JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Auditory deviance detection across time scales: Effects of local and global context

Matthew Blunt,[a] Stephanie Graceffo,[b] Nahaleh Fatemi,[c] and Mounya Elhilali[d]
Laboratory for Computational Audio Perception, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA

Abstract: Auditory deviance detection reflects the ability to identify violations of statistical regularities in sound sequences and is influenced by stimulus properties and contextual expectations. Using a priming paradigm in which listeners were exposed to different pitch distributions prior to deviant detection trials, we examined how performance varied across the experiment (global context) and relative to short-term priors (local context). Results show that listeners accumulate statistical information across distinct contextual time scales, where local and global context influence deviance detection. Computational modeling using a predictive coding framework is consistent with these effects, highlighting the importance of integrating contextual information across scales in dynamic acoustic environments.

[Editor: Qian-Jie Fu] https://doi.org/10.1121/10.0043222

Received: 17 December 2025 **Accepted:** 17 March 2026 **Published Online:** 2 April 2026

1. Introduction

Auditory salience refers to the inherent quality of a sound that captures attention in an involuntary, bottom-up fashion. It is what makes a sound "stand out" against the background, regardless of attentional focus or task goals. This property allows us to notice a wrong note in a melody, an unexpected word in speech, or a sudden noise in the environment. At the core of this process is the ability to detect when an incoming sound *violates expectations*, which is a process known as deviance detection (Näätänen *et al.*, 2007; Winkler *et al.*, 2009). Deviance detection provides a measurable behavioral and neural signature of expectation violations, and such violations are widely considered to be a key mechanism, contributing to perceptual salience. However, whereas salience is often defined in terms of automatic attentional orienting, deviance detection more directly quantifies sensitivity to expectation violation under explicit task demands. In the present study, we focus on behavioral deviance detection as an index of expectation sensitivity and examine how contextual information shapes detection performance across time. Neural correlates of deviance detection, particularly the mismatch negativity (MMN), have been widely used to study prediction error and adaptive expectation updating in the auditory system (Fitzgerald and Todd, 2020; Garrido *et al.*, 2009). Although MMN responses are frequently observed even when sounds are task irrelevant, their magnitude can be modulated by attentional state (Auksztulewicz and Friston, 2015), indicating that predictive processing and attention interact dynamically. Such work has informed broader theoretical accounts, linking expectation violation to perceptual salience and attentional orienting. Importantly, the strength of deviance detection is not fixed: An identical acoustic event can be perceived as highly salient in one context and nearly undetectable in another.

Early models of stimulus-driven auditory salience emphasized acoustic contrasts in pitch, intensity, and timbre, often borrowing frameworks from vision and treating the spectrogram as an "auditory image" (Kalinli and Narayanan, 2007; Kayser *et al.*, 2005). More recent models have incorporated temporal dynamics and predictive coding principles, emphasizing that deviance detection reflects the brain's estimates of statistical regularities and violations of those predictions (Kaya and Elhilali, 2014; Tsuchida and Cottrell, 2012). Within predictive coding accounts, salience can be understood as emerging when prediction errors signal meaningful departures from learned structure, thereby influencing perception and behavior. Despite these advances, many existing frameworks treat context as a short temporal window or static construct, limiting their ability to capture adaptation over longer time scales observed in natural listening (Barascud *et al.*, 2016; Heilbron and Chait, 2018). In reality, sensitivity to deviants is shaped by expectations operating across multiple time scales.

Evidence from music cognition and auditory scene analysis suggests that listeners integrate local and global statistics when evaluating incoming sounds. Here, local context refers to immediate regularities and short-term sensory history,

[a]Email: matthblunt@gmail.com
[b]Email: sgracef1@jhu.edu
[c]Email: sfatemi2@jhu.edu
[d]Corresponding author: mounya@jhu.edu

such as the pitch distribution established within a melody or phrase. Global context reflects accumulated exposure to regularities across longer time scales, shaping stable priors about what is likely or unlikely in a given environment (Pearce and Wiggins, 2012; Skerritt-Davis and Elhilali, 2021a). Neural studies indicate that the MMN is modulated by recent and long-term context, where sensitivity to deviants decreases under broad or repetitive exposure (Bendixen *et al.*, 2009; Costa-Faidella *et al.*, 2011). Similarly, behavioral evidence suggests that expectations derived from priming or accumulated experience bias perception, shifting the threshold for what counts as a deviant (Bianco *et al.*, 2020; Southwell *et al.*, 2017). Such findings are consistent with hierarchical accounts of statistical learning, in which listeners simultaneously track regularities over short and extended temporal windows (Fiser *et al.*, 2010).

The interaction of local and global statistics is particularly well illustrated in music. Music provides a useful testbed because it contains structured regularities at multiple hierarchical levels, but similar principles apply to speech and environmental sounds (McDermott *et al.*, 2013; Pearce, 2018). A deviant note may sound striking if the preceding context was narrow and predictable but is less noticeable if the listener has adapted to wide-ranging or repeated patterns. Predictive coding accounts suggest that local priors (sharp predictions based on recent input) and global priors (broad expectations built over longer exposure) jointly shape perception, where attention is directed toward events that violate either (Chennu *et al.*, 2013; Friston, 2005; Wacongne *et al.*, 2011). Therefore, understanding how these time scales interact is crucial for explaining how sensitivity to deviants adapts over time in real-world listening.

In this study, we test the influence of local and global context on auditory deviance detection using a priming paradigm. In each block, participants first heard a priming melody drawn from a specific pitch distribution, followed by a series of trial melodies in which a deviant note could occur. By varying the width of the priming distribution across blocks and repeating blocks with identical structure, the paradigm allowed us to test two hypotheses: (1) Broader priming distributions would reduce sensitivity to deviants within a block (local context effect), and (2) repeated exposure to deviants across blocks would reduce deviance detection sensitivity over time (global context effect), reflecting adaptation of expectations across extended exposure. By jointly examining behavioral performance and computational modeling results, this study aims to inform models of auditory salience by clarifying how expectation-driven deviance sensitivity evolves across interacting contextual time scales.

2. Methods

2.1 Experimental paradigm

This paradigm was designed to evaluate how deviance detection sensitivity changes with prior exposure to different piano note distributions across multiple time scales. Specifically, the design allowed local context to be manipulated within blocks via priming distributions, whereas global context was manipulated across blocks through repeated exposure to the same trial structure. The experiment was divided into four blocks: Figure 1(A), the *top row*, shows the distributions of notes used to establish the prior exposure for each block (the priming distribution), whereas the *bottom row* shows the baseline melody notes (black and white keys), the high-salience deviant notes (red keys), and the low-salience deviant notes (blue keys). Each block began with a priming melody composed of notes from the block-specific note distribution. During this phase, participants listened passively without responding. This was followed by 18 trial melodies in which participants judged whether a deviant (change) note was present. Responses ["yes/no" (*Y*/*N*)] were made via timed keypresses with no feedback provided. Trial order was randomized at the beginning of the experiment but remained fixed across all blocks for a given subject, ensuring that differences across blocks reflected contextual effects rather than changes in stimulus identity.

By introducing notes above or below the baseline distribution within a subset of the trial melodies, the paradigm captures the effect of salience based on local changes. Additionally, by exposing participants to different priming distributions for each block but keeping the 18 trial melodies fixed across the 4 experimental blocks, the paradigm allows us to assess how global context influences deviance detection performance. The use of repeated trial sets across blocks ensured that any changes in detection performance reflected accumulated contextual exposure rather than novelty or stimulus-specific effects. Finally, by using the same priming distribution for blocks 2 and 4, the experiment captures how global context, formed by the accumulated statistical patterns during exposure to repeated trials, influences perception, independent of specific priming melodies. Thus, by comparing detection performance between blocks primed with different note distributions prior to the set of trials, the paradigm allows for the evaluation of how shifts in global context (accumulated statistical patterns) influence deviance detection performance over extended exposure, whereas local context (immediate sensory features) shapes perception within individual blocks. Similarly, the two blocks primed with the same distribution enable the evaluation of shifts in deviance detection performance based on the global context formed through repeated exposure to the same statistical pattern.

2.2 Participants

Twenty-two subjects between the ages of 18 and 38 years old ($M = 22.0$ years, SD = 5.0 years), consisting of 14 female, 8 male, and 0 nonbinary subjects participated in the study. All participants were Johns Hopkins University affiliates. Participants reported between 0 and 21 years of musical experience ($M = 6.7$ years, SD = 5.4 years). All participants reported

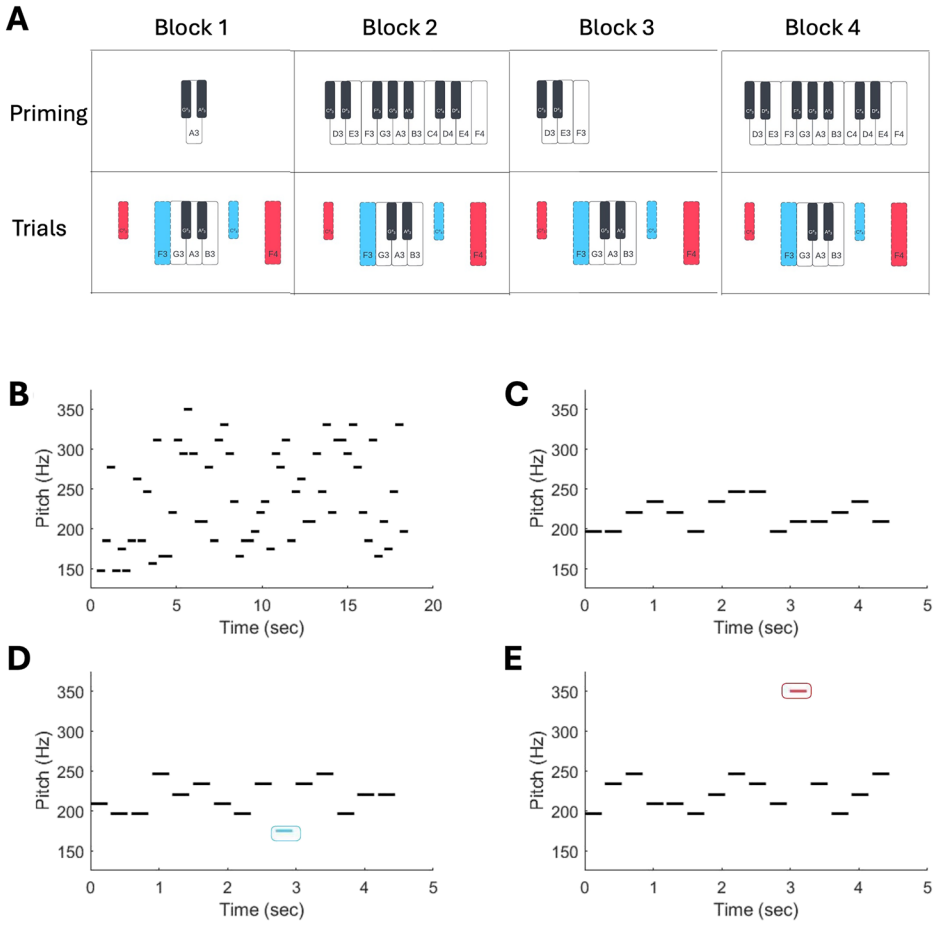

Fig. 1. (A) Priming melody note distributions (top row) and trial melody note distributions (bottom row) are depicted. Baseline note distributions are shown in black and white, low salience level notes are shown in blue, and high-salience level notes are shown in red. (B) Example priming melody, (C) example control trial melody, (D) example low salience level trial melody with introduced salient note outlined in blue, and (E) example high salience level trial melody with introduced salient note outlined in red are displayed.

normal hearing. Each participant provided written informed consent prior to participating and received financial compensation for taking part in the study. All procedures were approved by the Johns Hopkins Institutional Review Board.

2.3 Stimuli

Stimuli for this experiment were generated using 16-bit, 44.1 kHz piano note recordings with normal articulation and forte dynamics, obtained from the *Real World Computing (RWC) Music Database* (Goto *et al.*, 2003). Each note was 1.2 s in duration, with a 20 millisecond cosine-squared onset and offset ramp, and normalized by peak amplitude. Melodies were generated by overlapping the piano note samples every 0.3 s. All melodies with more than two repeated notes of the same pitch were excluded from consideration. Furthermore, each melody went through subjective evaluation to ensure it had musical, pleasant-sounding qualities.

Priming melodies consisted of 18.9-s audio clips generated by uniformly sampling 60 piano notes from a priming distribution, the range of which varied across experimental blocks: block 1, $(G_3 - A_3)$; block 2: $(C_3 - F_4)$; block 3, $(C_3 - F_3)$; and block 4: $(C_3 - F_4)$ [Fig. 1(A), top]. An example priming melody is shown in Fig. 1(B).

Trial melodies closely matched those used by Kaya and Elhilali (2014) and Kaya *et al.* (2020) and consisted of 5.4-s audio clips generated by uniformly sampling 15 piano notes from a baseline distribution with a 5-note range: $(G_3 - B_3)$ [Fig. 1(A), bottom, black and white keys]. Note that although the priming distribution for the priming melodies varied by block, the baseline distribution for the trial melodies was the *same* for all blocks. Six each of three types of trial melody variants were generated: control, low salience, and high salience. Control melodies consisted only of notes from the baseline distribution [Fig. 1(C)]. Low- and high-salience trial melodies also consisted of notes from the baseline distribution, *except* that one note from the baseline distribution was replaced by a note outside the distribution (salient note) at a random time between 50% and 80% of the trial length. The notes preceding and following the salient note fell within

1 semitone above or below the center A_3 note. In three low-salience melodies, the introduced salient note was 2 semitones above the baseline distribution, and in the other three, it was below the baseline distribution [Fig. 1(A) bottom, blue keys]. An example low-salience melody is depicted in Fig. 1(D). In three high-salience melodies, the introduced salient note was 6 semitones above the baseline distribution, and in the other three, it was below the baseline distribution [Fig. 1(A) bottom, red keys]. An example high-salience melody is shown in Fig. 1(E).

White text on a dark gray background served as the only visual stimuli in the experimental design. During presentation of priming stimuli, the text "Listen carefully to the melody." was displayed. During presentation of trial stimuli, the text "Listen for a change in the melody." was displayed.

2.4 Training block

Each experiment began with a training block designed to familiarize participants with the task. In the first phase, participants were presented with three example trial melodies: one control melody and two high-salience deviant melodies. Each was paired with task instructions and a visual representation of the melody, similar to what is depicted in Fig. 1(C) (control) and Fig. 1(E) (high salience). In the second phase, participants were presented with examples of control and high-salience melodies, randomly sampled from a balanced set of 12 example melodies, and asked to indicate if they heard a change in the melody. Feedback was provided after each trial, and participants had up to ten attempts to correctly identify three deviant melodies in a row.

2.5 Subject performance and statistical analyses

Detection was measured using sensitivity (d'), a common measure for deviance detection. The d' metric rewards correct detections, where subjects indicate a change in the melody when a salient note has been introduced, and penalizes false-alarm detections, where subjects indicate a change in a control melody. These values were mapped to the Z-distribution, and values were calculated separately for each experimental block and each subject in three ways, analyzing (1) low salience level melodies, (2) high salience level melodies, and (3) low and high salience level melodies, each referenced to melodies without salient notes. Correction for extreme proportions was handled via the *log-linear rule* outlined in Hautus (1995).

Listeners were tasked with indicating whether they heard a change in the melody in each trial. This note could be of high or low salience and occur above or below the baseline distribution. Because the response data were not normally distributed, significance tests were conducted using Wilcoxon signed-rank test. For multiple comparisons, Bonferroni corrections were applied. Unless otherwise specified, one- and two-way analyses of variance (ANOVAs) were computed using the aligned rank transform (ART) to correct for non-normality.

2.6 Predictive coding framework

To model how listeners track and adapt to statistical regularities in auditory sequences, we employed the dynamic regularity extraction (D-REX) model, which is a predictive coding framework designed to capture expectation formation and violation over time (Skerritt-Davis and Elhilali, 2021a). D-REX maintains a probabilistic estimate of the distribution of recent sensory input and updates this estimate as new sensory observations are acquired. This process generates a time-locked "surprisal," which reflects the degree to which each tone violates the model's current expectations. This formulation is well suited to the present paradigm as the perceptual salience of a deviant tone depends on the recent acoustic context and accumulated exposure across the experiment, possibly including priming sequences.

Using this framework, we implemented *three variants* of the D-REX model that differed only in how contextual information was accumulated or reset across priming sequences, experimental blocks, and the full experiment. These variants were implemented to test distinct hypotheses about the time scales over which listeners integrate contextual information, ranging from short-term, block-specific context to long-term, experiment-wide context. We evaluated model behavior by comparing model-derived predictions to human listener performance using a composite benchmark that captured overall detection performance and systematic changes across the experiment, incorporating sensitivity (d'), hit and false-alarm rates, and block-wise performance profiles, as outlined next.

2.7 Contextual model variants

Three variants of the D-REX model were implemented and differed only in how contextual information was accumulated or reset across priming sequences and experimental blocks [Fig. 4(A)]. All other aspects of the model, including parameter values and decision procedures, were held constant across variants.

No-context model: In this version of the model, contextual information was limited to statistics available within a single experimental block without considering priming sequences. The model was initialized with a common training prior at the beginning of each block and exposed only to the trial sequences for that block without priming tones. Contextual information was, therefore, not shared across blocks. A variant of this model was also considered, where the model was reset at the beginning of every trial in every block and yielded qualitatively similar results.

Local context model: In the local context model, the model was initialized with the same training priors at the start of each block and then exposed to the block-specific priming sequence followed immediately by the trial sequences for that block. Contextual information was retained within a block but reset between blocks, allowing priming to shape predictions locally.

Global context model: In the global context model, the model was initialized once using the training prior and run continuously over the entire experiment. Priming and trial sequences from all four blocks were concatenated in their original temporal order, allowing contextual information to accumulate across priming sequences and trial blocks and capturing long-term integration across the entire experimental session.

Across all three variants, the internal structure of the model and all parameter values were held constant and only the handling of contextual memory differed.

2.8 Model initialization and parameter selection

All model variants were initialized using a common training prior, which was estimated from the set of tutorial tone sequences that were presented to the listeners (see Sec. 2.4), and independent of the experimental priming and test stimuli. This prior represented baseline expectations about pitch statistics prior to experimental exposure and was used consistently across subjects and model variants.

Because trial order was randomized across participants, each listener was presented a unique sequence of stimuli and, therefore, a distinct history of contextual exposure. To reflect these differences, the models were run separately for each subject using that subject's actual trial order. This process resulted in 22 model instances, each corresponding to a different trial history.

In terms of degrees of freedom, the D-REX model includes a small number of free parameters that govern how statistical regularities are inferred and updated over time (Skerritt-Davis and Elhilali, 2021a). These parameters include the *hazard rate*, which controls the degree of contextual volatility (or likelihood of change of statistical context), the *update parameter*, which determines how strongly recent observations influence current predictions, and *observation noise*, which sets a lower bound on prediction uncertainty by reducing the precision of the estimated statistics. The model also assumes an underlying statistical distribution family, here, specified as a second-order Gaussian distribution. These parameters determine the scope, stability, and uncertainty of the regularities to which the model is sensitive.

The value of all parameters was selected empirically. Specifically, we optimized parameters using a composite metric composed of three equally weighted components: correspondence in sensitivity (d') between model and human behavior, correspondence in hit and false-alarm rates, and similarity in block-wise performance patterns (quantified using across-block correlations). Parameter optimization was performed across all three model variants simultaneously, and a single parameter set was selected to minimize this distance, on average, without bias toward any individual model. This approach ensured that differences observed between model variants reflected differences in contextual integration rather than differences in parameter tuning.

2.9 Surprisal computation and model responses

For each model variant, D-REX generated a time-resolved measure of surprisal for every tone in the stimulus sequence, which is defined as the negative log probability of the observed pitch given the model's current predictive distribution. Surprisal values were computed for all tones in priming and trial sequences; however, only surprisal values corresponding to trial tones were used for behavioral evaluation. To derive trial-level model responses, tone-level surprisal values within each trial were reduced to a single decision variable, defined as the largest positive change in surprisal between successive tones, therefore, emphasizing transient increases in prediction error within a trial.

Model (*Y*/*N*) responses were generated by comparing this decision variable to a block-specific threshold derived from the baseline distribution of trial responses within each block. The threshold was defined relative to the mean and variability of this baseline, allowing model sensitivity to be assessed with respect to typical, non-salient fluctuations in surprisal. Trials exceeding this threshold were classified as containing a detected deviant. Using this decision rule, each trial was labeled as a hit, miss, false alarm, or correct rejection, allowing block-wise hit rates, false-alarm rates, and sensitivity (d') to replicate the analysis applied to human responses.

3. Results

3.1 Subject performance

The results first characterize human behavioral performance and then compare these patterns with computational model predictions.

Figure 2(A) shows the average subject sensitivity (d') across the four experimental blocks. d' was highest in block 1, which had the tightest priming distribution (2.53 ± 0.13), and lower in later blocks (block 2, 2.05 ± 0.12; block 3, 2.12 ± 0.13; and block 4, 1.97 ± 0.12). A one-way ART ANOVA confirmed a significant main effect of block [$F(3, 348) = 5.43, p = 0.001$]. *Post hoc* tests showed that d' in block 1 was significantly higher than that in block 2 ($p = 0.032$), block 3 ($p = 0.003$), and block 4 ($p = 0.002$). Overall, performance varied across blocks but was generally lower after block 1,

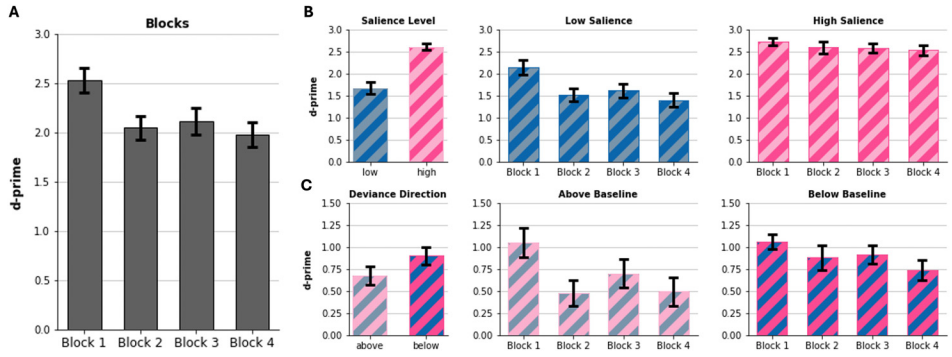

Fig. 2. (A) Sensitivity (d') for each experimental block (averaged across all deviant positions and salience levels); (B) d' for each salience level across all blocks (left), low salience for each block (middle), and high salience for each block (right); and (C) d' for each deviant position across all blocks (left), deviants above the baseline distribution (middle), and deviants below the baseline distribution (right) are displayed.

consistent with an influence of context accumulated across the entire experiment. To further examine whether performance differed among the later blocks, we directly examined block 2 and block 3, as this comparison isolates effects associated with changes in the priming distribution. Block 3 showed higher sensitivity than block 2, yielding a moderate within-subject effect size (Cohen's $d_z = 0.43$). The 95% confidence interval for the mean difference ranged between $(-0.01, 0.36)$, narrowly including zero, and the comparison was only marginally significant $[t(21) = 2.02, p = 0.056]$. This result suggests a rebound effect, although it should be interpreted cautiously given the limited statistical support. Examining subject response time (RT), participants responded slowest in block 1 (1.31 ± 0.16 s) and somewhat faster in later blocks (block 2, 1.21 ± 0.15 s; block 3, 1.08 ± 0.15 s; and block 4, 1.12 ± 0.13 s), although these differences were not significant [Fig. 3(A)].

Although block effects revealed a possible role for context in shaping performance across the experiment, salience levels of deviant tones provided a complementary lens by capturing how strongly bottom-up cues influenced detection overall. Figures 2(B) left and 3(B) left show average subject sensitivity (d') and RTs, respectively, for low- and high-salience deviants collapsed across blocks. High-salience deviants showed significantly greater sensitivity than low-salience deviants (2.61 ± 0.08 vs 1.67 ± 0.14; $W = 0.0, p < 0.0001$) and were also detected to be significantly faster (1.00 ± 0.12 s vs 1.32 ± 0.16 s; $W = 235.0, p = 0.0004$). To determine whether block and salience effects operated independently or in interaction, next, we examined whether the effect of salience level varied across blocks. Such an interaction would indicate that deviance detection performance is shaped by contextual influences operating at different time scales.

Figure 2(B) shows the average subject sensitivity (d') across blocks for *low*-salience deviants (middle) and *high*-salience deviants (right). For low-salience deviants, d' was highest in block 1 (2.14 ± 0.17) and lower in later blocks (block 2, 1.52 ± 0.15; block 3, 1.61 ± 0.16; and block 4, 1.40 ± 0.16). High-salience deviants, however, remained relatively stable across blocks: block 1 (2.73 ± 0.08), block 2 (2.59 ± 0.14), block 3 (2.58 ± 0.10), and block 4 (2.53 ± 0.11). A two-way ART ANOVA revealed a significant interaction between block and salience level $[F(3, 344) = 3.07, p = 0.028]$. Within each block, high-salience deviants had significantly greater d' than low-salience deviants (block1, $W = 9.5, p = 0.019$; block 2, $W = 11.5, p = 0.012$; block 3, $W = 0.0, p = 0.001$; and block 4, $W = 0.0, p = 0.001$). Follow-up Friedman tests show a block effect only for low-salience deviants $[\chi^2(3) = 23.08, p = 2.40 \times 10^{-5}]$ but not for high-salience deviants. *Post hoc* comparisons confirmed that d' in block 1 for low-salience deviants was significantly higher than that in block 2

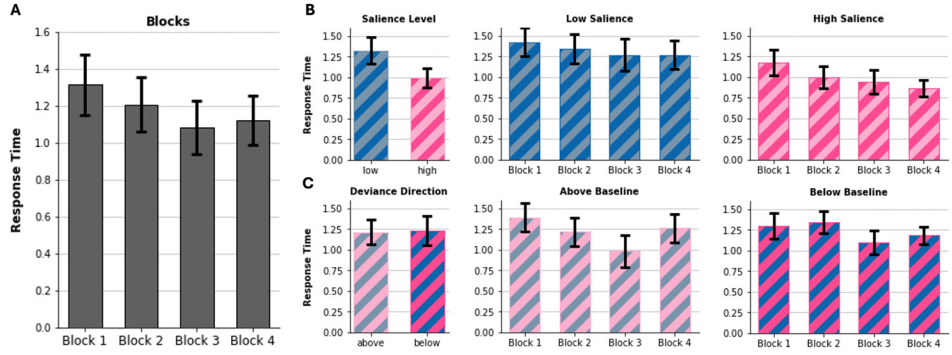

Fig. 3. (A) RT (s) for each experimental block (averaged across all deviant positions and salience levels); (B) RT for each salience level across all blocks (left), low salience for each block (middle), and high salience for each block (right); as well as (C) RT for each deviant position across all blocks (left), deviants above the baseline distribution (middle), and deviants below the baseline distribution (right) are shown.

($p = 0.009$), block 3 ($p = 0.010$), and block 4 ($p = 0.002$). Comparisons between the later blocks indicated that sensitivity for low-salience deviants was higher in block 3 than in block 2 (Cohen's $d_z = 0.42$), although this difference did not reach statistical significance ($p = 0.063$).

Figure 3(B) shows the average subject RT across blocks for *low*-salience deviants (middle) and *high*-salience deviants (right). For low-salience deviants, RTs were slowest in block 1 (1.42 ± 0.17 s) and became faster across later blocks (block 2, 1.34 ± 0.18 s; block 3, 1.27 ± 0.20 s; and block 4, 1.27 ± 0.17 s). RTs for high-salience deviants were consistently faster overall and declined more sharply across blocks: block 1 (1.18 ± 0.16 s), block 2 (1.00 ± 0.13 s), block 3 (0.94 ± 0.15 s), and block 4 (0.87 ± 0.10 s). No significant interaction between block and salience level was observed for RTs, and neither salience level showed a reliable block effect when analyzed separately. Within each block, high-salience deviants were detected significantly faster than low-salience deviants (block1, $W = 207.0, p = 0.036$; block 2, $W = 224.0$, $p = 0.006$; block 3, $W = 214.0, p = 0.018$; and block 4, $W = 221.0, p = 0.009$).

Overall, performance variability across blocks indicates an influence of longer-term contextual exposure, where block-related changes are primarily driven by low-salience deviants, whereas high-salience detections remained relatively stable.

Figures 2(C) left and 3(C) left show average subject sensitivity (d') and RTs, respectively, for above- and below-baseline deviants, collapsed across blocks. d' was higher, although not significantly, for below-baseline deviants (0.90 ± 0.10) than for above-baseline deviants (0.68 ± 0.10). RTs were nearly identical between position (below, 1.23 ± 0.18 s; and above, 1.21 ± 0.15 s). This analysis sheds light on deviance detection as a function of pitch direction independent of block-level context.

Next, to determine whether block and position effects operated independently or in interaction, we examined if the effect of deviant position varied across blocks. Such an interaction would indicate that detection performance is shaped by contextual influences—local (from the priming distribution), global (across blocks), or both. Figure 2(C) shows the average subject sensitivity (d') across blocks for *above* deviants (middle) and *below* deviants (right). For both positions, d' was highest in block 1 (above, 1.05 ± 0.17; and below, 1.06 ± 0.08) and lower in later blocks (*above*, block 2, 0.48 ± 0.15; block 3, 0.70 ± 0.16; block 4, 0.49 ± 0.16; *below*, block 2, 0.88 ± 0.14; block 3, 0.92 ± 0.10; and block 4, 0.74 ± 0.11). A two-way ART ANOVA confirmed significant main effects of block [$F(3, 344) = 5.15, p = 0.002$] and position [$F(1, 344) = 9.59$, $p = 0.002$] but no interaction. Within each block, no significant differences emerged between above and below deviants. Friedman's ANOVA revealed significant block effects for both above deviants [$\chi^2(3) = 25.14, p = 1.44e-5$], and below deviants [$\chi^2(3) = 11.86, p = 0.008$]. *Post hoc* tests indicated that above deviants were significantly higher in block 1 than block 2 ($p = 0.010$), block 3 ($p = 0.002$), and block 4 ($p = 0.0004$), whereas below deviants only differed between block 1 and block 4 ($p = 0.012$).

Figure 3(C) shows average RTs across blocks for *above* (middle) and *below* (right) deviants. For above deviants, RTs were slowest in block 1 (1.39 ± 0.17 s) and generally faster thereafter (block 2, 1.22 ± 0.18 s; block 3, 0.98 ± 0.20 s; and block 4, 1.26 ± 0.17 s). For below deviants, RTs peaked in block 2 (1.34 ± 0.13 s) but were lower in the other blocks (block 1, 1.30 ± 0.16 s; block 3, 1.10 ± 0.15 s; and block 4, 1.18 ± 0.10 s). A two-way ART ANOVA revealed no significant effects of block, position, or their interaction.

Overall, d' was generally lower and RTs were generally faster after block 1, which is consistent with an influence of global context. This pattern was especially pronounced for above deviants. Furthermore, for above deviants, we observed an increase in d' in block 2 to block 3, as well as an increase in RT for block 3 relative to block 4, although neither of these effects reached statistical significance. Given the absence of a significant block × position interaction, there is little indication that deviant position systematically modified performance across blocks.

3.2 Predictive model comparisons

To assess how different forms of contextual integration account for human auditory deviance detection, we compared behavioral performance to predictions from three variants of the D-REX model: a global context model, a local context model, and a no-context model. These models differed only in how contextual information was accumulated across priming sequences and experimental blocks [Fig. 4(A)], allowing differences in performance to be attributed directly to differences in context integration.

When sensitivity (d') was aggregated across all experimental blocks [Fig. 4(B)], human listeners showed overall performance that fell within the range spanned by all three model variants. At this coarse level, the global, local, and no-context models produced broadly similar average d' values, and no statistically significant differences were observed between human performance and any individual model. Because model parameters were selected to optimize average correspondence with human behavior across model variants, some overlap in aggregate performance is expected. This overall similarity, therefore, provides a baseline against which differences in block-level performance and contextual dynamics can be more meaningfully interpreted.

Differences between models became apparent when sensitivity was examined as a function of experimental block [Fig. 4(C)]. Human listeners exhibited the highest sensitivity in block 1, followed by reduced performance in later blocks. Although block 3 showed a moderate numerical increase relative to block 2, statistical support for a reliable rebound was limited. Interpreted cautiously, however, this direction of change is compatible with context-dependent modulation of

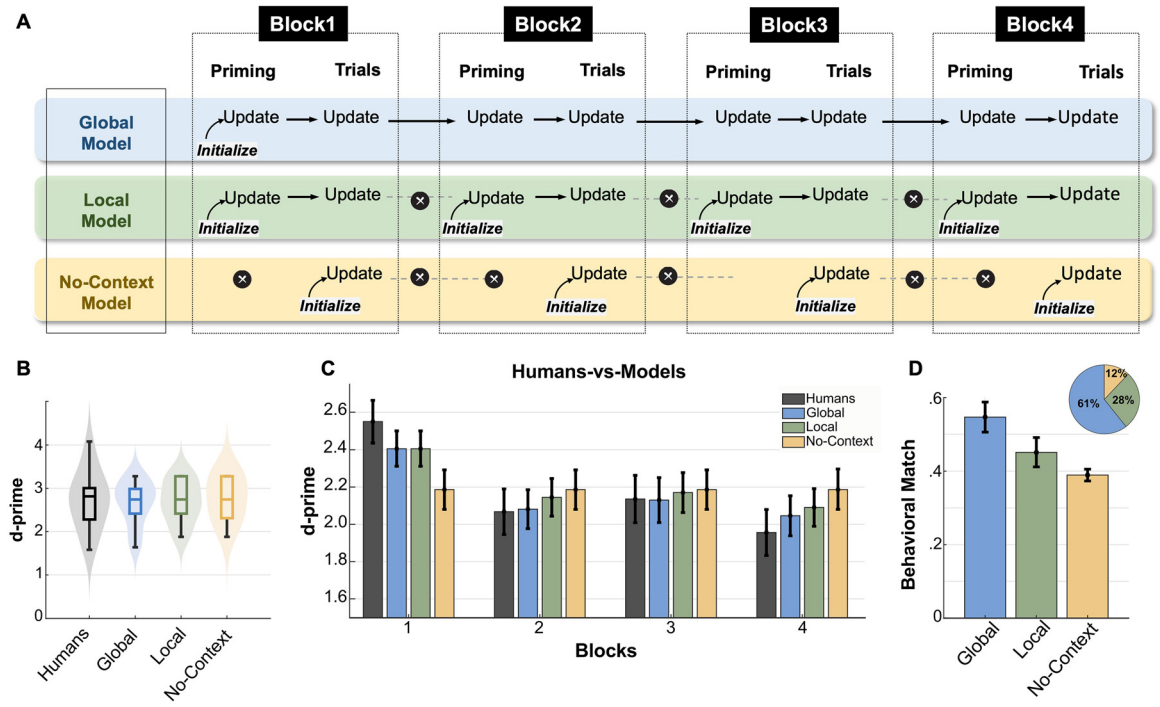

Fig. 4. (A) Schematic of the three D-REX model variants shows how contextual information is initialized and updated across priming sequences and experimental blocks for the global, local, and no-context models. (B) Overall sensitivity (d') is aggregated across the entire experiment for human listeners and each model variant. (C) Block-wise sensitivity (d') for humans and models across the four experimental blocks is depicted. (D) Model–human correspondence quantified using a composite behavioral match metric incorporates overall sensitivity (d'), hit and false-alarm rates, and block-wise performance profiles; the pie chart summarizes the proportion of bootstrap samples (1000 resamples) in which each model provided the best match.

sensitivity across blocks, consistent with global integration of accumulated statistical information. The global context model most closely reproduced this block-wise trajectory. It captured the overall pattern of block-wise modulation, including reduced performance after initial exposure and the relative ordering of the later blocks, which reflects sensitivity to accumulated statistical structure across blocks. The local context model reproduced block-level sensitivity more accurately in magnitude for some blocks but failed to capture the full across-block trajectory, whereas the no-context model showed flat performance across blocks and systematically underestimated human sensitivity especially in later blocks. By design, this model failed to capture global adaptation or block-specific effects, highlighting the necessity of contextual integration for explaining human performance. These patterns across blocks suggest that sensitivity to block-to-block changes in performance depends critically on the integration of global contextual information. It is also worth noting that all three models initially underperformed compared to human listeners in block 1. None of the three models incorporate any additional priors at the outset beyond the training block data. That is certainly not true of human listeners who are exposed to a lifetime of statistical variability across sound sequences that we interact with in everyday life.

To formally quantify model–human correspondence, a composite behavioral match metric was computed that incorporated overall sensitivity (d'), hit and false-alarm rates, and similarity in block-wise performance profiles [Fig. 4(D)]. Using this metric, the global context model provided a significantly better match to human behavior than the local context model [paired t-test, $t(21) = 3.35$, $p = 0.0003$]. The no-context model performed substantially worse than both context-sensitive models. A complementary bootstrap analysis further illustrated this result. Across 1000 bootstrap resamples, the global model yielded the best match to human behavior in 61% of samples compared to 21% for the local model and 12% for the no-context model [Fig. 4(D), pie chart]. This distribution indicates that the superiority of the global model yields more of a robust match across resamplings and is not driven by a small subset of subjects.

4. Discussion

The present results demonstrate that auditory deviance detection is shaped by contextual information operating across multiple time scales. Behaviorally, listeners showed sensitivity to short-term context established within blocks and longer-term context accumulated across the experiment. Computational modeling further revealed that these effects are not redundant: Whereas several model variants approximated human performance at an aggregate level, only models that integrated

contextual information over extended time scales captured the structure of performance across blocks. Together, these findings support the view that auditory salience, as inferred from deviance detection performance, emerges from the interaction of local and global expectations rather than from immediate stimulus properties alone. This interpretation aligns with broader accounts of perceptual adaptation, which emphasize that sensitivity to deviance reflects ongoing inference about environmental structure rather than fixed stimulus-driven responses (Nelken, 2014; Ulanovsky *et al.*, 2003).

Relation to prior work on long-term context and statistical learning: The influence of context on deviance detection has been well documented in behavioral and neural studies. MMN responses, for example, are known to depend on stimulus probability and recent history with reduced responses under conditions of repetition or broader contextual variability (Bendixen *et al.*, 2009; Costa-Faidella *et al.*, 2011; Näätänen *et al.*, 2007). Hierarchical Bayesian modeling work further suggests that MMN reflects prediction errors operating at multiple levels of statistical structure (Garrido *et al.*, 2013). In this study, Garrido *et al.* (2013) demonstrated that neural responses to auditory outliers are shaped by learned statistical regularities across different temporal scales, which is consistent with hierarchical predictive coding accounts. Such findings indicate that the neural signature of deviance detection reflects adaptive updating of expectations rather than a fixed response to physical stimulus differences. The present behavioral results extend this literature by showing that similar contextual adaptation operates over extended experimental time scales and influences perceptual sensitivity even when stimulus structure is held constant.

Short-term contextual effects have also been emphasized in studies of auditory scene analysis and statistical learning, where listeners rapidly adapt to local regularities in pitch, timing, or spectral structure (Saffran *et al.*, 1996; Southwell *et al.*, 2017; Winkler *et al.*, 2009). Our priming manipulation aligns with this work by demonstrating that local pitch distributions shape deviance detection within a block. However, the overall reduction in sensitivity across repeated blocks may reflect not only longer-term contextual adaptation but also more general factors such as reduced novelty, changes in motivation, or time-on-task effects. Notably, RTs did not show reliable slowing across blocks, which would typically accompany fatigue-related declines, suggesting that nonspecific performance factors alone are unlikely to fully account for the observed pattern. Even with these alternatives in mind, the block-wise modulation observed here indicates that performance is unlikely to be fully explained by local adaptation alone, which is consistent with the influence of broader contextual processes operating over extended exposure.

Connections to models of musical expectation and long-term learning: The role of global context observed here is also consistent with models of musical expectation that incorporate learning over extended exposure. Corpus-based approaches, such as the Information Dynamics of Music (IDyOM) framework, combine short-term learning from the unfolding sequence with long-term statistical knowledge acquired from large musical repertoires (Pearce, 2018; Pearce and Wiggins, 2012). In these models, long-term context reflects stylistic expectations accumulated over years of listening experience, shaping how surprising a given musical event is perceived to be. Importantly, related principles have been applied beyond music, including speech perception and environmental sound processing, suggesting that long-term statistical learning is a domain-general mechanism (McDermott *et al.*, 2013; Pearce, 2018).

Although the present study does not rely on culturally learned musical statistics, the behavioral and modeling results suggest that analogous mechanisms operate over much shorter time scales. Repeated exposure to the same trial structure across blocks was sufficient to alter perceptual sensitivity, even when local acoustic structure was held constant. This convergence suggests that the distinction between short-term and long-term context in musical expectation may be quantitative rather than qualitative, where similar computational principles govern learning across different temporal scales.

Implications for predictive coding accounts of auditory salience: Within a predictive coding framework, these findings support hierarchical models in which expectations are maintained and updated at multiple temporal levels (Chennu *et al.*, 2013; Friston, 2005; Wacongne *et al.*, 2011). Local context sharpens predictions based on recent input, enhancing sensitivity to deviations that violate immediate regularities, whereas global context adjusts baseline expectations about the likelihood and relevance of deviant events. The modeling results are consistent with this interpretation: Models that integrated contextual information across blocks provided a better overall match to human behavior, particularly in capturing block-to-block changes in sensitivity, whereas models limited to local context or lacking context altogether failed to reproduce these dynamics.

Importantly, the present results emphasize that auditory salience should be understood as an emergent consequence of expectation violation rather than as a direct function of stimulus novelty or acoustic contrast. Events become salient not simply because they differ physically from preceding sounds, but because they violate expectations shaped by recent and accumulated experience. This perspective aligns with broader accounts of perception as an adaptive, inference-driven process, operating across multiple time scales.

Overall, these findings demonstrate that auditory deviance detection reflects the integration of contextual information over short and extended time scales. By combining a controlled priming paradigm with computational modeling, the present study bridges experimental work on deviance detection with theoretical accounts of statistical learning and predictive coding. More broadly, the results highlight the flexibility of the auditory system in adapting to statistical structure and underscore the importance of considering local and global context when modeling perceptual salience in dynamic acoustic environments.

Acknowledgment

This research was supported by National Science Foundation (NSF) Grant No. 2444353-01. M.B., S.G., and N.F. contributed equally to this work.

Author Declarations

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

All experimental procedures involving human subjects were approved by the Institutional Review Board of Johns Hopkins University. Written informed consent was obtained from all participants prior to participation.

Data Availability

The behavioral data that support the findings of this work are available on the Laboratory for Computational Audio Perception (LCAP) website at https://engineering [Laboratory for Computational Audio Perception (LCAP), 2026]. The computational model is publicly available on GitHub at https://github.com/JHU-LCAP/DREX-model (Skerritt-Davis and Elhilali, 2021b).

References

Auksztulewicz, R., and Friston, K. (**2015**). "Attentional enhancement of auditory mismatch responses: A DCM/MEG study," Cereb. Cortex **25**, 4273–4283.

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., and Chait, M. (**2016**). "Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns," Proc. Natl. Acad. Sci. U.S.A. **113**(5), E616–E625.

Bendixen, A., Schroger, E., and Winkler, I. (**2009**). "I heard that coming: Event-related potential evidence for stimulus-driven prediction in the auditory system," J. Neurosci. **29**(26), 8447–8451.

Bianco, R., Harrison, P. M., Hu, M., Bolger, C., Picken, S., Pearce, M. T., and Chait, M. (**2020**). "Long-term implicit memory for sequential auditory patterns in humans," eLife **9**, e56073.

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., Owen, A., and Bekinschtein, T. A. (**2013**). "Expectation and attention in hierarchical auditory prediction," J. Neurosci. **33**(27), 11194–11205.

Costa-Faidella, J., Grimm, S., Slabu, L., Dőaz-Santaella, F., and Escera, C. (**2011**). "Multiple time scales of adaptation in the auditory system as revealed by human evoked potentials," Psychophysiology **48**(6), 774–783.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (**2010**). "Statistically optimal perception and learning: From behavior to neural representations," Trends Cognitive Sci. **14**(3), 119–130.

Fitzgerald, K., and Todd, J. (**2020**). "Making sense of mismatch negativity," Front. Psychiatry. **11**, 468.

Friston, K. J. (**2005**). "A theory of cortical responses," Philos. Trans. R. Soc. B **360**(1456), 815–836.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (**2009**). "The mismatch negativity: A review of underlying mechanisms," Clin. Neurophysiol. **120**(3), 453–463.

Garrido, M. I., Sahani, M., and Dolan, R. J. (**2013**). "Outlier responses reflect sensitivity to statistical structure in the human brain," PLoS Comput. Biol. **9**(3), e1002999.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (**2003**). "RWC music database: Music genre database and musical instrument sound database," in *Proceedings of International Symposium on Music Information Retrieval*, Baltimore, MD (October 26–30), pp. 229–230.

Hautus, M. J. (**1995**). "Corrections for extreme proportions and their biasing effects on estimated values of *d*," Behav. Res. Methods, Instrum., Comput. **27**, 46–51.

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

Kalinli, O., and Narayanan, S. (**2007**). "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *INTERSPEECH-2007*, Antwerp, Belgium (August 27–31, 2007), pp. 1941–1944.

Kaya, E. M., and Elhilali, M. (**2014**). "Investigating bottom-up auditory attention," Front. Hum. Neurosci. **8**, 327.

Kaya, E. M., Huang, N., and Elhilali, M. (**2020**). "Pitch, timbre and intensity interdependently modulate neural responses to salient sounds," Neuroscience **440**, 1–14.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (**2005**). "Mechanisms for allocating auditory attention: An auditory saliency map," Curr. Biol. **15**(21), 1943–1947.

Laboratory for Computational Audio Perception (LCAP) (**2026**). "Behavioral dataset for auditory deviance detection study," https://engineering.jhu.edu/lcap/.

McDermott, J. H., Schemitsch, M., and Simoncelli, E. P. (**2013**). "Summary statistics in auditory perception," Nat. Neurosci. **16**(4), 493–498.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (**2007**). "The mismatch negativity (MMN) in basic research of central auditory processing: A review," Clin. Neurophysiol. **118**(12), 2544–2590.

Nelken, I. (**2014**). "Stimulus-specific adaptation and deviance detection in the auditory system: Experiments and models," Biol. Cybern. **108**(5), 655–663.

Pearce, M. T. (**2018**). "Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation," Ann. N.Y. Acad. Sci. **1423**(1), 378–395.

Pearce, M. T., and Wiggins, G. A (**2012**). "Auditory expectation: The information dynamics of music perception and cognition," Top. Cognit. Sci **4**, 625–652.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (**1996**). "Statistical learning by 8-month-old infants," Science **274**(5294), 1926–1928.
Skerritt-Davis, B., and Elhilali, M. (**2021a**). "Neural encoding of auditory statistics," J. Neurosci. **41**(31), 6726–6739.
Skerritt-Davis, B., and Elhilali, M. (**2021b**). "D-REX model," GitHub repository, https://github.com/JHU-LCAP/DREX-model.
Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K. J., and Chait, M. (**2017**). "Is predictability salient? A study of attentional capture by auditory patterns," Philos. Trans. R. Soc. B **372**(1714), 20160105.
Tsuchida, T., and Cottrell, G. (**2012**). "Auditory saliency using natural statistics," in *Annual Meeting of the Cognitive Science Society*, Sapporo, Japan (August 1–4, 2012).
Ulanovsky, N., Las, L., and Nelken, I. (**2003**). "Processing of low-probability sounds by cortical neurons," Nat. Neurosci. **6**(4), 391–398.
Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., and Dehaene, S. (**2011**). "Evidence for a hierarchy of predictions and prediction errors in human cortex," Proc. Natl. Acad. Sci. U.S.A. **108**(51), 20754–20759.
Winkler, I., Denham, S. L., and Nelken, I. (**2009**). "Modeling the auditory scene: Predictive regularity representations and perceptual objects," Trends Cognit. Sci. **13**(12), 532–540.