Musical Expertise Is Associated with Improved Neural Statistical Learning in the Auditory Domain

Jacques Pesnot Lerousseau, Daniele Schön

Musical expertise is associated with improved neural statistical learning in the auditory domain.

Jacques Pesnot Lerousseau [1,*], Daniele Schön [1]

[1] Aix Marseille Univ, Inserm, INS, Inst Neurosci Syst, Marseille, France

* Correspondence: jacques.pesnot-lerousseau@univ-amu.fr

Corresponding Author and Lead Contact: Jacques Pesnot Lerousseau, Aix-Marseille Univ, INS, Inst Neurosci Syst, Marseille, France; jacques.pesnot-lerousseau@univ-amu.fr

Conflict of interests: The authors declare no competing interests.

Acknowledgments: We thank Céline Hidalgo and Patrick Marquis for helping with the data acquisition, Clement François, Benjamin Morillon, Maxime Maheu, Christopher Summerfield, and eLife for their Preprint Review service (Maria Chait, Timothy Behrens and three anonymous reviewers).

Funding sources: Work supported by APA foundation (RD-2016-9), ANR-11-LABX-0036 (BLRI), ANR-16-CONV-0002 (ILCB) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

Author contributions: Conceptualization J.P.L. and D.S.; Data curation J.P.L.; Formal Analysis J.P.L.; Funding acquisition D.S.; Investigation J.P.L.; Methodology J.P.L. and D.S.; Project administration D.S.; Resources D.S.; Supervision D.S.; Visualization J.P.L.; Writing – original draft J.P.L. and D.S.; Writing – review & editing J.P.L. and D.S.

Abstract.

It is poorly known whether musical training is associated with improvements in general cognitive abilities, such as statistical learning (SL). In standard SL paradigms, musicians have shown better performances than non-musicians. However, this advantage could be due to differences in auditory discrimination, in memory or truly in the ability to learn sequence statistics. Unfortunately, these different hypotheses make similar predictions in terms of expected results. To dissociate them, we developed a Bayesian model and recorded electroencephalography (EEG). Our results confirm that musicians perform ~15% better than non-musicians at predicting items in auditory sequences that embed either low or high-order statistics. These higher performances are explained in the model by parameters governing the learning of high-order statistics and the selection stage noise. EEG recordings reveal a neural underpinning of the musician's advantage: the P300 amplitude correlates with the surprise elicited by each item, and so, more strongly for musicians. Finally, early EEG components correlate with the surprise elicited by low-order statistics, as opposed to late EEG components that correlate with the surprise elicited by high-order statistics and this effect is stronger for musicians. Overall, our results demonstrate that musical expertise is associated with improved neural SL in the auditory domain.

Keywords: statistical learning, musical expertise, P300, modelling, EEG

Significance statement.

It is poorly known whether musical training leads to improvements in general cognitive skills. One fundamental cognitive ability, statistical learning (SL), is thought to be enhanced in musicians, but previous studies have reported mixed results. This is because such musician's advantage can embrace very different explanations, such as improvement in auditory discrimination or in memory. To solve this problem, we developed a Bayesian model and recorded electroencephalography to dissociate these explanations. Our results reveal that musical expertise is truly associated with an improved ability to learn sequence statistics, especially high-order statistics. This advantage is reflected in the electroencephalographic recordings, where the P300 amplitude is more sensitive to surprising items in musicians than in non-musicians.

Introduction.

Musical training provides exposure to rich statistical structure, which is thought to improve the ability to detect and use statistical regularities. The ability to learn sequence statistics is referred to as “statistical learning” (SL) (Dehaene et al. 2015). Having improved SL abilities in the auditory domain would allow musicians to make more accurate predictions about future events, thus improving perception (Summerfield and de Lange 2014), decision-making (Skinner 1953), and language acquisition (Saffran et al. 1996; Regnault et al. 2001; Koelsch et al. 2002; Kuhl 2004; Koelsch et al. 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Romberg and Saffran 2010; Kim et al. 2011). In this definition, statistics is employed in a broad sense, and can refer for example to the frequency of occurrence of individual items (*e.g.* the frequency of □, P(□), in □□□□□□□) or to the transition probability between items (*e.g.* the transition probability of □➡□, P(□|□), in □□□□□□□). In reference to discrete Markov chain analysis, those statistics reflect different orders of Markov chains. The probability of occurrence of an item given the preceding one, *e.g.* P(□|□), is known as 1st order Markov probability (see Figure 1A top). Similarly, the probability of an item given the preceding K items, *e.g.* P(□|□□), is known as Kth order Markov probability (see Figure 1A bottom). The concept of Markov chain order defines an ordering of SL, from low to high order as K increases.

Musicians perform better than non-musicians in different tasks of different SL orders. For example, numerous studies have shown that musicians have a mismatch negativity component (MMN) of higher amplitude, *i.e.* stronger neural responses to infrequent items than non-musicians (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Putkinen et al. 2014). Musicians are also better at segmenting words from an artificial language stream (Francois and Schön 2011). They have stronger neural responses to violations of 1st order Markov probability (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; François et al. 2012) and to higher order statistics (Vuust et al. 2009; Pearce et al. 2010; Daikoku 2018). However, it is not clear whether the advantage of musicians over non-musicians concerns low or high-order statistics because both have rarely been orthogonalized. Furthermore, it is also not clear whether these differences arise from an improved ability to learn sequence statistics — and if so, at which SL order — or from other processes. Indeed, given the probabilistic nature of the task and the multiplicity of the cognitive processes at stake, the fact that musicians performed better than non-musicians can receive different explanations. We isolated four alternative hypotheses. (**H0**: auditory discrimination) Musicians are better at discriminating between sounds. (**H1**: memory span) Musicians use a longer history of stimuli to make their predictions. (**H2**: SL) Musicians are able to estimate transition probabilities of higher order, such as 2nd order Markov probabilities. (**H3**: selection noise) Musicians have less noise in the selection process, i.e. they lose less information in late stages of the statistical learning process and/or are better at transforming statistical estimates into choices.

The major problem is that these hypotheses (**H0**, **H1**, **H2**, **H3**) make very similar predictions in terms of expected results: impaired auditory discrimination, lower memory, inappropriate statistics and increased selection noise all provoke a decrease in average performances. Confusion regarding these hypotheses could explain the discrepancies

observed in the literature, where the musician advantage is not always replicated (Regnault et al. 2001; Koelsch et al. 2002; Steinbeis et al. 2006; Koelsch et al. 2007; Miranda and Ullman 2007; Koelsch and Jentschke 2008; Koelsch and Sammler 2008; Jentschke and Koelsch 2009; Koelsch 2009a; Loui et al. 2010; Kim et al. 2011; Rohrmeier et al. 2011). Finally, previous computational models used to study trial-by-trial responses (Squires et al. 1976; Regnault et al. 2001; Koelsch et al. 2002, 2007; Mars et al. 2008; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Kolossa et al. 2012; Meyniel et al. 2016) were developed to uncover the general principles of SL. Being general by nature, they do not incorporate subject-specific parameters, and are thus unable to account for inter-individual differences (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Siegelman, Bogaerts, and Frost 2017; Siegelman, Bogaerts, Christiansen, et al. 2017). In the current work, rather than sticking to performance, we relied on modelling tools to tease apart the different hypotheses: we created a Bayesian model that makes explicit predictions, recorded electroencephalography (EEG) and assessed both trial-by-trial responses and inter-individual differences.

We asked the following questions: do musicians have better abilities to predict items in auditory sequences than non-musicians? If yes, can we tease apart the possible sources of this advantage? Can we identify a neural correlate of the ability to learn auditory sequence statistics? To answer these questions, we first evaluated whether musicians have better SL abilities in the auditory domain than non-musicians by measuring their ability to predict the forthcoming items of auditory sequences that embed either low or high-order statistics. We first ascertained that stimuli were easily discriminable by musicians and non-musicians. We then tested each potential source of this advantage by fitting a computational model and contrasting subject-specific parameters. We identified a neural correlate of the musician's advantage by correlating the model surprise with the EEG response amplitude at each time point, and compared the strength of this correlation in musicians and non-musicians. Finally, we explored the temporal structure of the EEG response by correlating the surprise elicited at very low, low and high-order SL with the EEG response amplitude at each time point. Based on the previous literature and because music provides a rich statistical structure, we made the hypothesis that musicians would be better at predicting items than non-musicians and that the advantage of musicians would be particularly pronounced for high-order statistics. We also made the hypothesis that the advantage of musicians would be explained by parameters relating to statistical learning, as opposed to parameters relating to memory or action selection. We also expected to observe a neural correlate of the musician's advantage, with a higher modulation of the EEG response by surprise in musicians than in non-musicians.

Methods.

STIMULI AND PARADIGM.

Participants.

We collected data from 27 musician participants (17 females, mean age 33.3 y, standard deviation ± 12.2, range [18, 62]) and 26 non-musician participants (15 females, 31.1 y ± 11.3 [20, 55]). All participated provided a written informed consent. All had normal hearing, reported no neurological deficits and received 20 euros for their time. Musicians had at least 10 years of intensive musical practice (19.9 y ± 11.4 range [10, 46], onset of practice, 7.8 y ± 4.7 [4, 25]). Non-musicians had no musical training. All participants were recruited at the university, among students and professors. This was done to increase the homogeneity in levels of education and in socioeconomic status between groups. Musicians were players in the university orchestra. Groups were matched on age (linear regression, □ = 2.06 ± 3.23, p = 0.53) and sex (logistic regression, □ = -0.22 ± 0.56, p = 0.70). No additional demographic variables were recorded. The experiment was approved by the Aix-Marseille University Ethics Committee on research on human subjects.

Stimuli.

Across the experiment, 10 sequences of 300 items (sounds) were presented to the participants. The vocabulary consisted of three items A, B and C. Each sequence contained the same amount of item types (100 As, Bs and Cs). The order of the items of the 10 sequences was designed to probe two levels of statistical learning order. Five sequences were 1st order Markov chains: each item was chosen only based on the previous item given the corresponding column in a transition probability matrix of size 3x3. This matrix described the probability of choosing a particular item given the preceding one, *e.g.* P(A|B). The matrix was biased so that each item was followed primarily by one item (p=0.8) compared to the other two (p=0.1). The other five sequences were 2nd order Markov chains: each item was chosen only based on the previous pair of items given the corresponding column in a transition probability matrix of size 9x3. This matrix described the probability of a particular item given the preceding pair, *e.g.* P(A|CB). The matrix was biased so that each pair was followed primarily by one item (p=0.8) compared to the other two (p=0.1). A new transition probability matrix was used for each sequence (in total five 1st order matrices and five 2nd order matrices for each participant). Before performing the experiment, we selected the sequences that allow proper parameter estimation (see Model parameter recovery). Note that the number of transition probabilities to be tracked grows exponentially with the order of the Markov chain, making high-order sequences (P(A|AA), P(A|BA), …, P(C|CC) → 27 probabilities in total) harder to predict than low-order ones (P(A|A), P(A|B), …, P(C|C) → 9 probabilities in total).

The three items A, B, and C were randomly assigned to three sounds for each participant. The same three sounds were used for each participant. The sounds were artificially generated impact sounds : wood, metal and glass (Aramaki et al. 2006). Importantly, all sounds had the same fundamental frequency, loudness and duration, and differed only in timbre (examples of “tuned” sounds available at http://www.lma.cnrs-mrs.fr/~kronland/Categorization/sounds.html). Each sound was 150 ms long, with cosine

ramp on and off of 10 ms, presented at a fixed rate of ~1.6 Hz (onset asynchrony of 600 ms). Each sequence lasted ~4.5 min.

Explicit prediction judgments were probed 20 times per sequence. Probes were randomly spaced by 9, 12, 15, 18 or 21 items. During a probe, participants had 4s of silent intervals to indicate the most likely forthcoming item, using key press on a keyboard.

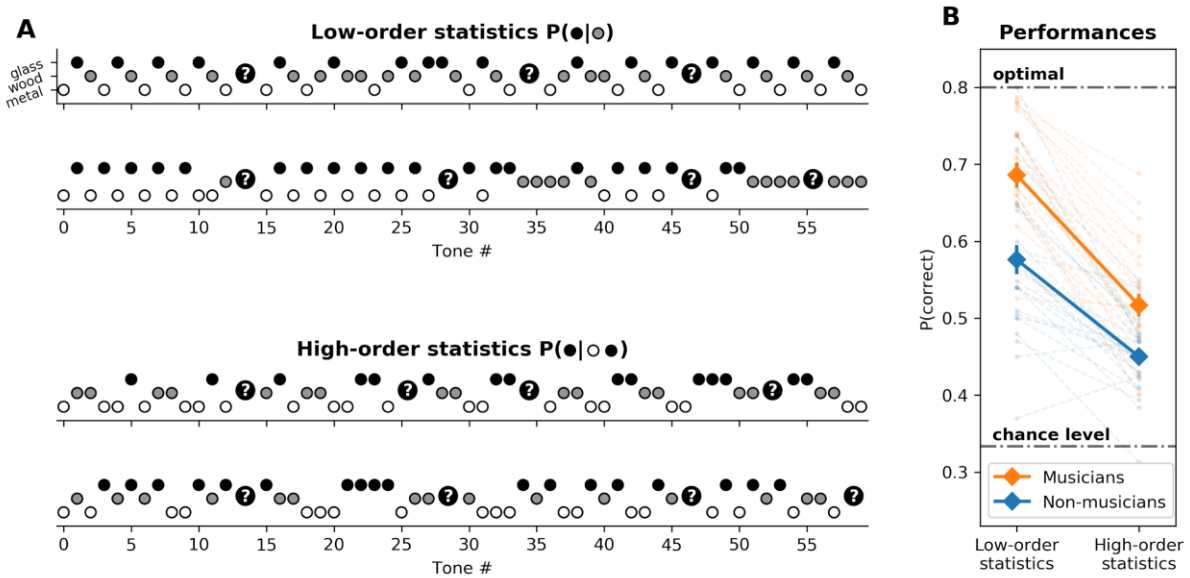

Figure 1. Musicians are better than non-musicians at predicting the forthcoming items of auditory sequences that embed either low or high-order statistics. A. Paradigm. Each sequence was composed of 300 sounds, chosen among 3 (impact sounds of glass, wood and metal; here respectively black, grey and white). Two types of sequences were generated, two examples of each are shown: (top sequences) Low-order statistics: 1st order Markov chains, defined by $P(s_t|s_{t-1})$. Each item is chosen based on the previous one. (bottom sequences) High-order statistics: 2nd order Markov chains, defined by $P(s_t|s_{t-2}s_{t-1})$. Each item is chosen based on the previous pair of items. Participants were randomly probed for an explicit prediction about the forthcoming tone 20 times per sequence, here symbolized by question marks. **B.** Behavioral results. Overall, the performances were higher than chance (33%) but lower than the theoretical optimum (80% — sequences are probabilistic therefore 100% is not achievable). Participant's predictions were closer to the generative statistics for low-order sequences, compared to high-order sequences ($p < 10^{-16}$). On both types of sequences musicians were better than non-musicians ($p < 10^{-7}$). Error bars represent standard error of the mean (s.e.m.).

Procedure.

Classical SL paradigms are usually composed of two phases (Saffran et al. 1996; Regnault et al. 2001; Koelsch et al. 2002; Perruchet and Pacton 2006; Koelsch et al. 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Romberg and Saffran 2010; Kim et al. 2011): a habituation block, consisting of the presentation of a sequence of items embedding the statistical regularity to be learned, and an evaluation block, consisting of the presentation of sequences that respect the regularity ("standard") or not ("deviant"). Learning is indirectly measured as the impact of the violation of the rule, typically differences in reaction time or accuracy between "standard" and "deviant" items. Our paradigm differed in two aspects. (1) We evaluated SL during the learning block and eliminated the evaluation block. Beyond considerably reducing the duration of the experiment, this solves critical problems related to forgetting and learning occurring in the evaluation block (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; François et al. 2012). As the stimuli were probabilistic sequences and as learning is a continuous process, there was no *a priori* way to classify items into two binary classes, like "standard" or "deviant". Instead, we relied on modelling tools and information theory to define the degree of expectation violation/fulfillment as the "theoretical surprise" elicited by each item. (2) The

task of the participant was to make explicit predictions, therefore SL was not indirectly measured *via* infrequent violations of the rule, but directly *via* the accuracy of the predictions and the adequation between participants' responses and model predictions.

Participants were seated in a soundproof room in front of a computer screen, a loudspeaker and a keyboard. Prior to the experiment, normal hearing was assessed using a rapid 5dB-step audiogram. Participants were then familiarized with the stimuli and the mapping between the items and response keys in a familiarization block of 50 items. During this block, items were presented in a random order and participants had to press the key corresponding to the heard sound at every item (three keys "left", "up" and "down"). The mapping between keys and sounds were fixed during the whole experiment and randomized across participants. The principal aim of the familiarization bloc was to train the participant to map the sounds and the keys. The second aim was to confirm that sounds were easily discriminable, by measuring the accuracy of the participants. On average, participants scored 97.8 % (± 0.5) of correct responses during this familiarization block.

Participants were then instructed to listen to the sequences, and to predict the forthcoming item using the keyboard whenever they saw the probe screen. Emphasis was put on accuracy and not speed. Each participant did 10 sequences, 5 of each type (1^{st} order and 2^{nd} order Markov chains), in a random order. Participants were informed of the unpredictable nature of the sequences and of the difference between the two types of sequences. This was done to minimize the use of incorrect strategies, such as trying to learn patterns or trying to uncover deterministic rules. More precisely, they were told that the sequences contained more or less regular patterns, but that there was always some variability engendering unpredictability. In piloting the experiment, most participants reported in the debriefing that some sequences were more difficult than others. We thus decided to inform all participants that some sequences were more difficult than others so as to prevent differences due to the awareness of the complexity level. They could take small breaks between each sequence. The whole experiment lasted ~50 min.

The sequences were presented binaurally to participants at an adjusted comfortable level (~70 dB) using loudspeakers. Stimuli presentation and data collection was controlled with Python custom scripts. Submillisecond synchrony between stimulus presentation and EEG acquisition was ensured using triggers embedded in the audio files and delivered to the acquisition computer via a dedicated channel.

BEHAVIORAL ANALYSES.

Outliers.

One participant (musician) was removed from the analyses because of poor performances in the familiarization block (80%, $z < -3$ on the z-scored performances scale of the group).

Model-free statistical analyses.

Statistical analyses were done using R and the package lme4 (Bates et al. 2014). The effects of sequence type, musicianship and their interaction on performances were estimated using logistic mixed-effect models. The random effect structure was set to take into account the experimental design, with a random intercept and a random "sequence type" slope for each participant. The probability of a correct response (0: incorrect, 1: correct) was modeled as a logistic regression with "sequence type" (0: 1^{st} order Markov chain, 1: 2^{nd} order Markov chain), "musicianship" (0: non-musicians, 1: musicians) and their

interaction as predictors. Model complexity was monitored using the Akaike Information Criterion, a standard measure to arbitrate between complexity and accuracy. The logistic mixed-effect model with the smallest Akaike Information Criterion (best model) was a model including “sequence type”, “musicianship” and their interaction as fixed effects. Reported p-values are Satterthwaite approximations.

MODELING ANALYSES.

Computational model.

The optimal model was based on a previously published model IDyOM (Pearce and Wiggins 2012; Harrison et al. 2020), an n-gram model (Chen and Goodman 1999) that we reframed and extended in a Bayesian framework. It is based on an “ideal observer” model, in the sense that it exploits all available information to compute predictions: it learns the correct generative model, without noise and with perfect memory. Its predictions are encoded in a “transition probability matrix”, that links immediate contexts and items. This matrix is continuously updated using Bayes rule, given the observed sequence. Formally, the model is exposed to a sequence of T items $s_{0:T-1}$ of a vocabulary Ω of size V. The context is given by the last K items of the sequence. As a Bayesian ideal observer, she uses Bayes rule to update her belief:

$$P(\theta|s_{0:T-1}) \propto P(s_{0:T-1}|\theta)P(\theta)$$

The transition probability matrix □ describes how likely each element is, given its preceding context. This learning process consists in estimating □ from the sequence $s_{0:T-1}$. This computation is based on the matrix N that contains the number of occurrences of each (K+1)-uplets of items in the sequence $s_{0:T-1}$. The matrix N is a matrix of size V x V^K where each row designates a particular item and each column a particular K-uplets of items. N_{ij} designates the number of occurrences of the (K+1)-uplet corresponding to the cell (i, j) of the matrix N (the j^{th} K-uplet followed by the i^{th} item). The full derivation of the likelihood term is given in Supplementary Methods. In the end, the predictive posterior probability of the model for each item is, rather naturally:

$$P(s_T = x_i|s_{0:T-1}) = P(s_T = x_i|s_{T-K:T-1}) = \frac{N_{ij} + 1}{\sum_{v=1}^{V} N_{vj} + 1}$$

where $P(s_t = x_i|s_{0:T-1})$ is the probability of the element $x_i \in \Omega$. It corresponds to the ratio of two scalars:

- N_{ij}+1: the number of times that the context (j^{th} column), *i.e.* the last K elements of the sequence, *i.e.* $s_{T-K:T-1}$, and the element x_i (i^{th} row) have been observed together plus one,
- $\sum_v (N_{vj}+1)$: the number of times that the context $s_{T-K:T-1}$ (the sum of the j^{th} column) has been observed plus V.

On top of this ideal observer, three sources of noise were added: imperfect memory (□), order of the estimated transition probability matrix (K) and selection noise (β). First, the size of the context K varied between 0 and 2. Second, a leak parameter was introduced to

account for memory decay. Previous observations were weighted by a weight $e^{-t/\lambda}$ for the t^{th} past stimulus. A small λ indicates quicker memory decay and worsen the performances. Note that λ is not related to K: K is the size of the chunk taken into account to compute the statistics (K = 2 means statistics on the form P(□|□□), while λ modulates how far in the sequence the observer looks to estimate this statistics). Finally, predictions were transformed into choice probability via a softmax function with inverse temperature β. A high β indicates high noise in the decision process and worsens the performances.

Formally, the predictive posterior probability of the model with noise is derived as follows:

$$I_{t,i,K} = 1[s_{t-K:t} = s_{T-K:T-1}x_i]$$

"1" is the indicator function. $I_{t,i,K}$ is a helpful indicator to count the number of times that the context ($s_{T-K:T-1}$) has been observed with the element x_i in the whole sequence ($s_{0:T-1}$). At any given point t in the sequence, $I_{t,i,K}$ is equal to one if and only if the substring $s_{t-K:t}$ is equal to the substring $s_{T-K:T-1}x_i$. Otherwise $I_{t,i,K}$ is equal to zero.

$$N_{i,\lambda,K} = \sum_{t=K}^{T-1} (e^{\frac{t-T+1}{\lambda}} I_{t,i,K})$$

$N_{i,\lambda,K}$ is the number of times that the context that the context ($s_{T-K:T-1}$) has been observed with the element x_i in the whole sequence. This counting is weighted by an exponential decay parameter. Past observations weigh less than recent observations. This weighting is the same as the PPM-Decay model (Harrison et al. 2020), an updated version of the IDyOM model.

$$P_{i,\lambda,K} = \frac{N_{i,\lambda,K} + 1}{\sum_{v=1}^{V} (N_{vj} + 1)}$$

Where $\sum_v (N_{vj}+1)$ is the number of times that the context ($s_{T-K:T-1}$) has been observed in the whole sequence, plus V.

$$P_{model}(s_T = x_i | s_{0:T-1}) = \frac{e^{-\beta P_{i,\lambda,K}}}{\sum_{v=1}^{V} e^{-\beta P_{v,\lambda,K}}}$$

$P_{i,\lambda,K}$ is transformed into a choice probability P_{model} by taking the softmax of parameter β over all possible elements in the vocabulary Ω.

The simulated effect of each parameter is displayed on Fig. Supp. 2.

The model was run on each sequence independently and re-initialized at the beginning of each sequence (it does not keep track of probabilities present in the other sequences).

Model parameter recovery.

We assessed our model selection procedure with a parameter recovery analysis. This standard procedure in modelling (Palminteri et al. 2017) ensures that there is no bias in

the parameter estimation, *i.e.* the values of the parameters do not suffer from a systematic overestimation or underestimation. We generated synthetic data for 10^3 models with random parameters (uniform between 2 and 1000, K, uniform choice between 0, 1 and 2, β exponential between 0 and 1). We then ensured that the estimated parameters from these synthetic data using our procedure were close to the original parameters. This was the case using our sequence and the same number of trials as our participants (spearman ρ = 0.91, ρ_K= 0.99, ρ_β = 0.96, see Fig. Supp. 8).

Model fitting.

The model fitting values reported in the paper are maximum likelihood estimates (see Fig. Supp. 3). Formally, for a set of responses $r_{0:N}$, a model $M_{,K,\beta}$ with parameters , K, β, and assuming that each response is independent, the likelihood was defined for each participant as:

$$\mathcal{L} = P(r_{0:N}|s_{0:N}, M_{\lambda,K,\beta}) = P(r_0|s_0, M_{\lambda,K,\beta})P(r_1|s_{0:1}, M_{\lambda,K,\beta}) \dots P(r_N|s_{0:N}, M_{\lambda,K,\beta})$$

$$log\mathcal{L} = log(P(r_{0:N}|s_0, M_{\lambda,K,\beta})) + \dots + log(P(r_N|s_{0:N}, M_{\lambda,K,\beta}))$$

As the logarithm is a monotonically increasing function, finding the maximum of the likelihood is the same as finding the maximum of the log-likelihood. As the number of parameters is low (, K, β), we relied on grid search to find estimates of this maximum. We computed the log-likelihood for each participant on trial-by-trial responses, for 200 logarithmically spaced values of (range 1, 1000), 3 values of K (range 0, 2), and 200 logarithmically spaced values of β (range 0, 2). For each participant, the argmax of the log-likelihood on this grid was defined as its parameters estimates.

Between-group parameter comparison.

The subject-specific parameters were then compared between groups. The group comparison for parameters and β was done using a linear regression, with “musicianship” as predictor (0: non-musicians, 1: musicians). Both parameters and β were log-transformed prior to the test to satisfy the assumptions of the linear model. The group comparison for parameter K was done using a logistic regression (0: K=1, 1: K=2), with “musicianship” as predictor (0: non-musicians, 1: musicians).

ELECTROENCEPHALOGRAPHY (EEG).

Apparatus.

EEG signal was recorded at 1000 Hz sampling rate using a BrainAmp amplifier and 64 preamplified Ag–AgCl electrodes mounted following the 10–10 international system (actiCap) in a soundproofed Faraday cage. The ground electrode was placed at AFz and the reference electrode at FCz.

Preprocessing.

Signal processing was done using MNE-python (Gramfort et al. 2013) and custom scripts written in Python. Continuous data were bandpass filtered (1-40 Hz, zero-phase Hamming window FIR filter) and major artifacts rejected by visual inspection. Independent component analysis (fastICA) was used to remove physiological artifacts such as eye blinks and muscular activity. Data were then segmented into epochs of 600 ms starting at 50 ms

prior to item onset and stopping 550 ms after. Epochs were zero-mean normalized to baseline ([–50, 0] ms) and re-referenced to the algebraic average of all electrodes.

Multiway canonical correlation analysis (MCCA).

Brain signals recorded with EEG have poor signal-to-noise ratio due to the presence of multiple competing sources and artifacts. A common remedy is to average recorded signals over multiple repetitions of the same stimulus and over multiple participants. However, averaging across participants is problematic, because differences in brain sources and geometry considerably increase variance. To deal with this problem, we relied on a powerful yet simple method recently developed (de Cheveigné et al. 2019). Multiway canonical correlation analysis (MCCA) consists of summarizing the data into individual spatial filters, named “summary components” (SC). These individual filters are built to maximize the temporal correlation between participants. MCCA was run on pooled data of musicians and non-musicians, in order to avoid spurious group differences. The first SC explained on average 55% of the variance of the ERP (peak correlation between SC time course and ERP at Fz, 93% explained variance) and was therefore selected for the rest of the study. Electrodes best explained by the SC were FC1, FC2, F1, Fz, F2, FC4, FC3, C2, C1, F4, F3, Cz, C3, which is consistent with an auditory response topography.

Model surprise regression.

As the stimuli are probabilistic sequences and as learning is a continuous process, there is no *a priori* way to classify items into two binary classes, like “expected” or “unexpected”. Instead, we relied on information theory (Shannon 1948) to define the degree of expectation as the “surprise” elicited by each item. Formally, the surprise is defined as $-\log_2 P(x_i)$ where $P(x_i)$ is the posterior probability $P(s_T = x_i | s_{0:T-1}, M_{\square,K,\beta})$ of the model $M_{\square,K,\beta}$ on the presented item s_T. Critically, the surprise depends on a particular model of the world $M_{\square,K,\beta}$, that ascribes a probability $P(s_T = x_i)$ to each possible item at each time step T. We defined $M_{\square,K,\beta}$ as our model fitted on behavioral responses. We then associated a level of theoretical surprise to each item of each sequence for each participant.

We therefore used the parameters extracted by model fitting from the behavioral responses (that concern only a small fraction of all items) to study the EEG responses to each item (280 items x 10 sequences per participant). A value of surprise was defined for each participant, each sequence, and each item, based on the behaviourally fitted parameters. We regressed the surprise against the SC amplitude in epochs that did not correspond to behavioral probes (280 items x 10 sequences). For that, we computed the linear regression of the surprise at each time point, ending up with an array of linear coefficients of size 52 x 600.

Statistical significance of the coefficients was assessed using cluster permutations in time (n = 2048). A t-test against 0 was performed and cluster-corrected for each group to assess the significance of the linear regression (the statistic was the sum of the t-values in the cluster). An independent t-test was performed and cluster-corrected between groups to assess the significance of the difference of the coefficients between musicians and non-musicians (the statistic was the sum of the t-values in the cluster).

Results.

Musicians perform better than non-musicians in an auditory task implicating SL.

Results were analyzed using mixed-effect logistic regression (see Figure 1B, see Methods). First of all, participants were better at predicting items in sequence embedding low-order statistics (63.2 ± 1.5 %) compared to sequences embedding high-order statistics (48.4 ± 1.0 %, □ = –0.51 ± 0.07, $p < 10^{-14}$). In other words, while all participants performed well above chance level (33%), their predictions are closer to the generative statistics for low-order sequences compared to high-order sequences. It should be noted that performances were intermediate between a random strategy (33 %) and the theoretical optimum (80 % - sequences are probabilistic therefore 100% is not achievable), indicating that the stimuli are well designed to study the imperfection (suboptimality) of cognitive processes and the inter-individual differences. On average, musicians (60.1 ± 2.2 %) were better than non-musicians (51.3 ± 1.9 %), and this effect is significant (between-group comparison □ = 0.48 ± 0.11, $p < 10^{-5}$). The interaction is significant (□ = -0.22 ± 0.10, p = 0.021). The negative sign indicates that the effect of musical expertise is less important for high-order than for low-order statistical regularities. Yet, this difference holds for both types of sequences: musical expertise is associated with higher performance in a statistical learning task, for both low-order (□ = 0.48 ± 0.11, $p < 10^{-4}$) and high-order (□ = 0.27 ± 0.07, $p < 10^{-3}$) statistical regularities. Control analysis revealed that musicians and non-musicians did not benefit from an overall increase in performance during the course of the experiment (effect of block rank □ = –0.001 ± 0.01, p = 0.84, interaction with musicianship □ = 0.02 ± 0.01, p = 0.10), ruling out the possibility of a difference between groups due to task learning.

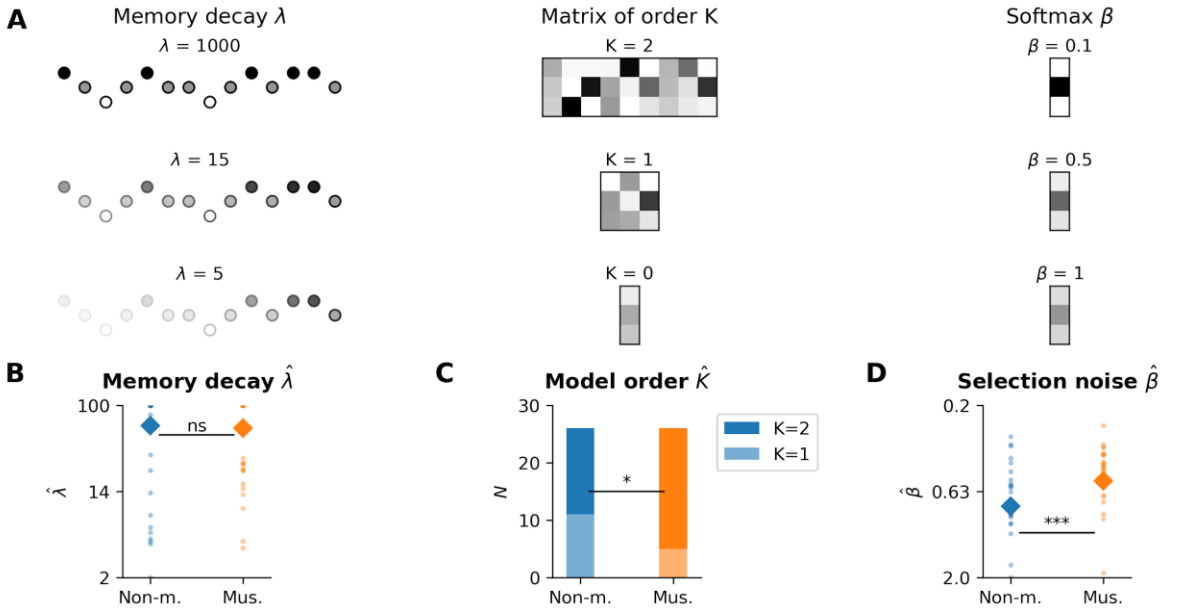

Figure 2. Modelling reveals key differences between musicians and non-musicians learning strategies. A. Model. The model comprises three components. (left) The sequence is weighted over time by an exponential decay function of parameter □, mimicking memory forgetting (low values deteriorate performances). (middle) The model learns the transition probability matrix between contexts and items. The context can be composed of K = 1 or K = 2 items. The model can also be sensitive to the probability of an item without any reference context (K = 0). (K < 2 deteriorates performances for high-order sequences). (right) The values of the transition probability matrix are converted into choice probability via a softmax rule, controlled by a selection noise parameter β (high values deteriorate performances). **(B)** Memory decay parameter □ did not differ significantly (p = 0.80) between musicians and non-musicians. **(C)** Model order parameter K was higher in musicians than in non-musicians

(p = 0.038), indicating a better match with a model that estimates higher order statistics. **(D)** Selection noise β was lower in musicians than non-musicians ($p < 10^{-3}$). Transparent dots represent individual data. Error bars represent standard error of the mean (s.e.m.).

Musicians estimate higher order transition probabilities (K) with a lower selection noise (β) compared to non-musicians.

We then used the model to tease apart which of the four alternative hypothesis explained the musicians's advantage: (**H0**: auditory discrimination) Musicians are better at discriminating between sounds. We eliminated this hypothesis by creating stimuli that were easily discriminable. Indeed, participants scored 97.8 % (± 0.5) of correct responses in an item identification task during the familiarization block, with no significant differences between musicians and non-musicians (= 0.57 ± 0.55, p = 0.29). Furthermore, EEG recordings during this task revealed no amplitude nor latency differences between musicians and non-musicians (see Fig. Supp. 1). (**H1**: memory span) Musicians use a longer history of stimuli to make their predictions. (**H2**: SL) Musicians are able to estimate higher order statistics. (**H3**: selection noise) Musicians have less noise in the selection stage.

The model we designed (see Figure 2A, see Fig. Supp. 2, see Methods, see Supp. Methods for a formal derivation of the model) encodes its predictions in a "transition probability matrix", that links immediate contexts and items. This matrix is continuously updated using Bayes rule, given the observed sequence. The context is given by the last K items of the sequence. We added three sources of noise to this model: imperfect memory (), order of the estimated transition probability matrix (K) and selection noise (β). The model was fitted to each participant's trial-by-trial responses (see Methods, see Figure 2B, Fig. Supp. 3). This procedure led to a set of three parameters per participant: , K, β. On average, the model explained 61.5 % (± 1.2) of the responses (musicians: 66.8 % ± 1.7; non-musicians: 56.3 % ± 1.5, low-order statistics sequences: 61.2 % ± 1.4, high-order statistics sequences: 62.1 % ± 1.4, between-group comparison = 10.5 ± 2.0 %, $p < 10^{-4}$, between-condition comparison = 0.96 ± 1.5 %, p = 0.53). It should be noted that this is higher than chance (33%, $p < 10^{-16}$) but also higher than the performances of the participants (55.8 ± 1.1 %, = 5.4 ± 0.5 %, $p < 10^{-13}$). This difference points to the fact that the model must be predicting more than just the participants' correct responses, *i.e.* it must be predicting their errors as well.

The comparison of the fitted parameters revealed key differences between the two groups. (**H1**) Memory decay fitted values were not significantly different between musicians and non-musicians (= 0.09 ± 0.34, p = 0.80). (**H2**) The order K of the estimated transition probability matrix was higher in musicians compared to non-musicians (= 1.39 ± 0.67, p = 0.038). This indicates that musicians tend to estimate higher order statistics. More specifically, the statistics they estimated (85 % of participants at K = 2) was the same as the statistics that generates the highest order sequences, *i.e.* 2nd order Markov chains. Non-musicians were closer to a model that estimates lower-order statistics (58 % of participants at K = 2). This suggests that, compared to musicians, non-musicians estimate lower-order statistics, leading to a loss of performance. (**H3**) Finally, the selection noise was lower in musicians compared to non-musicians (= -0.38 ± 0.11, $p < 10^{-3}$). Critically, the individual values of selection noise did not correlate with performances (= –1.38 ± 1.15, p = 0.23) nor with reaction times (= 0.21 ± 0.22, p = 0.37) in the item identification task of the familiarization block. This indicates that does probably not represent response mapping

confusion nor task engagement – global effects that should be observed similarly in the item identification task – but rather genuine noise in the late stages of the statistical learning. Without being exhaustive, we can hypothesize for example greater computational precision, less over/underestimation of small/large probabilities or application of accurate heuristics that marginally approximate Bayesian computations.

It should be noted that the negative interaction reported in Figure 1B seems to be inconsistent with the modeling results: K was higher for musicians, which intuitively predicts a larger effect of musicianship for higher order statistics. However, model simulations reveal that this inconsistency is solved if we take into account that the selection noise is changing as well. Indeed, a high selection noise reduces the difference in performance between low and high-order statistics – it "flattens" the line. As a consequence, the group that estimates low-order statistics with a high selection noise has less difference between high- and low-order sequences than a group that estimates high-order statistics with a low selection noise. Overall, a higher K combined with a lower selection noise predicts an interaction with a negative sign (see Fig. Supp. 2). This is what we observe in the data.

The P300 amplitude is more strongly correlated to model surprise in musicians relative to non-musicians.

We then used the model fitted on the behavioral data to shed light on the brain responses. Following previous work (Regnault et al. 2001; Koelsch et al. 2002, 2007; Mars et al. 2008; Jentschke and Koelsch 2009; Koelsch 2009a; Chase et al. 2011; Kim et al. 2011; Maheu et al. 2019), we hypothesized that brain signals linearly scale with the level of theoretical surprise. We relied on information theory (Shannon 1948) to formally define theoretical surprise as the negative log probability under the model M that, given a context, the forthcoming item will be a given item. This quantity corresponds to the intuitive notion of surprise: it is low when the item is expected and high when unexpected. We defined M as our model fitted on behavioral responses. Formally, it was defined as $-\log_2(P)$ where P is the posterior predictive probability of the presented item under the model M. We then associated a level of theoretical surprise with each item of each sequence for each participant. As only ~7% of items were behavioral probes, ~93% could be used to study EEG responses.

We relied on spatial filtering to reduce the dimension of the EEG dataset. Using multiway canonical correlation analysis (de Cheveigné et al. 2019), we computed spatial filters that maximize the temporal correlation between participants without diminishing inter-individual amplitude differences (see Methods, see Fig. Supp. 4). The filters are summary components (SC), ordered by explained variance. The first SC was a central (frontal positive, occipital negative) filter, with a standard auditory response topology. It explained on average 55% of the variance of the ERP (peak correlation between SC time course and ERP at Fz, 93% explained variance) and was therefore kept for the rest of the analyses. We fitted a linear regression across all items between the theoretical surprise level and the SC amplitude at each time point. The matrix of linear coefficients (number of participants x number of time points) was then submitted to a cluster permutation in time algorithm to assess its statistical significance (see Figure 3A-B). The linear regression was significant during a long-lasting late time window for musicians (210 - 320ms, $p < 10^{-3}$) and non-musicians (220 - 430ms, $p < 10^{-3}$). The associated function, topography and time window of this response coincided with a P300 response. During active processing of the sequence,

the amplitude of the P300 is therefore linearly strongly correlated with the theoretical level of surprise.

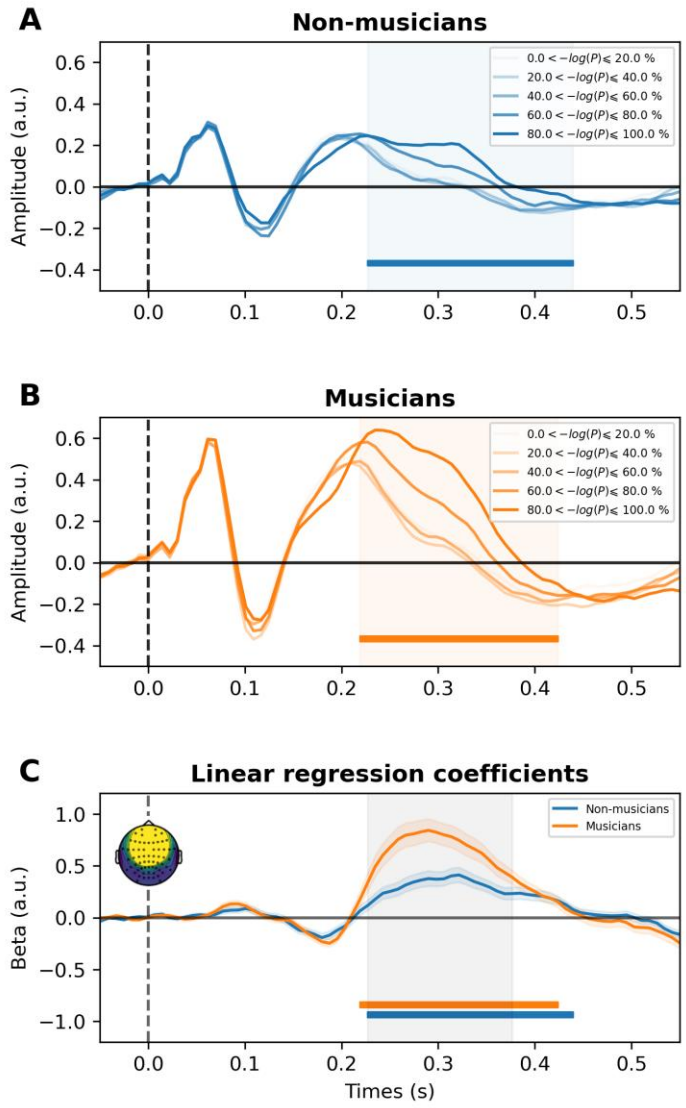

Figure 3. Model surprise correlates with single trial late auditory event related potential (ERP) amplitude. A. Average ERP of the main summary component (SC, see Methods) of non-musicians as a function of their model surprise quintiles. Model surprise was defined as $-\log_2(P)$ where P is the posterior predictive probability of the presented item under the model fitted to the participants behavioral responses. The significant time cluster is shown in blue. The linear regression was significant in a late time window, between 220 and 430 ms. **B.** Average ERP of the main SC of musicians as a function of their model surprise quintiles. The significant time cluster is shown in orange. The linear regression was significant in a late time window, between 210 and 420 ms. **C.** Coefficients of the linear regression between model surprise and SC amplitude for musicians and non-musicians. The difference was significant in the same late time window, between 220 and 380 ms. The significance of the effect (all $p < 10^{-3}$) was assessed using correction at the level of the cluster (blue and orange lines). The significance of the difference between groups ($p = 0.013$) was also cluster-corrected (grey area). Colored shaded areas represent standard error of the mean (s.e.m.). Inset plot shows the main SC topography.

We finally submitted the difference between musicians and nonmusicians coefficient matrices to a cluster permutation algorithm. This analysis revealed that the linear coefficients of the musicians were higher than the coefficients of the non-musicians in the same late time window (220 - 380ms, p = 0.013). This effect could be confounded with the overall ERP amplitude, artificially inflating the linear coefficients. Indeed, musicians had a larger root mean square (RMS) than non-musicians (= 0.67 ± 0.31, p = 0.035). This is a known phenomenon, that has been correlated to functional effect, such as better encoding of spectrally complex sounds (Regnault et al. 2001; Koelsch et al. 2002; Shahin et al. 2004, 2005; Koelsch et al. 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Kaganovich et al. 2013), and to anatomical differences, such as auditory cortex volume (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Seither-Preisler et al. 2014). In order to control for this potential confound, we normalized the ERP of each participant by the RMS of the ERP between 0 and 500 ms. This led to similar results (see Fig. Supp. 5), ensuring that the difference is not explained by the difference of overall ERP amplitude between musicians and non-musicians. The P300 amplitude is therefore modulated by the theoretical surprise to a higher degree in musicians than in non-musicians (see Figure 3C).

The ERP amplitude modulation enhancement in musicians is restricted to high-order statistical learning.

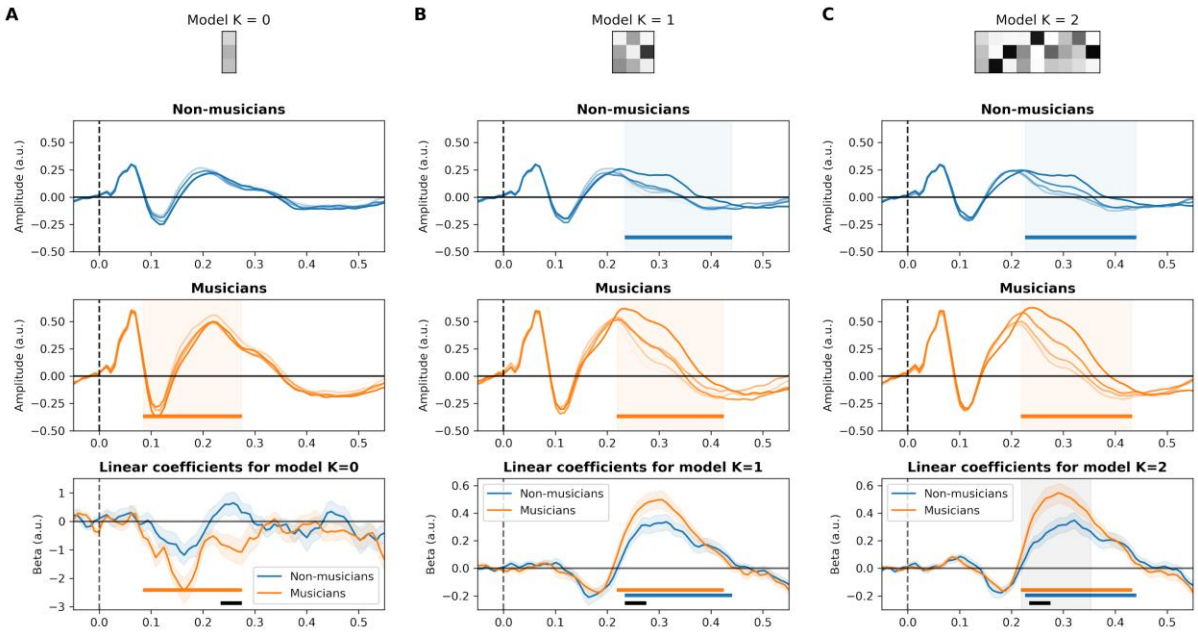

Figure 4. A cascade of low to high-order surprise correlates with the amplitude of single trial auditory event related potentials (ERP). A. Coefficients of the linear regression between surprise from the very low-order model (K=0) and SC amplitude for musicians and non-musicians. The linear regression is significant in a middle latency window for musicians, between 80 and 270 ms, and marginally significant for nonmusicians. The difference in linear coefficients between groups is not statistically significant (all clusters $p > 0.05$). **B.** Coefficients of the linear regression between the model (K=1) surprise and SC amplitude for musicians and non-musicians. The linear regression is significant in a late latency window, between 210 and 430 ms. There is no significant difference between groups (all clusters $p > 0.05$). **C.** Coefficients of the linear regression between the highest order model (K = 2) surprise and SC amplitude for musicians and non-musicians. The linear regression is significant in a late latency window, between 210 and 430 ms. The significant difference between groups ($p = 0.030$) is shown in the grey area, between 210 and 350 ms. Significance of the linear regression is shown in orange and blue (all $p < 10^{-3}$). Colored shaded areas represent standard error of the mean (s.e.m.). Significance of the interaction term Group x Model order is shown in black ($p = 0.041$). The topography of the results is the same as in Figure 3 (main SC).

The key component of the model is the estimation of the transition probability matrix. Critically, this matrix can be of any order (K), *i.e.* reflects the probability of observing an element given the preceding item (K = 1), bigram (K = 2), trigram (K = 3), or even *a priori*, without context (K = 0). This defines an ordering, from low ($K \leq 1$) to high-order predictions (K = 2). Modelling results on the behavioral data suggest that musicians compute higher order statistics. Unfortunately, behavioral data only reflect an aggregate of the neural processes. On the contrary, EEG data can give access to covert computations. We relied on the model to analyze the temporal structure of the EEG response, and uncover the succession of covert computations. We fitted the parameters □ and □ while fixing the order parameter K. This allowed us to define a level of theoretical surprise to each item, each sequence, each participant, for very low (K = 0), low (K = 1) and high-order statistics (K = 2). We then performed the same linear regression followed by cluster permutation analysis.

The regression with very low-order statistics (K = 0, see Figure 4A) showed a modulation peaking around 150 ms, significant for musicians and marginally significant for nonmusicians. It should be noted that even though when looking at the entire sequence the three items have the same frequency of occurrence, it is possible to look locally for surprising events. The topology and time window of the modulation was consistent with a

MMN (Näätänen 1995). This is coherent with the fact that a “K = 0” model is actually estimating the probability of an item irrespective of its immediately preceding context, *i.e.* the overall frequency of occurrence of this element. Indeed, MMN are typically elicited in an oddball paradigm, wherein the frequency of occurrence of items is manipulated (frequent *versus* rare). By contrast, the regression with higher order statistics revealed a significant correlation with the ERP amplitude later in time, between ~200 and ~400 ms for both "K = 1" (see Figure 4B) and "K = 2" (see Figure 4C) models. Group contrast analysis revealed that this modulation was larger for musicians compared to non musicians only for the highest model level (K = 2, p = 0.030). The interaction term Group x Model order revealed that this group difference was specific to the “K = 2” model. Using ERP normalized by the overall RMS led to similar results (see Fig. Supp. 6). In a nutshell, the very low statistics model correlates with ERP amplitude around 200 ms, similar to an MMN, and this correlation is similar in musicians and non-musicians. By contrast, higher order models correlate around 300 ms, similar to a P300, and this correlation is higher in musicians compared to non-musicians.

Discussion.

We presented auditory sequences drawn from a vocabulary of size 3 (glass, wood and metal) to a group of musicians and a group of non-musicians. We designed two types of sequences that embed either low or high-order statistics corresponding to 1st order and 2nd order Markov chains, respectively. Each sequence contained probes requiring participants to explicitly predict the most likely future item. In this task, musicians make more accurate predictions than non-musicians. This result is not explained by a sensory advantage, such as a better ability to discriminate or identify auditory targets. Indeed, the stimuli were chosen to be easily and unambiguously identifiable. Computational modelling further reveals that this advantage is best explained in terms of parameters governing the order of the Markov chain model and the selection noise, and no significant differences were revealed for the parameter governing the memory decay. EEG recordings during behaviorally unprobed items allow bridging modeling and electrophysiological signatures of the behavioral task. First, the amplitude of a central frontal cluster at 300 ms is strongly correlated on a single trial basis to the computationally modeled theoretical surprise. Second, this P300 model-based modulation is stronger for musicians than non-musicians, suggesting a difference in sensitivity to the probabilistic structure of the sequence. Last, neural responses to surprise with a low-order statistical structure ($K \leq 1$) are not statistically different between musicians and non-musicians, and diverge only for surprise with high-order statistical structures ($K = 2$). We conclude from these results that musicians have improved neural SL in the auditory domain compared to non-musicians.

Our results are relevant in the debate on musical training induced plasticity. It has been known for a long time that musical training is associated with low level sensory improvements, such as pitch or duration detection thresholds (Spiegel and Watson 1984; Regnault et al. 2001; Koelsch et al. 2002; Micheyl et al. 2006; Koelsch et al. 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Kuman et al. 2014). However, music expertise does not only require fine perception of isolated pitches and durations but it also puts high demands in terms of sequence processing (melody and harmony) that require both accurate temporal and spectral prediction (Patel 2011). In the current experimental design, we chose to control sensory-related gain biases by using large acoustic differences between stimuli. Thus, observed differences between musicians and non-musicians cannot be due to low level sensory differences. Another potential confound is that musicians may have a higher level of attention during the task. However, we observe a larger modulation of the P300 amplitude by model surprise in musicians compared to non-musicians. Attentional differences would mostly result in global amplitude differences and a main effect of group only. Moreover, we analyzed the ERP in the statistical learning task in response to the visual probes. This supplementary analysis revealed no amplitude nor latency differences between musicians and non-musicians (see Fig. Supp. 7). This suggests that the two groups were equally surprised by the visual probes, and thus were similarly attending the sounds. Finally, the fact that normalizing data by the global RMS does not change our results goes well in line with the fact that attentional differences, if they exist, do not fully account our results. By contrast and critically, the model captures differences related to downstream computations, directly involved in the inference process itself. It is important to note that the model not only captures the surprise associated with violations of the internal expectations but it also describes the continuous learning by constantly updating predictions as a function of the

context. Hence, the advantage of musicians over non-musicians does not solely rely on a greater sensitivity to predictions errors, as shown by the larger P300 modulation, it also reflects a higher order and more accurate continuous learning.

In the current experiment, we have chosen an explicit task with deliberately included probe trials to engage participants actively as well as to collect behavioural responses. Participants were active, aware of the unpredictable nature of the sequences, and explicitly doing a prediction task. This differs from more implicit tasks, where learning is inferred from indirect measures such as a reduction of reaction times in a serial reaction time task (Nissen and Bullemer 1987). There is a long standing debate concerning the distinction between implicit and explicit learning (Regnault et al. 2001; Cleeremans and Jiménez 2002; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Kim et al. 2009; Koelsch 2009a; Kim et al. 2011; Dale et al. 2012; Batterink et al. 2015). While our aim was not to address this question, our results can be compared with other studies using a similar approach. First, the linear relation between the theoretical level of surprise and the amplitude of the P300 has been reported in both implicit (Mars et al. 2008) and explicit prediction tasks (Maheu et al. 2019). Second, studies on music listening have reported that early EEG components, such as the N100 amplitude, are modulated by acoustic aspects of the signal while later EEG components, such as the MMN, early right anterior negativity (ERAN) and P300, are related to the formation of musical expectations. This was true in both passive (Koelsch 2009b; Vuust et al. 2012; Di Liberto et al. 2020; Quiroga-Martinez et al. 2020) and active (Koelsch et al. 2000; Omigie et al. 2013) listening conditions. Third, the ~300 ms latency and the fronto-central topography that we report indicate a probable modulation of the P3a subcomponent of the P300, which is earlier and more frontal than the P3b. P3a has been referred to as novelty P300 and possibly reflecting a rather automatic orientation of attention to unexpected context changes (Polich 2007). This is consistent with the recent study of (Quiroga-Martinez et al. 2020) that reported the same modulation of the P3a in musicians in a passive listening paradigm. Last, our results are also consistent with two studies that have shown an advantage of musicians over non-musicians in music listening (Hansen and Pearce 2014) and of jazz-specific expertise over classical-music expertise in jazz listening (Hansen et al. 2016) in a task of explicit uncertainty rating but not in a task of inferred uncertainty rating. These mixed results can be interpreted in the context of a recent proposal (Conway 2020) that suggested the existence of two interdependent systems for (statistical) learning. First, a "suite" of multiple automatic subsystems responsible for the learning of "simple" statistical regularities. Second, a central, attention-dependent system responsible for the learning of "complex" statistical regularities such as long term dependencies, and for the gating and control of the automatic subsystems. In this context, our results suggest that the learning of very low (K = 0) and low-order (K = 1) statistics depend on automatic subsystems, with no advantage of musicians over non-musicians. On the contrary, the learning of high-order statistics (K = 2) would depend on top-down control (Koelsch et al. 2019), with an advantage of musicians over non-musicians. Following (Conway 2020), this would in turn suggest that the advantage of musicians over non-musicians could possibly transfer to other cognitive functions, such as language, as the central system is thought of as a hub involved in multiple brain networks.

Our results also replicate and integrate in the Bayesian framework the notion of "abstract" MMN (Paavilainen 2013), *i.e.* the sensitivity to regularities beyond the mere frequency of appearance. Studies have demonstrated that musicians have larger MMN in

response to changes in melodic contours (Regnault et al. 2001; Tervaniemi et al. 2001; Koelsch et al. 2002, 2007; Fujioka et al. 2004; Jentschke and Koelsch 2009; Koelsch 2009a; Herholz et al. 2011; Kim et al. 2011; Paraskevopoulos et al. 2012) (Tervaniemi et al. 2001; Fujioka et al. 2004; Herholz et al. 2011; Paraskevopoulos et al. 2012), while having similar MMN in response to simple change of pitch (Fujioka et al. 2004). Similar findings have also been described for chord sequences (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011). Interestingly, changes of pitch concern the frequency of occurrence of items, *i.e.* K = 0 Markov chains, whereas changes of melody concerns the ordered structure, which is typically captured by a K = 1 or K = 2 Markov chain. We extend these results by using a task that does not rely on the pitch dimension, thus ensuring that sensory processing is not the origin of the musical expertise advantage. Furthermore, the model we developed sheds light on these results by drawing an explicit line between low-order statistics (frequency of occurrence, $K \leq 1$) and high-order ones ($K = 2$). Anecdotally, the musical advantage of high-order statistics over low-order ones is also present in music automatic generation, where high-order Markov chains "generate results with a sense of phrasal structure, rather than the 'aimless wandering' produced by a first-order system" (Roads and Strawn 1996) (see Iannis Xenakis' Analogique A and B for an example music generated by 1st order Markov chains).

Our model combines two approaches: an "ideal observer" and an "individualized modelling" strategy. This ideal observer allows defining a theoretical maximum on the probabilistic task, here defining performance bounds between a random (33%) and an optimal strategy (80%). However, ideal observer models are usually universal, and as such are unable to characterize inter-individual variability. By contrast, the added parameters included in our model precisely specify the multiple ways of being suboptimal and allow testing clear and separable hypotheses about different suboptimality sources. Using this combined approach, we reveal that humans do not accurately estimate 1st and 2nd order Markov statistics and provide insight on the possible limitations of this suboptimal estimation. If (Bayesian) optimality has been standard in psychophysics, *e.g.* in visual orientation discrimination (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Girshick et al. 2011; Kim et al. 2011) or multisensory integration (Ernst and Banks 2002) tasks, suboptimality is almost systematically observed in higher level cognitive tasks, such as discrete evidence accumulation (Drugowitsch et al. 2016) or explicit manipulation of probabilities (Kahneman and Tversky 1977). The errors are usually explained in terms of task-independent noise in sensory processing (Regnault et al. 2001; Koelsch et al. 2002, 2007; Osborne et al. 2005; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Brunton et al. 2013; Kaufman and Churchland 2013) or noise in the response selection, following the decision (Sutton and Barto 1998). However, recent proposals have put emphasis on limitations in the inference process itself (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Acerbi et al. 2014; Dayan 2014; Drugowitsch et al. 2016), arising from systematic biases (Beck et al. 2012) or from variability in the computational precision of variables represented in populations of neurons (Renart and Machens 2014). Our results are in line with these recent proposals. Indeed, the key parameters to separate musicians and non-musicians are the order of the estimated statistics (K) and the selection noise (β) after this computation. The interpretation of the parameter K is straightforward: the musician's estimates resemble high-order statistics more than non-musician's estimates. The selection

noise parameter □ is more subject to interpretation. We isolated three interpretations. First, it can reflect a loss of information in the inference process itself, due to imperfect computations. Second, it can reflect an imperfect transformation of statistical estimates into choices. This could depend on the task and on the type of ratings asked to the participants. For example, (Hansen and Pearce 2014) suggest that the IDyOM model is more accurate to account for inferred uncertainty rating than for explicit uncertainty ratings. Last, the parameter □ can also capture systematic deviations of the human behavior from the model's predictions. Indeed, the model we designed is limited in its ability to capture the full range of human sequence learning abilities. For example, humans learn syntactic and supra-regular rules (Fitch 2014; Dehaene et al. 2015) that cannot be captured by transition probabilities. It is also possible that humans learn higher-order Markov statistics. For example, it has been shown that 4-grams best approximate musicians' ratings to chord entropy (Hansen and Pearce 2014). Even though these other strategies would not result in improved behavioral accuracy, they could nonetheless be used by humans. Such deviations from the Markov strategy would all lead to an increase in the □ parameter. Further work is needed to precise how to interpret the difference in □ between musicians and non-musicians.

Finally, this study is cross-sectional, therefore caution should be taken when interpreting causal effects (Schellenberg 2019). A first interpretation is that good SL predispose individuals to pursue musical training. Indeed, large-scale twin studies have documented a genetic component to musical skill and extent of practicing (Mosing et al. 2014; Hambrick and Tucker-Drob 2015), suggesting that brain structure and function might predispose one to pursue musical training. This set of predispositions remains unknown. In light of the present results, we formulate the hypothesis that these genetic predispositions could directly concern SL abilities or a prerequisite for SL. A second interpretation is that other uncontrolled confounding factors, such as socio-economic status or levels of education, increase the probability of both pursuing musical training and having good SL abilities. We have tried to limit this by recruiting all participants at the university to increase the homogeneity between groups. A third interpretation is that musical training improves SL abilities. By providing a rich statistical structure and focusing attention on this structure, music could lead to beneficial effects on SL abilities in the auditory domain. Indeed, causal studies have previously demonstrated that musical training is effective for speech segmentation based on SL (François et al. 2013), for rehabilitation of reading and phonological skills (Flaugnacco et al. 2015) in children with dyslexia, turn taking in children with cochlear implant (Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Hidalgo et al. 2017, 2019) as well as in several neurological disorders (for a review, see (Sihvonen et al. 2017)). One open question relevant for the development of music-based remediation is whether different types of musical experience might provide different advantages, with more or less focus on the statistical structure. In our study, the musicians were practicing classical music, but it has been suggested for example that jazz players have larger responses to surprising auditory events than other musicians (Vuust et al. 2012). Overall, as a fundamental computation for any statistical structure representation and since statistical structuring is most often impaired, SL rehabilitation could possibly be at the core of the music remediation power.

Supplementary Methods.

DERIVATION OF THE MODEL.

The goal of the model is to infer the probability of each item given the preceding context. Formally, the model is exposed to a sequence of T items $s_{0:T-1}$ taken from a vocabulary Ω of size V. The context is given by the last K items of the sequence. As a Bayesian ideal observer, she uses Bayes rule to update her belief:

$$P(\theta|s_{0:T-1}) \propto P(s_{0:T-1}|\theta)P(\theta)$$

We can decompose the likelihood term using the chain rule :

$$P(s_{0:T-1}|\theta) = P(s_0|\theta)P(s_1|s_0,\theta)\dots P(s_{T-1}|s_0,s_1,\dots,s_{T-2},\theta)$$

Restricted case K = 1, v = 2

For simplicity, let's first suppose a restricted case in which K = 1, *i.e.* the sequence is a Markov chain of order K = 1, and V = 2, *i.e.* there are only two items: A and B. The derivation is inspired from (Brooks et al. 1996; Regnault et al. 2001; Koelsch et al. 2002, 2007; Jentschke and Koelsch 2009; Koelsch 2009a; Kim et al. 2011; Meyniel et al. 2016). The likelihood of a given observation depends only on the estimated transition probabilities and the previous item:

$$P(s_{0:T-1}|\theta) = P(s_0|\theta)\prod_{t=1}^{T-1} P(s_t|s_{t-1},\theta)$$

Assuming that for the first observation $P(s_0|\square) = 0.5$, we have:

$$P(s_{0:T-1}|\theta) = \frac{1}{2}\theta_{0,0}^{N_{A|A}}\theta_{0,1}^{N_{A|B}}\theta_{1,0}^{N_{B|A}}\theta_{1,1}^{N_{B|B}}$$

Where $\square$ denotes the “transition probability matrix” of order 1. It is a matrix of size 2 x 2 where each row designates a particular item (A or B) and each column a particular context item (A or B). $N_{A|B}$ designates the number of occurence of the bigram “BA” in the sequence, *i.e.* the number of times that the context “B” was followed by the item “A”. As each column sum to one, we can rewrite the equation as:

$$P(s_{0:T-1}|\theta) = \frac{1}{2}(\theta_{1,0}^{N_{B|A}}(1-\theta_{1,0})^{N_{A|A}})(\theta_{0,1}^{N_{A|B}}(1-\theta_{0,1})^{N_{B|B}})$$

A nice conjugate prior for this likelihood equation is the Beta distribution as the product of two Beta distributions is also a Beta distribution. Combined with a uniform prior distribution Beta(1, 1), the posterior distribution on $\square$ is therefore the product of two Beta distributions with parameters corresponding to the transition counts plus one.

$$P(\theta|s_{0:T-1}) = Beta(\theta_{1,0}|N_{B|A}+1, N_{A|A}+1)Beta(\theta_{0,1}|N_{A|B}+1, N_{B|B}+1)$$

General case.

The general case is similar to the restricted case. The main extension relies on the fact that the Dirichlet distribution generalizes the Beta distribution, with more than two parameters. The only difference will therefore be that the posterior is not any more a product of two Beta distributions but a product of V^K Dirichlet distributions with V parameters. As previously, the likelihood of a given observation depends only on the estimated transition probabilities and the previous K items:

$$P(s_{0:T-1}|\theta) = P(s_0|\theta)P(s_1|s_0,\theta)(s_{K-1}|s_0,s_1,\ldots,s_{K-2},\theta)\prod_{t=K}^{T-1}P(s_t|s_{t-K:t-1},\theta)$$

$$P(s_{0:T-1}|\theta) = \frac{1}{V^K}\prod_{v=1}^{V}\prod_{j=1}^{V^K}(\theta_{vj})^{N_{vj}}$$

Where □ denotes the "transition probability matrix" of order K. It is a matrix of size V x VK where each row designates a particular item and each column a particular K-uplets of items. N_{ij} designates the number of occurrences of the (K+1)-uplet corresponding to the cell (i, j) of the matrix (the j^{th} K-uplet followed by the i^{th} item). For simplicity, the first K observations are considered arbitrary such that $P(s_0) = P(s_1) = \ldots = P(s_{K-1}) = 1/V$. As each column sums to one, the derived likelihood corresponds to the product of V^K Dirichlet distributions. Combined with a uniform joint prior distribution Dir(1, 1, …, 1), the posterior distribution therefore results in a Dirichlet distribution with parameters corresponding to the transition counts N_{ij} plus one:

$$P(\theta|s_{0:T-1}) = \prod_{j=1}^{V^K} Dir(\theta_0,\theta_1,\ldots,\theta_V|N_{0,j}+1,N_{1,j}+1,N_{2,j}+1,\ldots,N_{V,j}+1)$$

The posterior distribution can then be turned into the likelihood of the next stimulus using Bayes' rule:

$$P(s_T|s_{0:T-1}) = \int P(s_T|\theta,s_{T-K:T-1})P(\theta|s_{0:T-1})d\theta$$

Which can be analytically solved and ends up being simply:

$$P(s_T = x_i|s_{0:T-1}) = \frac{N_{ij}+1}{\sum_{v=1}^{V}(N_{vj}+1)}$$

Where $P(s_t = x_i|s_{0:t-1})$ is the probability of the element $x_i \in \Omega$. It corresponds to the ratio of two scalars:

- N_{ij}+1: the number of times the (K+1)-uplet $s_{T-K:T-1}x_i$ has been observed plus one,
- $\sum_{v=1}$ (N_{vj}+1): the number of times that the K-uplet context $s_{T-K:T-1}$ has been observed plus V.

Supplementary Figures.

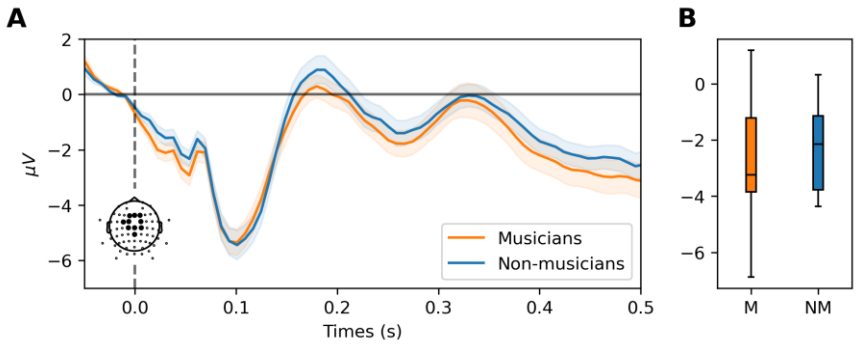

Figure Supp. 1. Auditory ERP in the familiarization task. **A.** Musicians and non-musicians have similar ERP in the central cluster (10 most activated electrodes between 50 and 150 ms, highlighted in the inset plot). Cluster permutations did not detect any difference (all cluster-corrected p-values > 0.05). **B.** Boxplot of the average activity of these central electrodes between 50 and 150 ms.

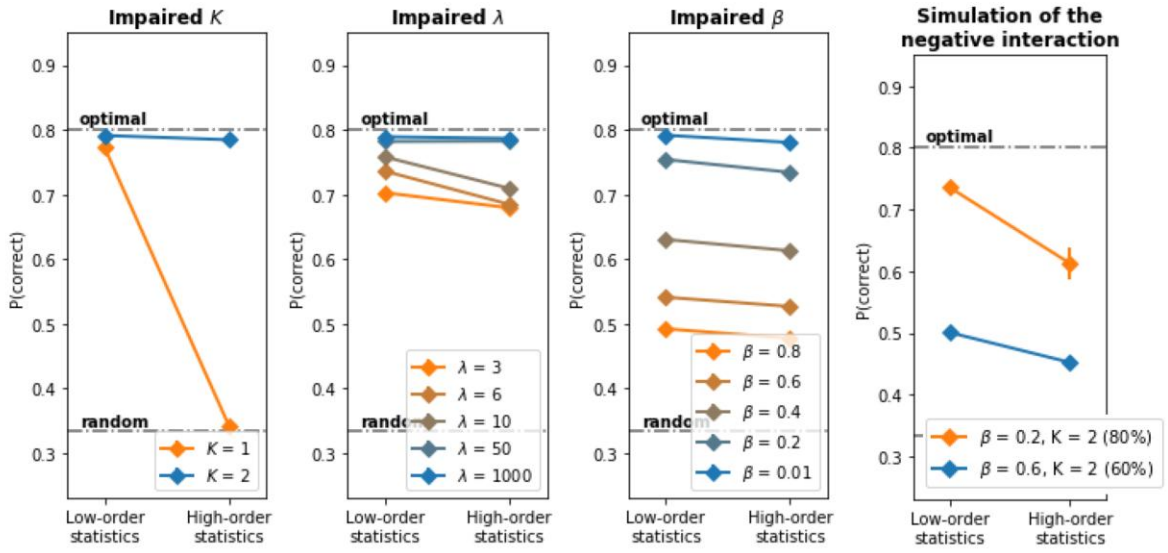

Figure Supp. 2. Model simulations. Performances can be impaired for multiple reasons. The model captures three sources of imperfections: low-order statistics (K), short memory (□) and high selection noise (β). The majority of participants have a value of K of 1 or 2, a value of □ between 5 and 50 and a value of β between 0.3 and 0.8. Within this range, each parameter has an impact on the performances. The last panel is showing the negative interaction. This pattern emerges when the two groups differ in proportion of K = 2 agents (here 80% orange vs 60% blue) and in selection noise β (here 0.2 orange vs 0.6 blue).

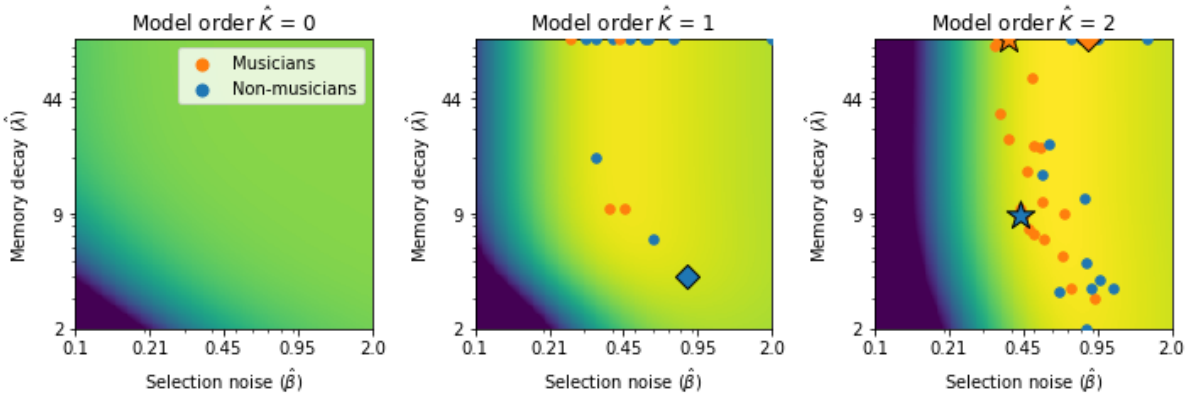

Figure Supp. 3. Maximum-likelihood fitting. Average log-likelihood of the data as a function of model parameters. Yellow indicates higher log-likelihood. Colored dots represent individual participants. Diamonds represent participants with the worst behavioral performance of each group. On the contrary, stars represent participants with the best behavioral performance of each group. Participants with good performance tend to have a low selection noise (β) and high model order (K).

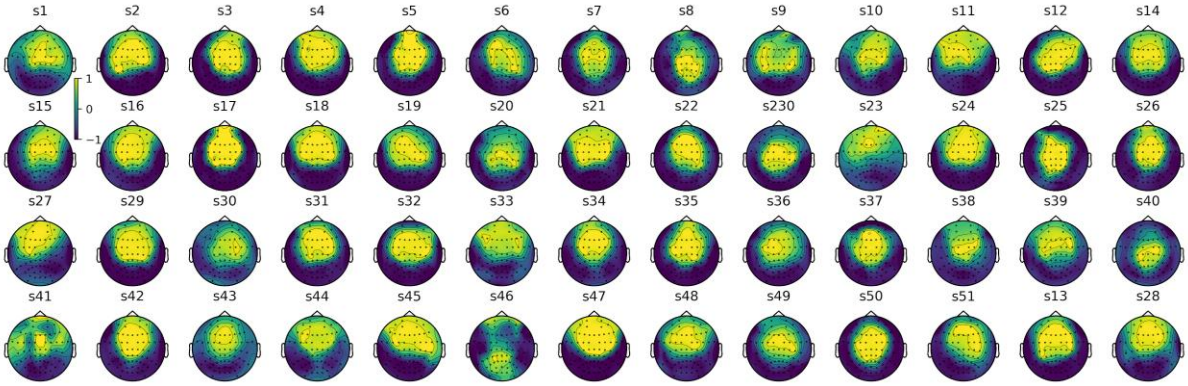

Figure Supp. 4. Topographies of the Summary Component (SC) #1 computed by the Multiway Canonical Correlation Analysis. Each topography represents the correlation between one individual SC #1 time course and each electrode time course. High correlation (yellow) indicates that the electrodes contributes strongly to the SC #1 time course.

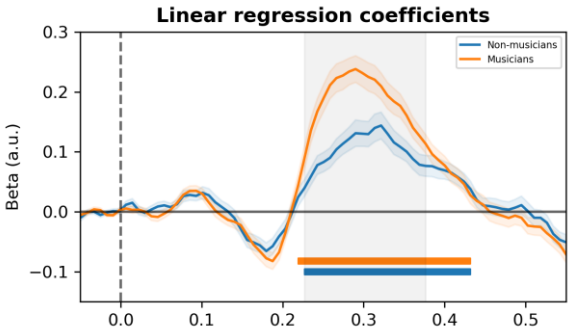

Figure Supp. 5. Model surprise correlates with single trial late auditory event related potential (ERP) amplitude. The ERP amplitude has been normalized on an individual basis by the root mean square of the ERP between 0 and 500 ms. Coefficients of the linear regression between model surprise and SC amplitude for musicians and non-musicians. The difference is significant in the same late time window, between 220 and 380 ms. Significance of the effect (all $p < 10^{-3}$) was assessed using correction at the level of the cluster (blue and orange lines). Significance of the difference between groups ($p = 0.014$) was also cluster-corrected (grey area). Colored shaded areas represent standard error of the mean (s.e.m.).

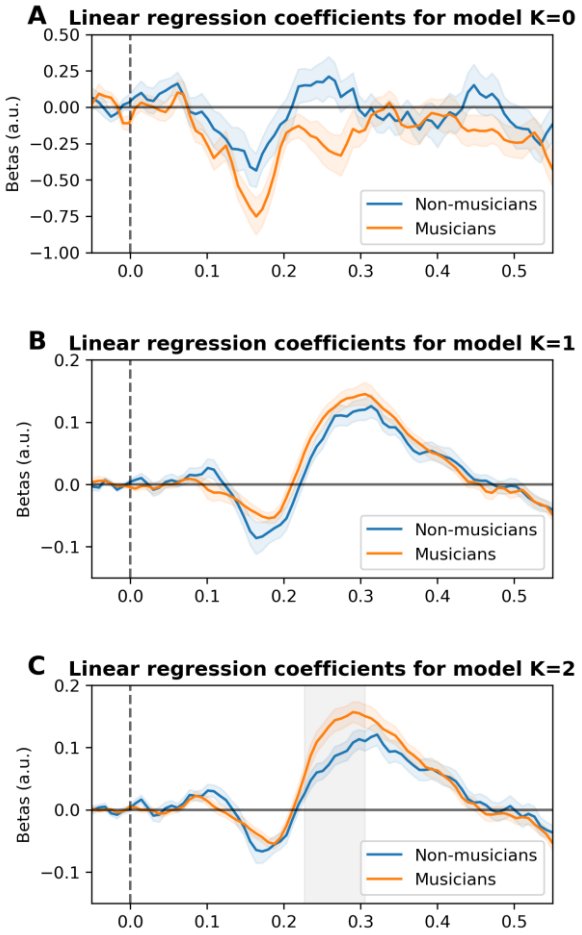

Figure Supp. 6. A cascade of low to high-order surprise correlates with the amplitude of single trial auditory event related potentials (ERP). The ERP amplitude has been normalized on an individual basis by the root mean square of the ERP between 0 and 500 ms. **A.** Coefficients of the linear regression between surprise from the very low-order model (K=0) and SC amplitude for musicians and non-musicians. The linear regression is significant in a middle latency window for musicians only, between 80 and 270 ms. There is no significant difference between groups (all clusters $p > 0.05$). **B.** Coefficients of the linear regression between the model (K=1) surprise and SC amplitude for musicians and non-musicians. The linear regression is significant in a late latency window, between 210 and 430 ms. There is no significant difference between groups (all clusters $p > 0.05$). **C.** Coefficients of the linear regression between the highest order model (K = 2) surprise and SC amplitude for musicians and non-musicians.

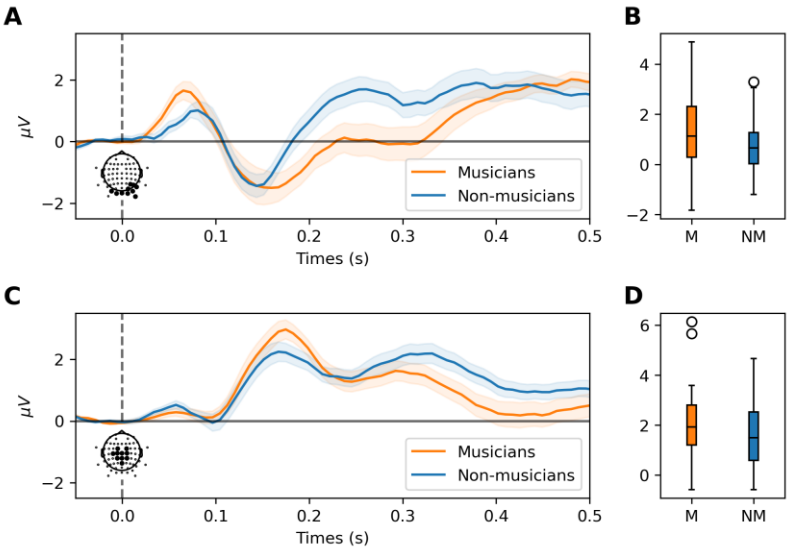

Figure Supp. 7. ERP in the statistical learning task in response to the visual probes. **A.** Musicians and non-musicians have similar ERP in the occipital cluster (10 most activated electrodes between 50 and 100 ms, highlighted in the inset plot). Cluster permutations did not detect any difference (all cluster-corrected p-values > 0.05). **B.** Boxplot of the average activity of these occipital electrodes between 50 and 100 ms. **C.** Musicians and non-musicians have similar ERP in the central cluster (10 most activated electrodes between 110 and 220 ms, highlighted in the inset plot). Cluster permutations did not detect any difference (all cluster-corrected p-values > 0.05). **D.** Boxplot of the average activity of these central electrodes between 110 and 220 ms.

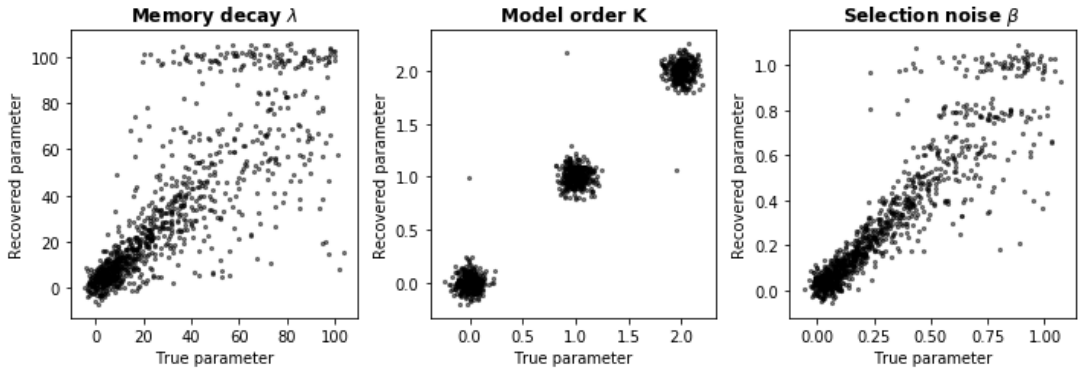

Figure Supp. 8. Parameter recovery analysis. We generated synthetic data for 10^3 models with random parameters (exponential between 0 and 100, K, uniform choice between 0, 1 and 2, β exponential between 0 and 1). Recovered parameters from these synthetic data using our procedure were close to the original parameters (spearman ρ = 0.91, ρ_K= 0.99, ρ_β = 0.96).

Bibliography.

Acerbi L, Vijayakumar S, Wolpert DM. 2014. On the origins of suboptimality in human probabilistic inference. PLoS Comput Biol. 10:e1003661.

Aramaki M, Kronland-Martinet R, Voinier T, Ystad S. 2006. A percussive sound synthesizer based on physical and perceptual attributes. Computer Music Journal. 30:32–41.

Bates D, Mächler M, Bolker B, Walker S. 2014. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823.

Batterink LJ, Reber PJ, Neville HJ, Paller KA. 2015. Implicit and explicit contributions to statistical learning. J Mem Lang. 83:62–78.

Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A. 2012. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron. 74:30–39.

Brooks S, Gelman A, Carlin JB, Stern HS, Rubin DB. 1996. Bayesian Data Analysis. The Statistician. 45:266.

Brunton BW, Botvinick MM, Brody CD. 2013. Rats and humans can optimally accumulate evidence for decision-making. Science. 340:95–98.

Chase HW, Swainson R, Durham L, Benham L, Cools R. 2011. Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. J Cogn Neurosci. 23:936–946.

Chen SF, Goodman J. 1999. An empirical study of smoothing techniques for language modeling. Comput Speech Lang. 13:359–393.

Cleeremans A, Jiménez L. 2002. Implicit learning and consciousness: A graded, dynamic perspective. Implicit learning and consciousness.

Conway CM. 2020. How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. Neurosci Biobehav Rev. 112:279–299.

Daikoku T. 2018. Neurophysiological markers of statistical learning in music and language: hierarchy, entropy, and uncertainty. Brain Sci. 8.

Dale R, Duran N, Morehead R. 2012. Prediction during statistical learning, and implications for the implicit/explicit divide. ACP. 8:196–209.

Dayan P. 2014. Rationalizable irrationalities of choice. Top Cogn Sci. 6:204–228.

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.

Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C. 2015. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. Neuron. 88:2–19.

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

Drugowitsch J, Wyart V, Devauchelle A-D, Koechlin E. 2016. Computational precision of mental inference as critical source of human choice suboptimality. Neuron. 92:1398–1411.

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 415:429–433.

Fitch WT. 2014. Toward a computational framework for cognitive biology: unifying approaches from cognitive neuroscience and comparative cognition. Phys Life Rev.

11:329–364.
Flaugnacco E, Lopez L, Terribili C, Montico M, Zoia S, Schön D. 2015. Music training increases phonological awareness and reading skills in developmental dyslexia: A randomized control trial. PLoS ONE. 10:e0138715.
François C, Chobert J, Besson M, Schön D. 2013. Music training for the development of speech segmentation. Cereb Cortex. 23:2038–2043.
Francois C, Schön D. 2011. Musical expertise boosts implicit learning of both musical and linguistic structures. Cereb Cortex. 21:2357–2365.
François C, Tillmann B, Schön D. 2012. Cognitive and methodological considerations on the effects of musical expertise on speech segmentation. Ann N Y Acad Sci. 1252:108–115.
Fujioka T, Trainor LJ, Ross B, Kakigi R, Pantev C. 2004. Musical training enhances automatic encoding of melodic contour and interval structure. J Cogn Neurosci. 16:1010–1021.
Girshick AR, Landy MS, Simoncelli EP. 2011. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat Neurosci. 14:926–932.
Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M. 2013. MEG and EEG data analysis with MNE-Python. Front Neurosci. 7:267.
Hambrick DZ, Tucker-Drob EM. 2015. The genetics of music accomplishment: evidence for gene-environment correlation and interaction. Psychon Bull Rev. 22:112–120.
Hansen NC, Pearce MT. 2014. Predictive uncertainty in auditory sequence processing. Front Psychol. 5:1052.
Hansen NC, Vuust P, Pearce M. 2016. “if you have to ask, you’ll never know”: effects of specialised stylistic expertise on predictive processing of music. PLoS ONE. 11:e0163584.
Harrison PMC, Bianco R, Chait M, Pearce MT. 2020. PPM-Decay: A computational model of auditory prediction with memory decay. PLoS Comput Biol. 16:e1008304.
Herholz SC, Boh B, Pantev C. 2011. Musical training modulates encoding of higher-order regularities in the auditory cortex. Eur J Neurosci. 34:524–529.
Hidalgo C, Falk S, Schön D. 2017. Speak on time! Effects of a musical rhythmic training on children with hearing loss. Hear Res. 351:11–18.
Hidalgo C, Pesnot-Lerousseau J, Marquis P, Roman S, Schön D. 2019. Rhythmic training improves temporal anticipation and adaptation abilities in children with hearing loss during verbal interaction. J Speech Lang Hear Res. 62:3234–3247.
Jentschke S, Koelsch S. 2009. Musical training modulates the development of syntax processing in children. Neuroimage. 47:735–744.
Kaganovich N, Kim J, Herring C, Schumaker J, Macpherson M, Weber-Fox C. 2013. Musicians show general enhancement of complex sound encoding and better inhibition of irrelevant auditory change in music: an ERP study. Eur J Neurosci. 37:1295–1307.
Kahneman D, Tversky A. 1977. Prospect theory. an analysis of decision making under risk. US Dept of the Navy.
Kaufman MT, Churchland AK. 2013. Cognitive neuroscience: sensory noise drives bad decisions. Nature. 496:172–173.
Kim R, Seitz A, Feenstra H, Shams L. 2009. Testing assumptions of statistical learning: is it long-term and implicit? Neurosci Lett. 461:145–149.
Kim S-G, Kim JS, Chung CK. 2011. The effect of conditional probability of chord progression on brain response: an MEG study. PLoS ONE. 6:e17337.
Koelsch S. 2009a. Music-syntactic processing and auditory memory: similarities and

differences between ERAN and MMN. Psychophysiology. 46:179–190.
Koelsch S. 2009b. Neural substrates of processing syntax and semantics in music. In: Haas R,, Brandes V, editors. Music that works. Vienna: Springer Vienna. p. 143–153.
Koelsch S, Gunter T, Friederici AD, Schröger E. 2000. Brain indices of music processing: “nonmusicians” are musical. J Cogn Neurosci. 12:520–541.
Koelsch S, Jentschke S. 2008. Short-term effects of processing musical syntax: an ERP study. Brain Res. 1212:55–62.
Koelsch S, Jentschke S, Sammler D, Mietchen D. 2007. Untangling syntactic and sensory processing: an ERP study of music perception. Psychophysiology. 44:476–490.
Koelsch S, Sammler D. 2008. Cognitive components of regularity processing in the auditory domain. PLoS ONE. 3:e2650.
Koelsch S, Schmidt B-H, Kansok J. 2002. Effects of musical expertise on the early right anterior negativity: an event-related brain potential study. Psychophysiology. 39:657–663.
Koelsch S, Vuust P, Friston K. 2019. Predictive processes and the peculiar case of music. Trends Cogn Sci (Regul Ed). 23:63–77.
Kolossa A, Fingscheidt T, Wessel K, Kopp B. 2012. A model-based approach to trial-by-trial p300 amplitude fluctuations. Front Hum Neurosci. 6:359.
Kuhl PK. 2004. Early language acquisition: cracking the speech code. Nat Rev Neurosci. 5:831–843.
Kuman PV, Rana B, Krishna R. 2014. Temporal processing in musicians and non-musicians. J Hear Sci.
Loui P, Wessel DL, Hudson Kam CL. 2010. Humans Rapidly Learn Grammatical Structure in a New Musical Scale. Music Percept. 27:377–388.
Maheu M, Dehaene S, Meyniel F. 2019. Brain signatures of a multiscale process of sequence learning in humans. elife. 8.
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.
Meyniel F, Maheu M, Dehaene S. 2016. Human Inferences about Sequences: A Minimal Transition Probability Model. PLoS Comput Biol. 12:e1005260.
Micheyl C, Delhommeau K, Perrot X, Oxenham AJ. 2006. Influence of musical and psychoacoustical training on pitch discrimination. Hear Res. 219:36–47.
Miranda RA, Ullman MT. 2007. Double dissociation between rules and memory in music: an event-related potential study. Neuroimage. 38:331–345.
Mosing MA, Pedersen NL, Madison G, Ullén F. 2014. Genetic pleiotropy explains associations between musical auditory discrimination and intelligence. PLoS ONE. 9:e113874.
Näätänen R. 1995. The mismatch negativity: a powerful tool for cognitive neuroscience. Ear Hear. 16:6–18.
Nissen MJ, Bullemer P. 1987. Attentional requirements of learning: Evidence from performance measures. Cogn Psychol. 19:1–32.
Omigie D, Pearce MT, Williamson VJ, Stewart L. 2013. Electrophysiological correlates of melodic processing in congenital amusia. Neuropsychologia. 51:1749–1762.
Osborne LC, Lisberger SG, Bialek W. 2005. A sensory source for motor variation. Nature. 437:412–416.
Paavilainen P. 2013. The mismatch-negativity (MMN) component of the auditory event-related potential to violations of abstract regularities: a review. Int J Psychophysiol.

88:109–123.
Palminteri S, Wyart V, Koechlin E. 2017. The importance of falsification in computational cognitive modeling. Trends Cogn Sci (Regul Ed). 21:425–433.
Paraskevopoulos E, Kuchenbuch A, Herholz SC, Pantev C. 2012. Musical expertise induces audiovisual integration of abstract congruency rules. J Neurosci. 32:18196–18203.
Patel AD. 2011. Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. Front Psychol. 2:142.
Pearce MT, Ruiz MH, Kapasi S, Wiggins GA, Bhattacharya J. 2010. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. Neuroimage. 50:302–313.
Pearce MT, Wiggins GA. 2012. Auditory expectation: the information dynamics of music perception and cognition. Top Cogn Sci. 4:625–652.
Perruchet P, Pacton S. 2006. Implicit learning and statistical learning: one phenomenon, two approaches. Trends Cogn Sci (Regul Ed). 10:233–238.
Polich J. 2007. Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol. 118:2128–2148.
Putkinen V, Tervaniemi M, Saarikivi K, de Vent N, Huotilainen M. 2014. Investigating the effects of musical training on functional brain development with a novel Melodic MMN paradigm. Neurobiol Learn Mem. 110:8–15.
Quiroga-Martinez DR, Hansen NC, Højlund A, Pearce M, Brattico E, Vuust P. 2020. Decomposing neural responses to melodic surprise in musicians and non-musicians: Evidence for a hierarchy of predictions in the auditory system. Neuroimage. 215:116816.
Regnault P, Bigand E, Besson M. 2001. Different brain mechanisms mediate sensitivity to sensory consonance and harmonic context: evidence from auditory event-related brain potentials. J Cogn Neurosci. 13:241–255.
Renart A, Machens CK. 2014. Variability in neural activity and behavior. Curr Opin Neurobiol. 25:211–220.
Roads C, Strawn J. 1996. The computer music tutorial.
Rohrmeier M, Rebuschat P, Cross I. 2011. Incidental and online learning of melodic structure. Conscious Cogn. 20:214–222.
Romberg AR, Saffran JR. 2010. Statistical learning and language acquisition. Wiley Interdiscip Rev Cogn Sci. 1:906–914.
Saffran JR, Aslin RN, Newport EL. 1996. Statistical learning by 8-month-old infants. Science. 274:1926–1928.
Schellenberg EG. 2019. Correlation = causation? Music training, psychology, and neuroscience. Psychol Aesthet Creat Arts.
Seither-Preisler A, Parncutt R, Schneider P. 2014. Size and synchronization of auditory cortex promotes musical, literacy, and attentional skills in children. J Neurosci. 34:10937–10949.
Shahin A, Roberts LE, Pantev C, Trainor LJ, Ross B. 2005. Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. Neuroreport. 16:1781–1785.
Shahin A, Roberts LE, Trainor LJ. 2004. Enhancement of auditory cortical development by musical experience in children. Neuroreport. 15:1917–1921.
Shannon CE. 1948. A mathematical theory of communication. Bell System Technical Journal. 27:379–423.
Siegelman N, Bogaerts L, Christiansen MH, Frost R. 2017. Towards a theory of individual

differences in statistical learning. Philos Trans R Soc Lond B Biol Sci. 372.
Siegelman N, Bogaerts L, Frost R. 2017. Measuring individual differences in statistical learning: Current pitfalls and possible solutions. Behav Res Methods. 49:418–432.
Sihvonen AJ, Särkämö T, Leo V, Tervaniemi M, Altenmüller E, Soinila S. 2017. Music-based interventions in neurological rehabilitation. Lancet Neurol. 16:648–660.
Skinner BF. 1953. Science and Human Behavior.
Spiegel MF, Watson CS. 1984. Performance on frequency- discrimination tasks by musicians and nonmusicians. J Acoust Soc Am. 76:1690–1695.
Squires KC, Wickens C, Squires NK, Donchin E. 1976. The effect of stimulus sequence on the waveform of the cortical event-related potential. Science. 193:1142–1146.
Steinbeis N, Koelsch S, Sloboda JA. 2006. The role of harmonic expectancy violations in musical emotions: evidence from subjective, physiological, and neural responses. J Cogn Neurosci. 18:1380–1393.
Summerfield C, de Lange FP. 2014. Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci. 15:745–756.
Sutton RS, Barto AG. 1998. Reinforcement Learning: An Introduction. IEEE Trans Neural Netw. 9:1054–1054.
Tervaniemi M, Rytkönen M, Schröger E, Ilmoniemi RJ, Näätänen R. 2001. Superior formation of cortical memory traces for melodic patterns in musicians. Learn Mem. 8:295–300.
Vuust P, Brattico E, Seppänen M, Näätänen R, Tervaniemi M. 2012. The sound of music: differentiating musicians using a fast, musical multi-feature mismatch negativity paradigm. Neuropsychologia. 50:1432–1443.
Vuust P, Ostergaard L, Pallesen KJ, Bailey C, Roepstorff A. 2009. Predictive coding of music--brain responses to rhythmic incongruity. Cortex. 45:80–92.