0.07$, 0.16, 0.19 for repetitions 1, 2, and 3, respectively; Di Liberto et al., 2020).

EEG experiment 2

Data acquisition and experimental paradigm. Twenty-one healthy subjects (6 female between 17 and 35 year old, median = 25) participated in the EEG experiment. All participants were highly trained musicians with a degree in music. Each subject 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 Health Research Ethics Evaluation Board 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 and digitized at 2048 Hz using a BioSemi ActiveTwo system. Three additional electrodes were placed on the upper midline of participants' neck, jaw, and right wrist to control for motor movements of the tongue, masseter muscle, and forearm fingers extensors, respectively. Audio stimuli were presented at a sampling rate of 44,100 Hz using a Genelec 8010 loudspeaker and custom Python code. Testing was conducted at École Normale Supérieure in a dimmed room. Participants were instructed to minimize motor activities while performing the task.

The experiment consisted of 88 trials in which participants were asked to either listen or perform mental imagery of ~35 s melodies from a corpus of Bach chorales (see below, Stimuli and procedure). The entire stimulus set consisted of four such melodies, with each melody being presented 11 times per condition (listening and imagery) over the duration of the experiment. The presentation order of the resulting 88 trials was randomized. Participants were asked to read the music scores placed at the center of the desk during both listening and imagery conditions. Participants were provided with the scores before the experiment and were asked to become familiar with the melodies. This pre-exposure to the music material was planned to maximize the imagery performance. A tactile metronome (Peterson Body Beat Vibe Clip) marking the start of 100 bpm bars (each 2.4 s) was placed on the left ankle of all participants to allow them to perform the mental imagery task with high temporal precision. A constant lag of 35 ms was determined during the pilot experiments based on the subjective report on the participants, who reported that the metronome with lag 0 ms was not in sync with the music. That correction was applied for all participants with the same lag value. Neural data from 0 to 500 ms after each metronome onset were excluded from the main analyses in Figures 2 and 3 to ensure that the results do not reflect tactile responses. The metronome responses were analyzed separately to assess the dynamics of the tactile response (Fig. 3*G*). Note that the EEG response to the metronome reflects a mixture of tactile and auditory responses in the listening condition.

Before the experiment, musical imagery skills (or audiation skills) were assessed for every subject with the Advanced Measures of Music Audiation test (https://giamusicassessment.com/).

Stimuli and procedure. Four melodies were selected from a monophonic MIDI corpus of Bach chorales (BWV 349, BWV 291, BWV 354, BWV 271). All chorales use similar compositional principles. The composer takes a melody from a Lutheran hymn (cantus firmus) and harmonizes three lower parts (alto, tenor, and bass) accompanying the initial melody on soprano. The monophonic version of those melodies consist of the canti firmi. Original keys were used. The four melodies are based on a common grammatical structure and show very similar melodic and rhythmic patterns. The audio stimuli were synthesized using Fender Rhodes simulation software (Neo-Soul Keys) with 100 bpm, each corresponding to the start of a bar (every 2.4 s).

EEG data preprocessing

Neural data from both experiments were analyzed offline using MATLAB software (MathWorks). EEG signals were digitally filtered between 1 and 30 Hz using a Butterworth zero-phase filter (low- and high-pass filters both with order 2 and implemented with the function filtfilt), and down sampled to 64 Hz. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. Channels were then rereferenced to the average of the 64 channels. The temporal response function (TRF) weights did not qualitatively change when using high-pass filters down to 0.1 Hz. Low-frequencies below 1 Hz were crucial for the melodic expectations analysis (see Fig. 5), which was based on EEG data filtered between 0.1 and 30 Hz [Marion et al. (2021) has a more extensive analysis on the EEG frequency band.].

Information dynamics of music model

The information dynamics of music (IDyOM; Pearce, 2005) model is a framework based on variable-order, hidden Markov models. Given a note sequence of a melody, the probability distribution over every possible note continuation is estimated for every n-gram context up to a given length k (model order). The distributions for the various orders were combined according to an entropy-based weighting function (IDyOM; Pearce, 2005, Section 6.2). Here, we used an unbounded implementation of IDyOM that builds n-grams using contexts up to the size of each music piece. In addition, predictions were the result of a combination of long-term models (LTM) and short-term models (STM), which yields better estimates than either model alone. The LTM was the result of a pretraining on a large corpus of Western music that did not include the stimuli presented during the EEG experiment, thus simulating the statistical knowledge of a listener that was implicitly acquired after a lifetime of exposure to music. The STM, on the other hand, is constructed online for each individual music piece that was used in the EEG experiment.

Our choice of IDyOM was motivated by the empirical support that Markov-model-based frameworks received as a model of human melodic expectation (Pearce and Wiggins, 2006; Pearce et al., 2010; Omigie et al., 2013; Quiroga-Martinez et al., 2019). Furthermore, a previous study from our laboratory demonstrated robust coupling between the melodic expectations calculated with this configuration of IDyOM and cortical responses to music (Di Liberto et al., 2020).

Music features

In the present study, we have assessed the coupling between the EEG data and various features of the music stimuli. The note onset time

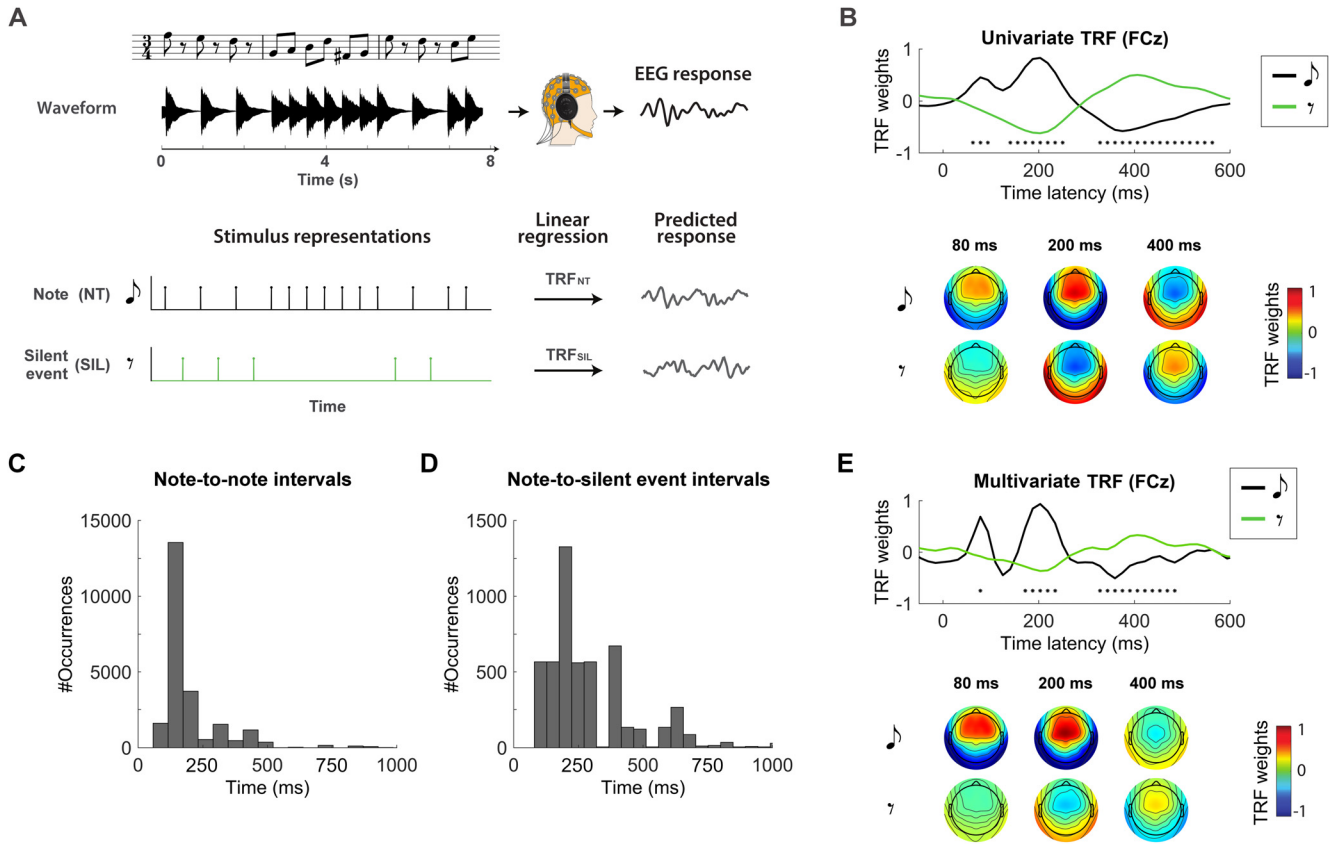

Figure 2. Robust cortical response to silence during music listening. ***A***, Experiment 1 setup. The EEG signal was recorded as participants listened to monophonic piano music. Univariate vectors were defined that mark with value 1 the onset of either notes (NT) or silent-events (SIL). A system identification procedure based on lagged linear regression was performed between each vector and the neural signal that minimizes the EEG prediction error. ***B***, The regression weights represent the TRF describing the coupling of the EEG signal TRF_{NT} and silent-events TRF_{SIL}. TRFs at the representative channel FCz are shown (top), revealing significant differences (FDR corrected Wilcoxon test, $*q < 0.001$) between the neural signature of note and silent-event because of inverted polarities, as clarified by the topographies of the TRF components (bottom). ***C***, ***D***, Overall distribution of time intervals between notes and between silent-events and the immediately preceding note. The *y*-axis indicates the number of occurrences for a given bin of time intervals when considering all trials. The data show that a large number of silent-events occurred <200 ms after a note, implying that in experiment 1, TRF_{SIL} could have potentially been affected by the late response to the previous note. ***E***, The analysis from ***B*** was rerun by using multivariate TRF models, i.e., considering note and silent-event vectors simultaneously with multivariate lagged regression to account for possible interaction between the two. The figure shows the regression weights corresponding to the two regressors at the selected channel FCz, and the topographies show the regression weights. As for the univariate TRF result, significant differences were found between note and silent-event TRFs (FDR corrected Wilcoxon test, $*q < 0.001$). TRF_{NT} showed qualitatively more pronounced early TRF components.

information was extracted from the MIDI files and encoded into time-series marking with an impulsemarking all note onsets (NT) with an impulse with value one. The time-series had length matching that of the corresponding music piece and had the same sampling frequency as the EEG data (Fig. 2*A*). We then used IDyOM to identify silent-events, that is, time instants without a note but where a note could have plausibly occurred. IDyOM does not encode silent-events explicitly, so we applied custom changes to the original Lisp programming language to extract the information of interest on the silent-events without changing the way IDyOM operates. Specifically, for each note, with a quantization of 1/16th of a bar, IDyOM was used to search for the time for the next most likely event. The search continued for progressively longer latencies until the model predicted a note with high likelihood (>0.3). We called those instants silent-events. The procedure was repeated on the silent-event instants to predict where the next note would occur by knowing that there was no note where the model had predicted one. This information was then encoded into time-series marking with an unit impulse each silent-event onset (SIL). Experiment 1 had a total of 23,514 notes and 5202 silent-events. In experiment 2, 1548 notes and 271 silent-events were used to fit the TRF in each condition (listening and imagery). Note that such events co-occurring with the tactile metronome were excluded. Figure 2, *C* and *D*, and Figure 3, *E* and *F*, report additional information on the distribution of notes and silent-events in the two experiments.

To investigate the cortical processing of note and silence expectations, we estimated the surprise and entropy values for each individual note of a given music piece by using IDyOM. Given a note e_i, a note sequence $e_{1.n}$, which immediately precedes that note, and an alphabet E describing the possible onset time values for the note, surprise $S(e_i|e_1.i-1)$ refers to the inverse probability of occurrence of a particular note at a given position in the melody (MacKay, 2003; Pearce et al., 2010) as follows:

$$S(e_i|e_{1.i-1}) = log_2 \frac{1}{p(e_i|e_{1.i-1})}.$$

The entropy in a given melodic context was defined as the Shannon (1948) entropy computed by averaging the surprise over all possible continuations of the note sequence, as described by E in the following:

$$H(e_{1.i-1}) = \sum_{e \in E} p(e|e_{1.i-1}) S(e|e_{1.i-1}).$$

In other words, the entropy provides an indication on the uncertainty on the upcoming music event given the preceding context.

IDyOM simulates implicit melodic learning by estimating the probability distribution of each upcoming note. This model can operate on multiple viewpoints, meaning that it can capture the distributions of various properties of music. Here, we focused on the onset time viewpoint. IDyOM generates predictions of upcoming music events based on what

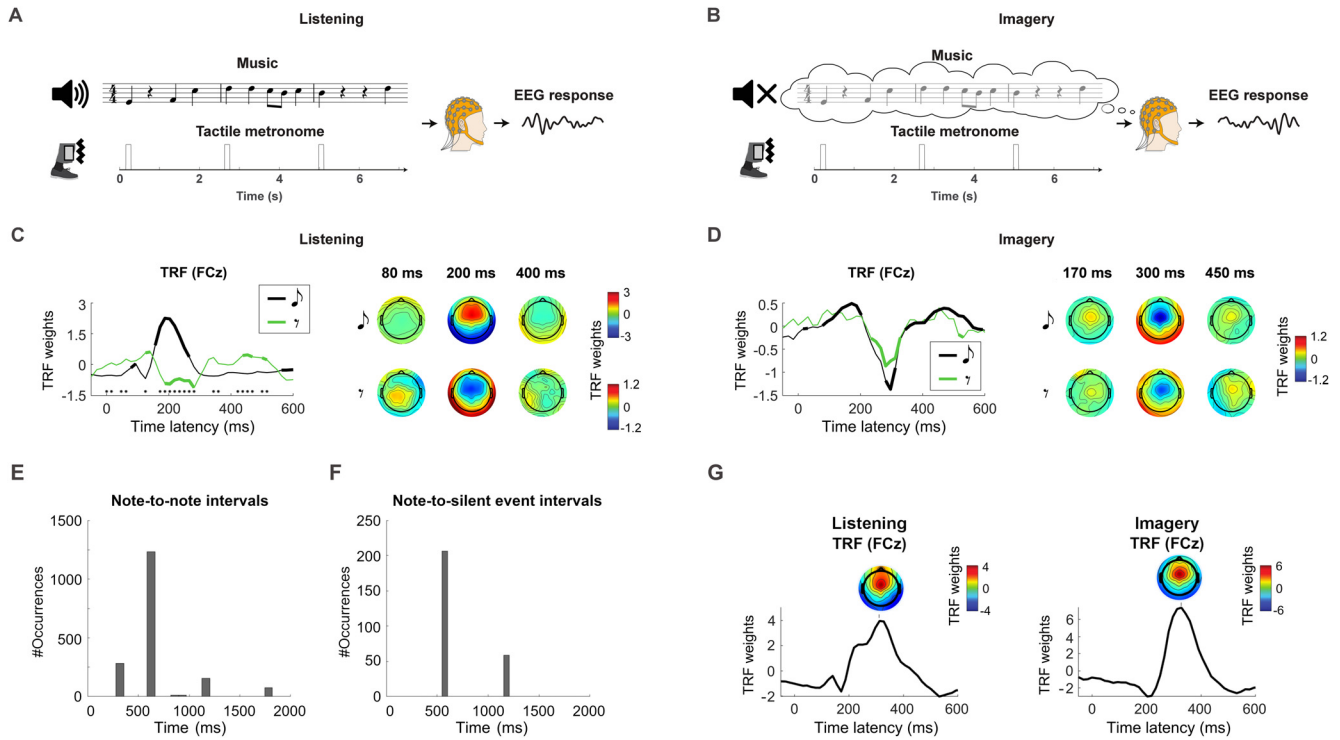

Figure 3. Comparable cortical encoding of music silence and note during imagery. ***A***, ***B***, EEG signals were recorded as participants listened to and imagined piano melodies (experiment 2). A vibrotactile metronome placed on the left ankle allowed for the precise execution of the auditory imagery task. ***C***, TRFs at the channel FCz (left) and topographies of the TRF at selected time latencies (right) are reported for the listening condition. Thick lines indicate TRF weights that are larger than the baseline at latency zero (FDR corrected Wilcoxon signed rank test, $q < 0.01$). Black asterisks indicate significant differences between NT and SIL (FDR corrected Wilcoxon signed-rank test, $q < 0.01$). ***D***, The TRF results are reported for the imagery condition, showing a significant component centered at ~300 ms for both note and silent-events with, as hypothesized, no significant difference between NT and SIL, which had the same polarity in this case. ***E***, ***F***, Overall distribution of time intervals between notes and between silent events and the immediately preceding note in experiment 2. The *y*-axis indicates the number of occurrences for a given bin of time intervals when considering all trials. ***G***, TRFs were fit for the listening and imagery conditions using a univariate stimulus regressor marking the metronome with unit impulses (and zero at all other time points). TRFs are shown at the EEG channel FCz. Topographies depicting the TRF weights at all channels are also shown at the peak of the dominant TRF component.

is learned, allowing the estimation of entropy values for the properties of interest. Each of these features was encoded into time series by using the values of the features to modulate the amplitude of a note onset vector. This resulted in four time series: surprise and entropy of the onset time for notes (S_{NT} and H_{NT}) and silences (S_{SIL} and H_{SIL}).

TRF analysis

A system identification technique was used to compute the channel-specific music-EEG mapping for analyzing the EEG signals from both experiments. This method, here referred to as the TRF (Lalor et al., 2009; Ding et al., 2014), uses a regularized linear regression (Crosse et al., 2016) to estimate a filter that optimally describes how the brain transforms a set of stimulus features into the corresponding neural response (Fig. 2*A*, forward model). Leave-one-out cross-validation (across trials) was used to assess how well the TRF models could predict unseen data while controlling for overfitting. Specifically, we implemented a nested-loop cross-validation, with the inner-loop consisting of a leave-one-out cross-validation where TRF models were fit on the training fold and used to predict the EEG signal of the left-out trial. The purpose of the inner loop was to determine the optimal regularization parameter ($\lambda \in [10^{-9}, 10^{5}]$) by selecting the one maximizing the EEG prediction correlation (averaged across all electrodes and validation trials). The outer loop iterated over each left-out test trial, where the TRF model was fit on all other trials (using the optimal regularization parameter identified with the inner loop), and the quality of the model was quantified by calculating the Pearson's correlation between the preprocessed recorded signal and the prediction of the signal at each scalp electrode.

The interaction between stimulus and recorded brain responses is not instantaneous; in fact, a sound stimulus at time t_0 affects the brain signals for a certain time window $[t_1, t_1+t_{win}]$, with $t_1 \geq 0$ and $t_{win} > 0$. The TRF takes this into account by including multiple time lags between the stimulus and neural signal, providing us with model weights that can be interpreted in both space (scalp topographies) and time (music-EEG latencies). The relative long interonset interval (IOI) between music events (e.g., the most common note duration was 300 ms in experiment 2) could constitute a challenge for the TRF analysis, which may erroneously associate a neural response to a note n to the previous note n-1 because of the intrinsic regularity of music. To overcome this limitation, a broad time lag window of −300 to 900 ms was selected to fit the TRF models, which enabled the regression model to more reliably distinguish the response to the current and neighboring events.

A univariate forward TRF analysis was used to assess the neural response to music notes and silent-events. TRF models were fit for relating NT and SIL with the EEG signal from experiments 1 and 2. Note that note and silent-event TRFs were fit separately. The temporal dynamics of the neural response to music were then inferred from the TRF model weights for latencies that were considered physiologically plausible according to previous work (Freitas et al., 2018; Jagiello et al., 2019; Di Liberto et al., 2020), as shown in Figures 2*B*, 3*C*, and 3*D*. A multivariate TRF analysis was also conducted for experiment 1 by combining NT and SIL, which allowed to assess the neural signature corresponding to silent-events while regressing out the possible impact of the evoked responses to the preceding notes (Fig. 2*E*).

In experiment 2 a multivariate TRF analysis was also used to assess the cortical encoding of melodic surprise for note and silence events separately. Specifically, given either note or silence events, forward TRF models were fit by representing the stimulus as the concatenation of the corresponding (1) onset time vector, (2) entropy time vector, (3) and surprise time vector. Then, the TRF analysis was repeated after shuffling the expectation values (entropy and surprise) in the multivariate regressor. Specifically, a random permutation was applied to shuffle the entropy and surprise values of the events while preserving the onset time.

This allowed for the comparison of the TRF models with shuffled modes with the same dimensionality but with meaningless melodic expectation value sequences (see below, Statistical analyses for additional details on the permutation analysis). The rationale was that the inclusion of melodic expectation information improves the EEG prediction correlations only if the EEG responses to music are modulated by such expectations, a phenomenon that was already shown for notes (Di Liberto et al., 2020; Marion et al., 2021) but not for silences.

Multiway canonical correlation analysis

The TRF analysis has some limitations, such as working under the assumption of time invariance of the neural responses to notes and silent-events. That could be an issue because it is possible that the responses to silence change depending on the position (e.g., two consecutive silences). However, ERP analysis make the same assumption, and the high level of noise in the EEG hampers our ability to study questions on the raw data. We tackled this issue in experiment 2 with multiway canonical correlation analysis (MCCA), a tool that merges EEG data across subjects to improve the signal-to-noise ratio (SNR). MCCA is an extension of canonical correlation analysis (Hotelling, 1936) to the case of multiple (> 2) datasets. Given N multichannel datasets Y_i with size $T \times J_i$, $1 \leq i \leq N$ (time × channels), MCCA finds a linear transform W_i (sizes $J_i \times J_0$, where $J_0 < \min(J_i)_{1 \leq i \leq N}$), which, when applied to the corresponding data matrices, aligns them to common coordinates and reveals shared patterns (de Cheveigné et al., 2019). These patterns can be derived by summing the transformed data matrices as follows: $Y = \sum_{i=1}^{N} Y_i W_i$. The columns of the matrix Y, which are mutually orthogonal, are referred to as summary components (SC; de Cheveigné et al., 2019). The first components are signals that most strongly reflect the shared information across the several input datasets, thus minimizing subject-specific and channel-specific noise. Here, these datasets are EEG responses to the same task for 21 subjects. Note that EEG data were averaged across the 11 repetitions of each musical piece to improve the SNR before running the MCCA analysis.

This technique allows extraction of a consensus EEG signal, which is more reliable than that of any subject. This methodology is a better solution than averaging data across subjects, which in the absence of appropriate coregistration leads to loss of information because of topographical discrepancies. MCCA accommodates such discrepancies without the need for coregistration. Under the assumption that the EEG responses to music and music imagery share a similar time course within a homogeneous group of young adults, the MCCA procedure allows us to extract such common cortical signals from other more variable aspects of the EEG signals, such as subject-specific noise. For this reason, our analysis focuses on the first N_{SC} summary components, which we can consider as spanning the most reliable EEG response to music and music imagery. N_{SC} was set to the number of components with the largest (fifth percentile) correlation with the original EEG data ($N_{SC} = 10$ and $N_{SC} = 8$ for the listening and imagery conditions, respectively). Denoised EEG data were then calculated by inverting the MCCA mapping and projecting the N_{SC} summary components back to the subject-specific EEG channel space. The latter procedure allowed us to study the MCCA results in the same space of the TRF results (EEG channels) and to assess the robustness of the result across participants.

This last step was executed twice. First, denoised EEG data were calculated by using only the first summary component, which intuitively represent the strongest and most correlated response across subjects—the sensory response. A second denoised EEG dataset was calculated based on the remaining N_{SC}-1 components, which were expected to include the residual sensory response but, importantly, to encode the neural prediction signal. A time-locked average analysis was conducted on the two resulting signals, allowing us to derive an average response for notes and silent-events for each of the signals (first component and the combination of the remaining N_{SC}-1 components) and for each condition (listening and imagery). Baseline correction was not applied for the time-locked average as the MCCA procedure should have substantially reduced subject-specific noise (e.g., temporal drifts). Thus, we were interested in assessing the exact average signals corresponding with notes and silent-events, including possible nonzero activity before the event. This also allowed us to more clearly visualize the potential impact of previous notes to the average signal corresponding with notes or silent-events. Different from the TRF analysis, the time intervals corresponding to the metronome response were included in the MCCA procedure, allowing us to extract components related to the corresponding sensory response.

This analysis was conducted on EEG data that was filtered between 1 and 30 Hz. We also ran the procedure by including frequencies down to 0.1 Hz. However, the separation between sensory and prediction components was not as clear cut as in the 1–30 Hz case as the sensory response contributed to the first several components.

Statistical analyses

Consistent statistical procedures were applied to the datasets from the two experiments.

Linear mixed-model analyses were performed when testing for significant effects in case of multiple factors over multiple groups. This statistical test was conducted when studying the TRF results in Figure 3 and the MCCA results in Figure 4 to assess effects of event type (notes and silent-events) and condition (listening and imagery).

Pairwise comparisons were assessed via the (nonparametric) Wilcoxon signed-rank test. Correction for multiple comparisons was applied where necessary via the false discovery rate (FDR) approach. In that case, the q value is reported, that is, FDR-adjusted p value. This FDR-corrected Wilcoxon analysis was used when testing the TRF weights for significance in experiment 1 by comparing each data point of the TRF global field power with a baseline at latency zero. The same FDR-corrected analysis was also run when conducting a *post hoc* analysis on the TRF weights in experiment 2 (Fig. 3), again with a baseline at latency zero, and in the lateralization analysis in Figure 5.

A permutation procedure was used to test for a significant neural encoding of melodic expectations in experiment 2 (Fig. 5). That procedure consisted of rerunning 100 times per participant the forward TRF procedure, each time after random shuffling of the expectation values, while preserving the timing information (see above, TRF analysis for further details on the shuffling procedure). A null distribution of the mean EEG prediction correlation across participants was estimated with bootstrap resampling to assess whether melodic expectations improved the EEG prediction correlations. The null distribution was composed of $N =$ 10,000 data points, each derived by selecting a random data point per subject among the 100 shuffles, averaging the corresponding EEG prediction correlations across participants, and repeating the procedure 10,000 times. The uncorrected p values are reported in this case as several p values were smaller than the sensitivity of the test ($p < 10^{-4}$). The 100 data points per participants were also used to estimate a null distribution to assess significance for individual participants. Note that the analyses for both the group subject level and individual subject level were conducted after averaging the EEG prediction correlations across all electrodes.

Data Availability

The music listening Bach dataset is available to download via Dryad at (https://doi.org/10.5061/dryad.dbrv15f0j). The music listening and imagery dataset will be available to download via Dryad. The TRF analysis was conducted using the freely available mTRF-toolbox version 2.0, which can be downloaded from https://sourceforge.net/projects/aespa/. The MCCA analysis was conducted using the freely available MATLAB toolbox NoiseTools, which can be downloaded from http://audition.ens.fr/adc/NoiseTools/. Melodic expectation information was calculated with the Lisp implementation of the IDyOM, which is freely available for download at https://code.soundsoftware.ac.uk/projects/idyom-project.

Results

Neural data were recorded from participants as they alternately performed a music listening (experiments 1 and 2) and a music imagery task (experiment 2) based on monophonic piano

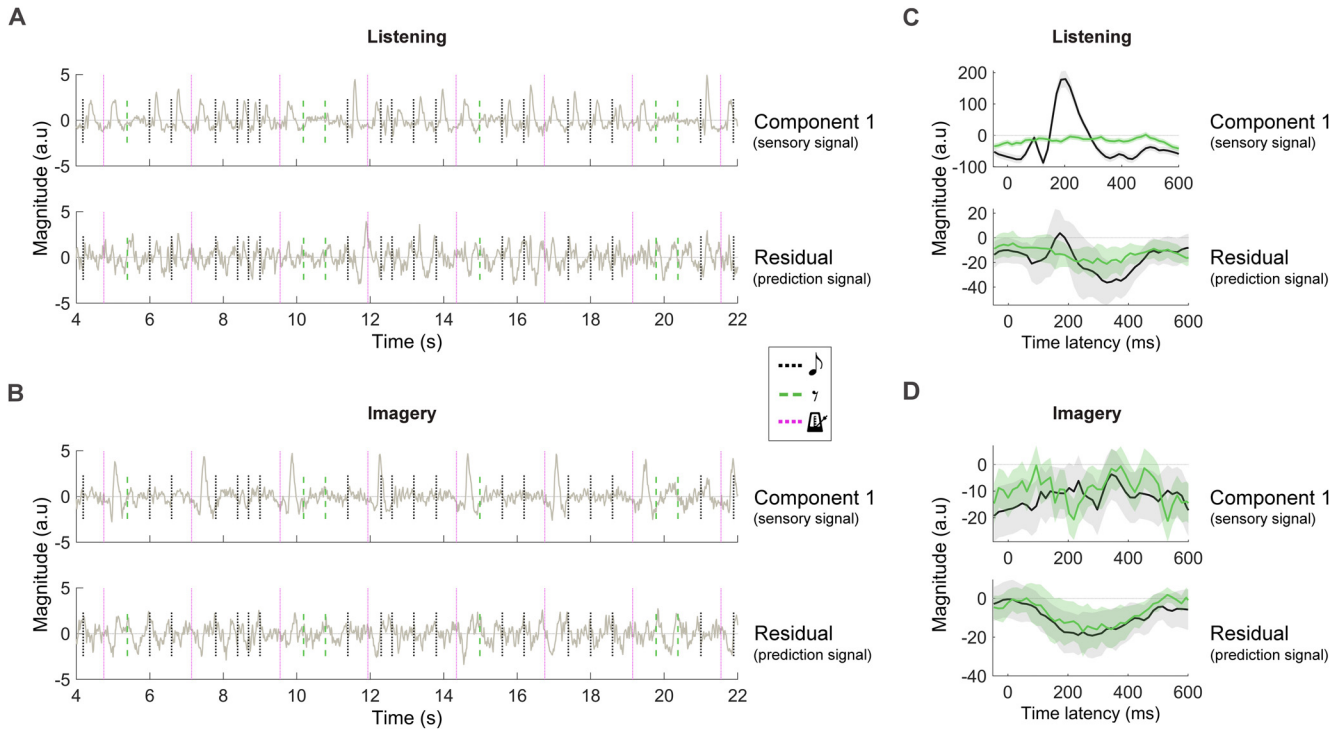

Figure 4. Disentangling sensory and prediction neural signals with unsupervised correlation analysis. MCCA was used on all EEG data to identify components of the EEG signal that were consistent across subjects. N_{SC} SCs with the largest intersubject correlation were preserved. The first SC represents the EEG response that is the most correlated signal across subjects. Here, we hypothesized the first SC and the residual N_{SC}-1 SCs to capture sensory and prediction cortical signals, respectively. ***A***, ***B***, The first SC (top) and the residual N_{SC}-1 SCs (bottom) were back projected onto each participant's EEG channel space for each condition. The average signals at the EEG channel FCz were shown for a selected portion of Melody 4 (olive green lines). Vertical lines mark music events, notes (black dotted lines), silent-events (green dashed lines), and vibrotactile metronome onset (purple dotted lines). Note that sensory responses could exist only for note and metronome in the listening condition and for metronome only in the imagery condition. ***C***, ***D***, First SC (top) and the residual N_{SC}-1 SCs (bottom) at the EEG channel FCz after time-locked averaging to note and silent-event onsets. Shaded areas indicate the 95% confidence interval calculated across participants.

melodies from Bach. A computational model of melodic structure based on Markov chains (Pearce, 2005) was used to identify silent-events, that is, silent instants where a note could have plausibly occurred (see above, Materials and Methods). Our analyses aimed at testing the hypothesis that an endogenous prediction signal emerges in correspondence to silent-events. We parametrized the onset times of notes and silent-events in univariate vectors (NT and SIL, respectively) and related them with the neural data by means of three distinct analysis procedures. In the listening task, the S response, which was present in NT but not SIL, was anticipated to account for most of the variance of the EEG responses to melodies. The residual nonsensory response was instead hypothesized to reflect top-down neural P signals. According to the predictive processing framework, P was expected to be measured in combination with the sensory response in correspondence of notes in the listening condition (δ_{sur} = S−P) and in isolation in correspondence to notes in the imagery condition and silent-events in both conditions (S = 0, δ_{sur} = −P).

Experiment 1: Robust cortical response to silence during music listening

In the first EEG experiment, 20 healthy participants were instructed to listen to 10 monophonic piano excerpts from Bach sonatas and partitas, each repeated three times and played in random order. The cortical responses to music were assessed by means of a multivariate linear regression framework known as the TRF, which takes into account the interactions and overlap between a succession of notes. Given a property of interest of a sensory stimulus encoded in a time vector, the TRF estimates an optimal linear transformation of those vectors that minimizes the EEG prediction error (Fig. 2*A*). The TRF weights can then be interpreted to assess the spatiotemporal dynamics of the underlying neural system.

First, the cortical response to music notes was assessed by calculating the TRF between a time vector marking note onsets with value 1 (NT) and the corresponding EEG signal (1–30 Hz). Then, the same procedure was repeated when considering the onset time of silent-events (Fig. 2*A*). The global field power of TRF_{NT} indicated that three components centered at ~80, 200, and 400 ms were significantly larger that the baseline at lag 0 ms of sound EEG latency (FDR-corrected Wilcoxon tests, $q < 0.001$; NT, significant effects for the time-latencies in the windows of 62.5–250 and 297–516 ms). Instead, only two significant components emerged for TRF_{SIL} at 200 and 400 ms. The regression weights for TRF_{NT} and TRF_{SIL} are shown in Figure 2*B* for a representative electrode (FCz), with the corresponding topographies for all electrodes. Interestingly, note and silent-event responses showed inverse polarity, showing a large negative correlation between the two curves ($r = -0.946$, $p = 4.0 * 10^{-30}$) and leading to significant differences for the three components at 80, 200, and 400 ms (FDR-corrected Wilcoxon tests, $q < 0.001$; SIL, significant effects for time latencies in the windows of 125–282 and 359–484 ms).

The large difference between the neural responses to notes and silent-events is likely because of the absence of auditory stimulation for music silence. As such, TRF_{SIL} was expected to reflect the effect of top-down predictions, which could include the prediction signal itself as well as the update of internal priors on the upcoming music event after detecting a silence. Indeed,

the present design comes with a potential confound: TRF_{SIL} could be capturing an average late response to a previous note. Although this risk is minimized by our choice of 10 music stimuli with various tempo, the majority of silent-events occurred <300 ms after a note (Fig. 2*C*,*D*). To assess the likely interaction between silent-events and the preceding notes, we conducted a multivariate forward TRF analysis where both the NT and SIL regressors were used to predict the EEG signal simultaneously. In this context, the NT vector could be seen as a nuisance regressor when studying TRF_{SIL} and vice versa. The result of this analysis (Fig. 2*E*) indicates that the inclusion of NT as a nuisance regressor does not change the main TRF result (polarity inversion between TRF_{NT} and TRF_{SIL}, with a large negative correlation between the two curves: $r = -0.616$, $p = 1.6 * 10^{-7}$). Furthermore, the dynamics of TRF_{SIL} did not change compared with the univariate analysis (Pearson's $r = 0.95$, $p = 3.3 * 10^{-31}$ between the TRF_{SIL} curves in the univariate and multivariate TRF analyses), despite a reduction in power over time latencies (Wilcoxon tests, $p = 3.3 * 10^{-11}$), which reflects the expected smaller magnitude silent-event neural signals compared with evoked responses to notes, an effect that is magnified in the multivariate model as the two neural responses are assessed simultaneously.

We designed a second experiment aiming at overcoming the limitations with experiment 1. The following section describes that experiment 2, whose novel design based on musical imagery enables the isolation of endogenous neural signatures of both notes and silent-events.

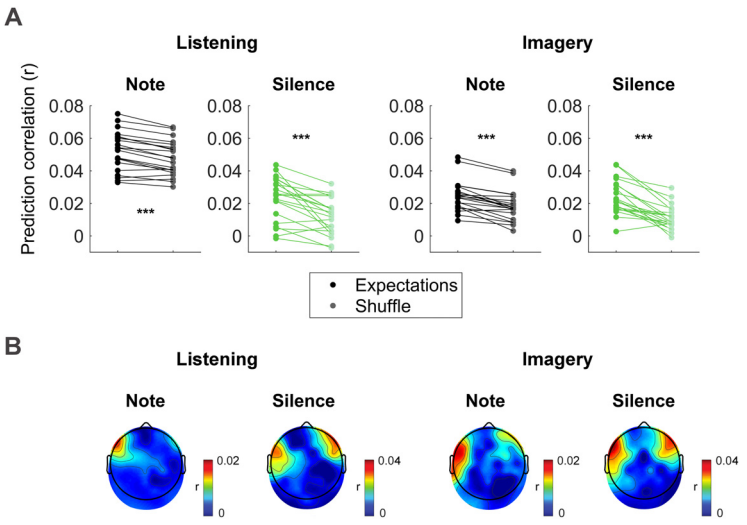

Figure 5. Notes and silence expectation encoding in low-frequency EEG. A multivariate TRF analysis was conducted to identify the linear transformation that best predicts low-frequency EEG data (0.1–30 Hz) based on a three-dimensional stimulus representation indicating for either note or silent-events the event onset time, entropy at that position, and surprise of that event. ***A***, EEG prediction correlations (r) of the TRF using the note or silence expectation values estimated with IDyOM are compared with a null model, where the EEG prediction correlations were obtained with a TRF that was fit after a random shuffling of the expectation values (event onset times were preserved). Results averaged across all electrodes are reported for both listening and imagery conditions. Each dot indicates the result for a single subject. Significant differences were measured for notes and silent-events in both conditions (permutation test, $^{***}p < 10^{-4}$). ***B***, Topographical maps indicating the EEG prediction correlation increase (expectation minus null model) at each EEG channels.

Experiment 2: Cortical encoding of music silences during listening and imagery tasks

A second EEG experiment was conducted on 21 expert musicians. In the listening condition (Fig. 3*A*), participants were presented with four ~35 s piano melodies from Bach chorales. In the imagery condition (Fig. 3*B*), participants were instructed to imagine hearing the same melodies. Each piece was presented and imagined 11 times, for a total of 88 trials with random order. A vibrotactile metronome placed on participants' left ankle marked the beginning of 100 bpm measures (every 2.4 s) in both conditions, allowing participants to perform the auditory imagery task with high temporal precision. Therefore, the neural signal was expected to reflect sensory responses to auditory and tactile sensory inputs for the listening condition and to the tactile input only for imagery. Neural data within 500 ms from the metronome input were excluded from the TRF analyses that follow to ensure that the results do not reflect tactile responses.

First, we replicated the result from experiment 1 by running a forward TRF analysis on the EEG data (1–30 Hz) for the listening condition. The TRF weights showed spatiotemporal dynamics consistent with the previous result, with inverse polarities for NT and SIL (Fig. 3*C*). Then, we ran the same TRF procedure on the auditory imagery condition. Although the investigation was conducted in a manner consistent with the previous experiment, the analyses largely focused on the EEG channel FCz, where the main effect (a polarity inversion in the listening condition) was expected based on the results from experiment 1. A linear mixed-model analysis indicated significant effects of condition (listening vs imagery) and regressor (notes vs silent-events) on the TRF weights, with a significant interaction effect between condition and regressor (The dependent variable was the average magnitude of the TRF component at FCz for the latencies 250 ± 100 ms, condition and regressor were the independent variables, and subjects a random intercept; effect of regressor: estimate = −2538, *tStat* = −12.4, $p = 2.4 * 10^{-20}$; effect of condition: estimate = −2746, $t = -11.4$, $p = 1.7 * 10^{-18}$; interaction effect: estimate = 1307, $t = 10.1$, $p = 6.3 * 10^{-16}$.). A *post hoc* analysis on the individual TRFs indicated components larger than the baseline at latency zero for all conditions (FDR-corrected Wilcoxon tests, $q < 0.01$; see above, Materials and Methods). Figure 3*C* shows significant TRF components at FCz in the listening condition. TRF traces for notes and silent-events were negatively correlated ($r_{NT\_LIST,SIL\_LIST} = -0.60$, $p = 7.0 * 10^{-5}$), thus replicating the result from experiment 1. As we showed in part 1 of this study, the result in Figure 3*D* indicates (Marion et al., 2021) robust neural correlates of auditory imagery in correspondence to notes. Crucially, the EEG dynamics of auditory imagery corresponding to silent-events showed shape and latencies comparable to those measured for imagery of notes ($r_{NT\_IMAG,SIL\_IMAG} = 0.89$, $p < 1.2 * 10^{-13}$), with the same polarity measured for silent-events in the listening condition ($r_{SIL\_LIST,SIL\_IMAG} = 0.57$, $p = 2.1 * 10^{-4}$, $r_{SIL\_LIST,NT\_IMAG} = 0.52$, $p = 7.4 * 10^{-4}$). Conversely, TRF_{NT} in the listening condition had inverse polarity, which was consistent with the polarity of tactile responses, that is, the only other sensory response in the EEG data (Fig. 3*G*, TRF result for the metronome-only vector).

The result in Figure 3*D* indicates that the inverted cortical polarity measured for TRF_{NT} and TRF_{SIL} during music listening (Figs. 2*B*, 3*C*) depends on the presence or absence of a sensory stimulus, respectively, rather than a different encoding of notes and silent-events per se. In fact, that difference was not present in the imagery condition, where there was no auditory stimulation. This result is in line with a predictive processing account of

auditory perception whereby the brain constantly attempts to predict sensory signals (Fig. 1). The analyses that follow aim to provide further support to this result by (1) disentangling sensory and prediction signals in both the listening and imagery conditions with a methodology that different from the TRF does not use explicit knowledge on the position of notes and silent-events and (2) assessing whether the prediction signal encodes the precise melodic expectation values as estimated by a computational model of musical structure.

Disentangling neural sensory responses and neural prediction signal

The TRF analysis showed robust note and silent-event encoding in both listening and imagery conditions. However, the TRF analysis did not account for differences between responses to individual events. For example, neural responses change with the listener's expectation of a note based on the proximal music context (Omigie et al., 2013; Di Liberto et al., 2020). The following investigates the neural signature of individual music events across the time domain of a musical piece.

Investigating the cortical processing of individual music notes requires an approach that is effective despite the low SNR of EEG recordings. The TRF procedure described previously summarizes information across the time domain, providing a summary neural trace for each participant representing the typical response to a note or a silent-event. However, that approach does not provide us with a view at the level of individual events (notes and silences). To do so, we used MCCA, an approach that denoises the EEG data by preserving components of the signal that are consistent across participants.

An MCCA analysis was run on EEG data from all participants simultaneously, preserving the first N_{SC} SCs with largest inter-subject correlation (see above, Materials and Methods). This approach enables the investigation of neural data in the original EEG channel with a remarkably high SNR, allowing us to assess the neural signature of each individual event in a melody. SC_1 was expected to capture the sensory response, which is likely the strongest and most consistent signal across participants. As we had hypothesized, SC_1 showed strong neural activation corresponding to sensory events, that is, notes and metronome in the listening condition and metronome only in the imagery condition (Fig. 1, hypothesis; Fig. 4, result). That result was visible both at the level of individual music events (Fig. 4*A*,*B*) and on the time-locked average signals (Fig. 4*C*,*D*). Next, the first sensory component (SC_1) was removed from the EEG data to study the residual N_{SC}–1 components, which were expected to capture neural predictions and, therefore, to be active in correspondence of both notes and silent-events, as depicted in Figure 1. The result in Figure 4 confirms that hypothesis by showing neural activation for all music events, with negative components corresponding to both notes and silent-events between ~200 and 400 ms in the imagery conditions. A linear mixed-model analysis confirmed such observations. Significant effects of condition (listening vs imagery) and event-type (notes vs silent-events) were measured on the time-locked averages for SC_1, with a significant interaction effect between condition and event-type (The dependent variable was the average magnitude of the time-locked average component at FCz for the latencies 250 ± 100 ms, condition and event type were the independent variables, and subjects a random intercept; effect of event type: estimate = −6.1, *tStat* = −6.6, $p = 4.5 * 10^{-9}$; effect of condition: estimate = −4.3, $t = -4.7$, $p = 9.5 * 10^{-6}$; interaction effect: estimate = 3.0, $t = 5.1$, $p = 2.1 * 10^{-6}$.). This result is in line with the interpretation of the first component as a sensory response signal, which is present only for notes in the listening condition. Significant effects were also measured on the residual N_{SC}–1 components for condition but not event type nor the interaction between the two (effect of event type: estimate = 3.6, *tStat* = 1.9, $p = 0.058$; effect of condition: estimate = 4.6, $t = 2.5$, $p = 0.016$; interaction effect: estimate = −1.9, $t = -1.6$, $p = 0.11$), which is in line with the interpretation of the residual N_{SC}–1 components as a prediction signal. Overall, this result is consistent with our initial hypothesis in Figure 1.

Cortical encoding of silence expectations during music listening and imagery

Recent studies indicated that low-frequency neural signals encode melodic expectations when participants listen to monophonic music (Omigie et al., 2013; Di Liberto et al., 2020). Specifically, melodic expectations modulate the magnitude of the auditory responses, with larger neural responses for less expected events. In line with those results and with the hypothesis that cortical responses to music reflect a combination of sensory and prediction signals (Fig. 1), we anticipated EEG responses to notes and silent-events to be modulated by melodic expectations during both listening and imagery conditions. To test that, we first estimated the expectation of a note with IDyOM (Pearce, 2005), the model of melodic structure based on variable-order Markov chains, which was also used to identify the silent-events. Expectation values were calculated from the music score based on both a long-term model of Western music and short-term proximal information on the current piece. As a result, IDyOM provided us with measures of surprise and the Shannon entropy of the onset time of each upcoming note and silent-event. Surprise informs us how unexpected a note (or a silent-event) was at a given time point, whereas the entropy indicates the uncertainty at a particular position in a melody before the music note is observed. Each of these features was encoded into time series by using the values of the features to modulate the amplitude of note and silent-event onset vectors. This resulted in four time series: surprise for notes (S_{NT}) and silent-events (S_{SIL}), and entropy for notes (H_{NT}) and silent-events (H_{SIL}). We then called EXP_{NT} and EXP_{SIL} the concatenation of the surprise, entropy, and onset vectors for notes and silent-events, respectively. Note that EXP_{NT} and EXP_{SIL} were calculated by using timing but not pitch information as silent-events do not have a pitch value.

Forward TRF models were fit to assess the coupling between low-frequency EEG (0.1–30 Hz) and the onset time expectation vectors. Shuffled versions of the expectation vectors ($N = 100$ per subject), with surprise and entropy values randomly permuted while preserving the temporal information of the event onsets, were used as a baseline for the assessment of the expectation EEG encoding. Both EXP_{NT} and EXP_{SIL} could predict the EEG better than the shuffled versions in both the listening and imagery conditions [EEG prediction correlation was averaged across all EEG channels; the mean across subjects was compared with a bootstrap resampling distribution of the mean across subjects derived from the shuffled data; $N = 10{,}000$; $p < 10^{-4}$ for notes and silent-events in both conditions (see above, Materials and Methods; Fig. 5*A*).]. A significant EEG encoding of expectations was also measured at the individual subject level, with 12/21 and 10/21 subjects above chance level in the listening condition for note and silent-event TRFs, respectively, and 10/21 and 17/21 subjects above chance level in the imagery condition (one-tailed permutation test, $N = 100$ permutations per subject per condition, $q < 0.05$, FDR correction). Although this effect of expectation was assessed on the average of all EEG channels,

Figure 5*B* shows the topographical distribution of that effect (contrast between EEG prediction accuracies for expectation and the 95th percentile of the shuffles). Similar but weaker effects were measured for EEGs filtered between 1 and 30 Hz for all conditions but silent-events in the imagined condition (EEG prediction correlation values were averaged across all channels; the mean value across subjects was compared with a bootstrap resampling distribution of the mean across subjects derived from the shuffled data; $N = 10{,}000$; NT listening, $p < 10^{-4}$; SIL listening, $p = 0.021$; NT imagery, $p = 0.009$; SIL imagery, $p = 0.541$.).

These results indicate a fine-grained encoding of melodic expectations in the cortical signals corresponding to music listening and imagery. We also tested whether the effect of onset time expectations on the EEG prediction increase showed significant lateralization. We found a weak left-lateralization effect for notes in the listening condition, which, however, did not survive correction for multiple comparisons (FDR-corrected Wilcoxon test, $q = 0.100$).

Discussion

Predictive processing explains rhythmic and melodic perception as a continuous attempt of our brain to anticipate the timing and identity of upcoming music events. Although previous research investigated such predictive mechanisms indirectly by measuring how expectation modulates sensory responses, this study used neural measurements of music processing in the absence of a sensory input. In particular, we combined for the first time low-frequency EEG measurements corresponding to silent music events during music listening with EEG signals recorded during musical imagery, both in the context of natural melodies. In doing so, we could demonstrate that robust neural activity consistent with prediction error signals emerge during the meaningful silence of music and that such neural activity is modulated by the strength of the music expectations. We propose a unifying perspective on auditory predictions, where endogenous auditory predictions have a central role in music perception both during listening and imagery.

EEG encoding of music-silence reveals neural auditory predictions

Existing computational models of music structure, such as IDyOM, generally consider silences as time intervals without notes [stimulus onset asynchronies (SOA) or IOI; Pearce, 2005]. However, melodies contain silent instants where a note could have plausibly occurred. The present study demonstrates that the human brain encodes these music silent-events, suggesting that the physiological validity of those models can be augmented via an explicit account of silent-events, rather than just an implicit encoding of that same information as in IDyOM. The finding that musical imagery elicits robust neural activity (Tibo et al., 2020; Marion et al., 2021) laid the foundations of the present study, providing us with an experiment that allows us to discern endogenous neural processes from exogenous auditory perception mechanisms. Our results are summarized by the model in Figure 6, in line with the predictive processing framework. The neural encoding of sound and silence corresponds to S−P and, as such, to −P when there is no stimulus (S = 0), that is, silent-events during listening and imagery and notes during imagery. This result emerged both when using forward TRFs (Crosse et al., 2016) and MCCA (de Cheveigné et al., 2019), two methods with different assumptions and rationale. One crucial difference between the two is that TRFs describe the neural responses for an individual subject with an impulse response, one for each

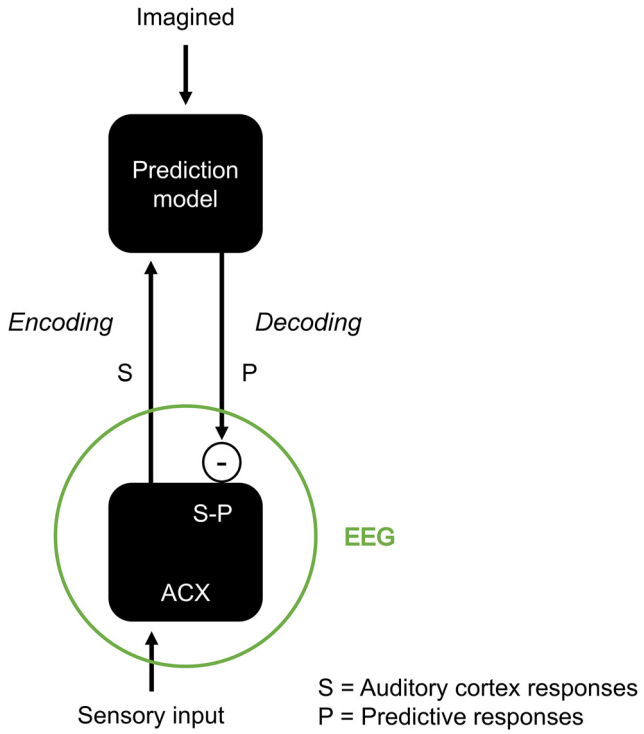

Figure 6. Computational model for how predictions influence neural signals corresponding to auditory listening and imagery. Auditory inputs elicit bottom-up S responses through the auditory cortex (ACX). A prediction model generates a top-down P signal that is more similar to S for more predictable sensory events. That prediction is compared with S, producing an error signal $\delta_{sur} = S - P$. The EEG response is hypothesized to capture a combination of δ_{sur} and S, meaning that some level of EEG activation is expected even when S is fully predictable (Margulis, 2014). When a sound is imagined, S = 0, and therefore $\delta_{sur} = -P$, as for our hypothesis in Figure 1.

stimulus feature set (notes vs silence). Instead, MCCA does not require any explicit knowledge of the timing and identity of notes and silent-events. This unsupervised approach combined the data from multiple subjects to extract neural components that were sensitive to individual events (note or silence) within a continuous music piece, with a remarkable signal-to-noise ratio. Crucially, the two methods converged to consistent results, revealing that silent-events correspond to robust neural responses and that similar neural signals emerge for imagined notes and silent-events. This internally generated music of silence is in line with a continuous attempt of the brain to predict upcoming plausible notes. Altogether, this study provides direct evidence for the duality of sensory and prediction signals suggested by the predictive processing framework.

Our results suggest that both listening and auditory imagery entail the transformation of external or imagined sounds by an internal predictive model that encodes our conceptions and expectations of the sound, which is then compared with the sensory stimulus, if present. The finding, which is captured by the model in Figure 6, is in line with previous fMRI and PET results on auditory imagery (Halpern and Zatorre, 1999; Meister et al., 2004; Kraemer et al., 2005; Zhang et al., 2017). In fact, such studies showed robust neural activation in correspondence with auditory imagery as we measured here with EEG. Crucially, our results linked the neural activation for auditory imagery to a general predictive mechanism that applies to both listening and imagery. Specifically, the model in Figure 6 explains both imagery and silence activations as the result of the subtraction of sensory responses and prediction signals, leading to a change in response

polarity when the sensory input is absent. Indeed, further work with multiple technologies (e.g., fMRI, EEG, electrocorticography) is needed to conclusively link our finding with studies based on hemodynamic measurements and test our model. One challenge is to clarify what exactly each neural measurement can capture within that model. EEG recordings provide us with macroscopic measurements that are likely to include a variety of neural components. Although the evidence points to a strong sensitivity to prediction errors (or surprise), there may be additional components that encode S and P separately.

Although the music of silence allows to clearly separate the neural prediction signal from sensory responses, technologies with higher spatial resolution may be able to uncover more precise details on the neural implementation of this predictive mechanism. One unsolved question regards the possibility that prediction processes could be at the core of the ability to perform auditory imagery. Based on our finding that prediction processes explain a significant portion of EEG variance during auditory imagery and that imagined notes and silent-events corresponded to similar neural activation, one possibility is that auditory imagery may rely on the same endogenous prediction mechanisms that are engaged during auditory listening rather than involving a separate imagery process. Finally, additional research should be conducted to investigate possible links between our model and beat perception. In fact, the present design was optimized for the imagery task, thus working with relatively simple melodies. Experiments with a broader set of music stimuli are needed to tackle that question, for example, by using syncopated music stimuli, which would allow for a more distinct separation of beat and notes (Tal et al., 2017).

Silence neural signals are graded by expectations

The TRF analysis in Figure 5 confirmed the hypothesis that low-frequency EEG responses to naturalistic music encode melodic expectation in correspondence with prospectively predictable silent-events. The responses to silent-events were shown to covary with the expectation strengths, which were drawn from a note onset time statistical model (Pearce, 2005), as it was previously shown for music notes (Omigie et al., 2013; Di Liberto et al., 2020). These results are in line and go beyond previous measurements of the neural responses to sensory omissions that focused on scenarios where strong expectations on the upcoming occurrence of a stimulus were built artificially [missing stimulus potentials (MSPs); Bendixen et al., 2009)]. Mismatch negativity responses (MMN) to omitted tones were measured for SOAs up to 150 ms (Yabe et al., 1997), whereas studies with longer asynchronies, closer to those of the present study, were shown to elicit MSPs with a modality-specific (auditory) negativity at ~230 ms and a modality-independent (both auditory and visual) positivity at 465 ms (Simson et al., 1976; Joutsiniemi and Hari, 1989). Silent-events in melodies differ from omissions in that they have a much lower probability of corresponding to a sound. Furthermore, omission cannot be predicted; although the participants of experiment 2 were pre-exposed to the four melodies, silent-events were not unexpected per se. In other words, the participants were certainly not surprised in the traditional sense when they encountered a silent-event as they had heard the melody before. Instead, our results are different from the unexpectedness investigated with sensory omission paradigms as they reveal prediction errors related to the processing of melodic structure based on the melody statistics.

Further work should be conducted to directly explain the overlap and interaction of the two phenomena. Our finding contributes to that question by suggesting a unifying view linking MSP (omission response), expectation modulation (EM) of sensory responses, and auditory imagery using naturalistic music listening. Our results suggest that the MSP negativity and EM are results of the same prediction process. In addition to providing new direct evidence on the neural substrate of MSPs, we show that such responses can be measured when the music is internally generated (imagery). This result is in line with a view of the auditory system where predictions are simultaneously computed at multiple time scales (e.g., hierarchical predictive coding) and, crucially, where local expectations (at short time scales) are performed by our brain even in presence of exact prior knowledge of the upcoming stimulus (e.g., repetition of a song, production or imagination of a song). In fact, the TRF analysis in Figure 5 indicates a robust encoding of melodic expectations, although the stimuli were precisely known by the participants (Only four repeated stimuli were presented, and participants were exposed to the pieces before the start of the EEG experiment.).

We contend that the present finding has implications for computational models of sensory perception. For example, neural signals have been modeled by focusing on evoked responses (Ferezou and Deneux, 2017; Doelling et al., 2019), thus describing the neural signal as a sum of fixed-latency sensory evoked responses while generally ignoring prediction processes. Instead, as highlighted by this study, prediction signals emerge in correspondence to both notes and silent-events in the neural signal. As in Figure 6, evoked-response models could be extended by including such prediction mechanisms both in the presence and absence of a sensory event. The resulting model would describe the S and P duality and would explain the neural responses to music silence that were measured in the present study. We conclude that our brain considers silent-events as temporally precise and information-rich events that provide our brain with valuable information (i.e., that a note was not present at a particular plausible time point) contributing to the subsequent predictions. Our results may reflect a general property of sensory perception and, as such, we expect similar neural responses to emerge during meaningful silences in other auditory stimuli such as speech. Specifically, expectation signals similar to the predictive melodic expectations in music sequences have been demonstrated in the neural responses to phoneme sequences, the fundamental units of speech (Brodbeck et al., 2018; Di Liberto et al., 2019). Therefore, we anticipate that future studies may reveal predictive responses that closely resemble those we identified in music silences but that would reflect the linguistic model of the listener, confirming that the findings of the current study are indicative of general auditory perception mechanisms.

In summary, the present study shows robust neural signatures of music silence, suggesting that silent-events have great importance in the neural encoding of music. Furthermore, we provide evidence that the encoding of silent-events reflects a neural prediction signal with results that are in line with the predictive processing framework.

References

Bendixen A, Schröger E, Winkler I (2009) I heard that coming: event-related potential evidence for stimulus-driven prediction in the auditory system. J Neurosci 29:8447–8451.

Brodbeck C, Hong LE, Simon JZ (2018) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28:3976–3983.e5.

Chennu S, Noreika V, Gueorguiev D, Shtyrov Y, Bekinschtein TA, Henson R (2016) Silent expectations: dynamic causal modeling of cortical prediction and attention to sounds that weren't. J Neurosci 36:8305–8316.

Clark A (2016) Surfing uncertainty: prediction, action, and the embodied mind. Oxford, England: Oxford UP.

Crosse MJ, Di Liberto GM, Bednar A, 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.

de Cheveigné A, Di Liberto GM, Arzounian D, Wong DDE, Hjortkjær J, Fuglsang S, Parra LC (2019) Multiway canonical correlation analysis of brain data. Neuroimage 186:728–740.

den Ouden HEM, Kok P, de Lange FP (2012) How prediction errors shape perception, attention, and motivation. Front Psychol 3:548.

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

Di Liberto GM, Wong D, Melnik GA, de CA (2019) Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage 196:237–247.

Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88:41–46.

Doelling KB, Florencia Assaneo M, Bevilacqua D, Pesaran B, Poeppel D (2019) An oscillator model better predicts cortical entrainment to music. Proc Natl Acad Sci U S A 116:10113–10121.

Ferezou I, Deneux T (2017) Review: how do spontaneous and sensory-evoked activities interact? Neurophotonics 4:031221. 031221.

Freitas C, Manzato E, Burini A, Taylor MJ, Lerch JP, Anagnostou E (2018) Neural correlates of familiarity in music listening: a systematic review and a neuroimaging meta-analysis. Front Neurosci 12:686.

Friederici AD, Pfeifer E, Hahne A (1993) Event-related brain potentials during natural speech processing: effects of semantic, morphological and syntactic violations. Brain Res Cogn Brain Res 1:183–192.

Grisoni L, Mohr B, Pulvermüller F (2019) Prediction mechanisms in motor and auditory areas and their role in sound perception and language understanding. Neuroimage 199:206–216.

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.

Heeger DJ (2017) Theory of cortical function. Proc Natl Acad Sci U S A 114:1773–1782.

Heilbron M, Chait M (2018) Great expectations: is there evidence for predictive coding in auditory cortex? Neuroscience 389:54–73.

Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377.

Jagiello R, Pomper U, Yoneya M, Zhao S, Chait M (2019) Rapid brain responses to familiar vs. unfamiliar music—an EEG and pupillometry study. Sci Rep 9:15570.

Joutsiniemi S-L, Hari R (1989) Omissions of auditory stimuli may activate frontal cortex. Eur J Neurosci 1:524–528.

Koelsch S, Vuust P, Friston K (2019) Predictive processes and the peculiar case of music. Trends Cogn Sci 23:63–77.

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

Kutas M, Hillyard SA (1980) Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207:203–205.

Kutas M, Hillyard SA (1984) Brain potentials during reading reflect word expectancy and semantic association. Nature 307:161–163.

Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol 62:621–647.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol 102:349–359.

MacKay D (2003) Information theory, inference and learning algorithms. Cambridge, England: Cambridge UP.

Margulis EH (2014) On repeat: how music plays the mind. New York, NY: Oxford UP.

Marion G, Liberto D, Shamma GM (2021) The music of silence. part 1: responses to musical imagery accurately encode melodic expectations and acoustics. J Neurosci, In press.

Mars RB, Debener S, Gladwin TE, Harrison LM, Haggard P, Rothwell JC, Bestmann S (2008) Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J Neurosci 28:12539–12545.

Meister IG, 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.

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

Pantev C, Roberts LE, Schulz M, Engelien A, Ross B (2001) Timbre-specific enhancement of auditory cortical representations in musicians. Neuroreport 12:169–174.

Pearce MT (2005) The construction and evaluation of statistical models of melodic structure in music perception and composition. London, England: City University.

Pearce MT, Wiggins GA (2006) Expectation in melody: the influence of context and learning. Music Percept 23:377–405.

Pearce MT, Müllensiefen D, Wiggins GA (2010) The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. Perception 39:1365–1389.

Pouget A, Beck JM, Ma WJ, Latham PE (2013) Probabilistic brains: knowns and unknowns. Nat Neurosci 16:1170–1178.

Quiroga-Martinez DR, Hansen NC, Højlund A, Pearce MT, Brattico E, Vuust P (2019) Reduced prediction error responses in high-as compared to low-uncertainty musical contexts. Cortex 120:181–200.

Rabovsky M, Hansen SS, McClelland JL (2018) Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nat Hum Behav 2:693–705.

Seer C, Lange F, Boos M, Dengler R, Kopp B (2016) Prior probabilities modulate cortical surprise responses: A study of event-related potentials. Brain Cogn 106:78–89.

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423.

Simson R, Vaughan HG, Ritter W (1976) The scalp topography of potentials associated with missing visual or auditory stimuli. Electroencephalogr Clin Neurophysiol 40:33–42.

Spratling MW (2017) A review of predictive coding algorithms. Brain Cogn 112:92–97.

Strauß A, Kotz SA, Obleser J (2013) Narrowed expectancies under degraded speech: revisiting the N400. J Cogn Neurosci 25:1383–1395.

Sutton S, Braren M, Zubin J, John ER (1965) Evoked-potential correlates of stimulus uncertainty. Science 150:1187–1188.

Tal I, Large EW, Rabinovitch E, Wei Y, Schroeder CE, Poeppel D, Golumbic EZ (2017) Neural entrainment to the beat: the "missing-pulse" phenomenon. J Neurosci 37:6331–6341.

Tibo M, Geirnaert S, Bertrand A (2020) EEG-based decoding and recognition of imagined music. bioRxiv 2020.09.30.320176.

Walsh KS, McGovern DP, Clark A, O'Connell RG (2020) Evaluating the neurophysiological evidence for predictive processing as a model of perception. Ann N Y Acad Sci 1464:242–268.

Yabe H, Tervaniemi M, Reinikainen K, Näätänen R (1997) Temporal window of integration revealed by MMN to sound omission. Neuroreport 8:1971–1974.

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.