2020-03

Predictive coding in auditory perception: challenges and unresolved questions

denham, susan

10.1111/ejn.13802
European Journal of Neuroscience
Wiley

PROF. SUSAN L DENHAM (Orcid ID : 0000-0002-0988-5672)

PROF. ISTVÁN WINKLER (Orcid ID : 0000-0002-3344-6151)

Article type : Special Issue Review

Reviewers: Ross Maddox, University of Rochester, USA
Stefan Koelsch, Freie Universität Berlin, Germany

Title: Predictive coding in auditory perception: challenges and unresolved questions

Article type: Review

Section: Perception

Authors: Susan L. Denham[1], István Winkler[2]

Affiliations: [1] School of Psychology, University of Plymouth, Plymouth, UK;
orcid.org/0000-0002-0988-5672

[2] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

orcid.org/0000-0002-3344-6151

Running title: Predictive coding in auditory perception

Total pages: 34, figures: 1

Total words in i) whole manuscript: 9855, ii) abstract: 200

Keywords:

auditory object representation, auditory scene analysis, pattern detection, computational modelling

Corresponding Author mail id:-sdenham@plymouth.ac.uk

Abstract

Predictive coding is arguably the currently dominant theoretical framework for the study of perception. It has been employed to explain important auditory perceptual phenomena, and it has inspired theoretical, experimental, and computational modelling efforts aimed at describing how the auditory system parses the complex sound input into meaningful units (auditory scene analysis). These efforts have uncovered some vital questions, addressing which could help to further specify predictive coding and clarify some of its basic assumptions. The goal of the current review is to motivate these questions, and show how unresolved issues in explaining some auditory phenomena lead to general questions of the theoretical framework. We focus on experimental and computational modelling issues related to sequential grouping in auditory scene analysis (auditory pattern detection and bistable perception), as we believe that this is the research topic where predictive coding has the highest potential for advancing our understanding. In addition to specific questions, our analysis led us to identify three more general questions that require further clarification: 1) What exactly is meant by prediction in predictive coding? 2)What governs which generative models make the predictions? and, 3) What (if it exists) is the correlate of perceptual experience within the predictive coding framework?

Introduction

Perception seems so simple. I look out of the window to see houses, trees, people walking past, the sky above, the grass below. I hear birds in the trees, cars going past, the distant sound of an alarm, a symphony playing on the radio. The world is full of objects that make their presence known to me through my senses – what could be more simple? Yet the apparent simplicity and efficacy of perceptual experience hides a host of difficulties and questions for which we do not yet have the answers. The problem is that the information reaching our senses is generally incomplete, ambiguous, distributed in space and time and not neatly sorted according to its source ; but see (Gibson, 1979). It is clear that to support effective interactions with the world and the objects in it, one of the key functions of our perceptual systems is to discover the likely sources of our sensations (Brunswik, 1956). This points towards the need for creating and maintaining representations that can partition (or segregate) incoming information and integrate source-related information appropriately through time and across modalities, while allowing us to infer details that may not have been detected. Indeed, we typically maintain a sense of a relatively stable environment within which discrete objects have some persistence even when we are not currently sensing them.

This view of perception also resonates with that of Gestalt psychology (Köhler, 1947), i.e. that the whole (Gestalt), formed through integration of component parts, is primary in perception, and influences the processing and perception of the parts. In terms of auditory perception, the Gestalt of the melody or utterance we perceive relies upon integrating discrete sound events often separated by relatively long intervals (in terms of neural processing time constants). In becoming aware of the Gestalt, subjectively, we usually experience a sense of expectation for what might come next. Functionally, the proposal is that our representations generate predictions that guide the integration of component parts. That is, predictions, whether for the continuation of the object in time or its (currently) hidden parts, are integral to object perception. This notion led to the proposal that object perception is a process akin to the generation of scientific hypotheses (Gregory, 1980). The

predictive view of perception has been computationally formalised in the predictive coding framework, e.g., (Mumford, 1992; Dayan *et al.*, 1995; Rao & Ballard, 1999; Friston, 2005; Hohwy, 2007; Bastos *et al.*, 2012)

In this paper, our aim is to review some of the key principles of predictive coding in relation to the evidence base in auditory perception, and to highlight what we believe are important challenges and unresolved questions. Our intention is to complement the more extensive review of the neuroscientific evidence for predictive coding in auditory cortex (Heilbron & Chait, 2017) and neuroscientific critiques of predictive coding; e.g. (Kogo & Trengove, 2015).

Predictive coding has actually been applied to auditory signal processing since the 1960's in the form of linear predictive coding (Atal & Hanaver, 1971) and the method became a mainstay of artificial speech recognition systems for some time (Rabiner & Gold, 1975). However, here we use the term in a somewhat different sense. Rather than predicting the signal on a sample by sample basis as a linear function of previous samples, with offline training devoted to minimising residual error and function coefficients being used to represent components of interest, here, in considering predictive coding as a model of neural processing, we are interested in an online version in which residual errors become important signals of new information (Spratling, 2017b). In this version, which aims at capturing the processing principles of the biological systems underlying perception, representations are continuously formed and maintained as a combination of previous experience and current context. They are considered to act as generative models that signal expectations of future events and are refined through comparisons between actual and predicted input; in essence, these representations instantiate inferences regarding the causes of the sensory input.

The predictive coding framework is attractive as it describes a system, capable in principle of autonomously controlling its own learning in a continuous manner. Predictive coding has been implemented in a number of qualitatively different ways, primarily in relation to visual processing e.g., (Rao & Ballard, 1999; Spratling, 2008; 2017a), but also for other modalities, e.g., (Kilner *et al.*, 2007; Kiebel *et al.*, 2009; Friston & Frith, 2015; Rubin *et al.*, 2016). However, there are key features common to all (Heilbron & Chait, 2017). These are the existence of separate units (neural populations) that signal predictions and prediction errors, embedded within a hierarchical structure. For example, in one of the earliest models of predictive coding, used to account for 'non-classical' responses in primary visual cortex (Rao & Ballard, 1999), signals from higher order visual areas to lower order areas were interpreted as predictions and forward signals from lower to higher areas as prediction errors, calculated as the difference between the top-down predictions and incoming signals from the sensor. It had been shown that cells in primary visual cortex that respond to a short bar of preferred orientation show a reduced response as the bar is lengthened to extend beyond their 'classical' receptive fields, a phenomenon known as end-stopping (Hubel & Wiesel, 1968). In the model, forward responses (prediction errors) are reduced as a result of predictive accuracy increasing with bar length. The reason is that the receptive fields in both levels of the model reflect the statistics of the natural scenes used for training; prediction neurons in the higher level generate more accurate predictions of stimuli that more closely resemble their prior experience, in this case long rather than short line segments. This early model also highlights a number of questions that we will revisit later in the paper: Although there is a clearly documented hierarchical organisation in sensory cortices, what evidence do we have from auditory perception for the operational necessity for a hierarchy of generative processes; i.e., do higher order statistics influence auditory perception?

What exactly is meant by prediction? What governs which generative models make the predictions? What (if it exists) is the correlate of perceptual experience within the predictive coding framework?

The idea that predictability influences auditory perception is often assumed; e.g., many accounts of music perception hold that expectancies and their violation are intrinsic to music appreciation (Meyer, 1956; 1967; Huron, 2006; Rohrmeier & Koelsch, 2012). Indeed, neural activity continues when a musical piece is abruptly silenced (even when the listener is not familiar with it), suggesting that the brain considers the likely continuation of the melody (Kraemer *et al.*, 2005). Similarly, in speech processing predictability arising from semantic and syntactic cues; e.g., (Kutas & Hillyard, 1984; Kamide *et al.*, 2003), has been studied for some time, and has been argued to underlie observed native language benefits (Kutas, 2011). However, clear evidence of a role for predictability in auditory perceptual organisation, termed auditory scene analysis by Bregman (1990), is less easy to find (Bendixen, 2014), although there are some studies we discuss below that address this issue. It should also be noted that demonstrating predictability influences or facilitates perception does not in itself provide sufficient evidence for predictive processing, as it is possible that predictability is inferred *post hoc*, rather than involving the generation of signals in anticipation of a predicted event; few studies actually allow this differentiation to be probed.

Nevertheless, there are several paradigms that have been used to investigate processing in the auditory system for which the measured responses have been interpreted in terms of predictions and prediction errors. For example, in the phenomenon known as stimulus specific adaptation (SSA) (Ulanovsky *et al.*, 2003), neural responses are found to increase as an inverse function of stimulus probability, i.e. the less predictable (more surprising) a stimulus is, the larger the measured response (larger prediction error). A similar interpretation has also been adopted to explain the differential electroencephalogram (EEG) signal, known as the mismatch negativity (MMN) (Näätänen *et al.*, 2011; Winkler & Czigler, 2012). In this case, large scale brain responses (rather than individual neuron activity) tend to increase in response to an unexpected (less predictable) stimulus. Even though extensive, the MMN literature cannot resolve the issue of prediction versus post hoc predictability. Although much of this literature is compatible with a hierarchical predictive coding framework, the effects of explicit knowledge are equivocal. For example, Sussman *et al.* (2002) found that explicit knowledge about upcoming deviants can eliminate MMN (i.e., knowing about the higher-order structure of a sequence allows one to predict violations of local regularities), whereas others, such Rinne *et al.* (2001) and Horvath *et al.* (2011) found that prior knowledge had no effect on the MMN. Similarly, music-related expectancy violations (regular versus irregular chord progressions) elicit an early right anterior negativity, and continue to do so even when individuals are informed that a specific violation is about to occur (Guo & Koelsch, 2016).

Somewhat stronger support for predictions *per se* comes from the phenomenon known as the omission response; when a sound is unexpectedly omitted from a predictable sequence, there is a response to the missing sound (Joutsiniemi & Hari, 1989). This response, it has been argued, provides clear evidence for a role of predictability in auditory processing as it is elicited in the absence of a stimulus and thus indicates that the auditory system must be generating some form of prediction in order to elicit the signal of prediction error, i.e. the difference between the expected event and the actual null input. Indeed, perhaps the strongest evidence for prediction comes from the study of Bendixen *et al.* (2009), who showed that the EEG signal in response to the first 50ms of a predictable omitted tone is indistinguishable from the response to the tone itself. A similar claim,

regarding prediction, is made for the closely related offset response (Hillyard & Picton, 1978). However, there are some caveats; omission responses to a pure tone sequence are only elicited when the inter-onset interval is less than ca. 170 milliseconds (Yabe *et al.*, 1998), and the offset response is strongly affected by the isochrony of the sequence (Andreou *et al.*, 2011). Neither of these limitations is easily explained by predictive coding, but we do not explore them further here.

Are predictions necessary to explain auditory perceptual experience?

Theoretical perspectives

Are auditory object representations necessarily predictive? While the concept of an object, and its key role as the unit of attention (Duncan, 1984) and prediction (Zhao *et al.*, 2013), is firmly established in the visual domain, the same is not true for auditory perception. The term object is sometimes used to refer to single sound events, such as individual tones or even sub-components of speech like phonemes, e.g. (Kral, 2013), and sometimes to sequences of sound events, such as a melody (Wightman & Jenison, 1995). Another distinction can be made between the sound source (concrete object) and the pattern of sounds emitted by it (abstract object), as representations of either of these possess the main attributes expected for object representations (e.g. a violin and the tune played by the violin both have object properties in that they can enter into various cognitive processes and are invariant to many transformations). (Kubovy & van Valkenburg, 2001). Recent definitions of the term also adopt different views on the necessity for prediction. Griffiths and Warren (2004), describe auditory object processing as the separation of information relating to a thing (sound source) in the world from the rest of the world, and the abstraction of that information across different occurrences of the same object and across modalities. In contrast, Winkler *et al.* (2009), more inclined towards the Gregory (1980) notion of perception, conceive of an auditory object as a persistent representation of a putative thing in the world that is derived from patterns (or regularities) in the sensory input and generates predictions of parts of the object not yet available. In this definition, object representations are mental constructions of inferred sound sources that provide the means for predicting parts of the 'whole', while Griffiths and Warren's discussion of objects focusses more on discrete events and their boundaries and makes no mention of inference or prediction; see also (Kubovy & van Valkenburg, 2001). The terms pattern and regularity are often used interchangeably, but here we use *pattern* to refer to a short recurrent temporal sequence and *regularity* to refer to more general statistical predictability; e.g., the overall timbre or smooth changes in the pitch of a talker's voice.

This lack of consensus regarding auditory objects is reflected in current models of auditory scene analysis. While grouping of simultaneously present components, based on harmonicity, common onsets (and offsets), and directionality has been extensively modelled (e.g., see (Wang & Brown, 2006) for a range of work in this area), grouping of object components separated in time has suffered from a lack of clarity regarding whether the result of such grouping should be considered a stream (Bregman, 1990), a figure versus ground (Teki *et al.*, 2011), or an object (Winkler *et al.*, 2009; Winkler & Schröger, 2015).

Auditory streaming has long been a favoured experimental paradigm for exploring sequential grouping in auditory scene analysis (Bregman, 1990) because it allows one to investigate the cues as well as the temporal dynamics of grouping. The two-tone auditory streaming stimulus most commonly used consists of a sequence of pure tones with alternating high (H) and low (L) frequency, arranged either as a simple alternating sequence, HLHLHL…, or as series of triplets separated by a silent interval (_), HLH_HLH_HLH_ …(van Noorden, 1975). Due to the ambiguous nature of this stimulus, listeners generally experience perceptual bi- or multistabilty (Schwartz *et al.*, 2012). That is, listening to long segments of the stimulus leads to perception switching back and forth between alternative sound organizations (Anstis & Saida, 1985; Denham & Winkler, 2006; Pressnitzer & Hupe, 2006). The possibility for alternative interpretations provides a good testbed for exploring influences of predictability on perceptual organisation.

Disagreement over the need for predictions to account for auditory streaming has a long history. Bregman (1990) did not include it in his theoretical framework; see also (French-St George & Bregman, 1989; Rogers & Bregman, 1993). In contrast, in her dynamic attending theory (Jones, 1976; Jones & Boltz, 1989), Jones suggests that temporal predictions determine sound grouping through dynamic modulation of attention (Demany & Semal, 2002; Devergie *et al.*, 2010). Some more recent accounts, e.g., (Winkler & Schröger, 2015), argue that the 'old+new' strategy (i.e., that the auditory system first assigns parts of the input to previously discovered streams and treats the residue as a new one), described by Bregman as an essential principle of auditory scene analysis, relies on predictive processing in order to achieve sound segregation and the formation/maintenance of auditory object representations in a single pass. Although situated within a different literature, the 'old+new' strategy is essentially a restatement of the 'explaining away' principle that is at the core of many predictive coding systems.

Modelling perspectives

The implementation of theoretical models in computational form raises questions of its own regarding the complexity of the representations needed to explain the target phenomena and the necessity for explicit predictions. By 'explicit' predictions here we mean that there is neural activity that can be interpreted as conveying predictive information in anticipation of unobserved events or parts of an object. In contrast, the current intrinsic state of the system can be interpreted as embodying 'implicit' predictions without any overt activity.

Two computational models of auditory streaming can be used to illustrate contrasting views on the need for explicit predictions and the nature of the representations necessary for explaining the perceptual phenomenon. In the auditory streaming model of Barniv and Nelken (2015), incoming sound events are assigned to one of two classes and perceptual decisions are expressed in term of streams (Bregman, 1990). The perceptual decision is determined by which sounds are assigned to the currently dominant class: if both high and low tones are assigned to the dominant class then the 'perceptual' decision is one stream, and if they are assigned to different classes the decision is two streams. The model works through evidence accumulation in favour of the non-dominant class, and perceptual switching is determined by the dynamics of the class centroids and the strength of the

evidence for each class. Predictions in this model are implicit, manifesting in the accumulated probability distributions, which represent the featural expectations for future events explained by each class. However, the model does not make explicit temporal predictions nor does it represent information about event timing or ordering.

In contrast, (relative) timing, ordering, and featural information are all retained in the model of Mill *et al.* (2013). In this model, representations of repeating (periodic) patterns discovered in the incoming sound sequence are extracted on a continuous basis. The set of representations is dynamic and in principle unlimited; as a result, many representations, each containing information about a repeating embedded pattern, may be held in parallel. Each representation (termed a proto-object) makes an explicit prediction of the next event it expects together with its timing. The dynamics and contents of perceptual experience are represented by the changes in dominance (through bifurcation) of one or more proto-objects. Prediction is fundamental to the model and plays three key roles: 1) competition between proto-objects that predict the same event leads to the emergence of the perceptual organisations reported by listeners, 2) predictions guide decisions about event ownership, determining the likelihood that an event was generated by a known source (object), and 3) predictions are used to verify representations, which are deleted if their predictions fail.

Whereas both models attempt to simulate perceptual decisions in auditory streaming, only the proto-object model (Mill *et al.*, 2013) integrates auditory stream segregation with auditory object formation. The models also differ in the nature of the sequence memory they maintain. In the evidence accumulation model (Barniv & Nelken, 2015), the focus is on feature distributions, and temporal patterns affect the class representations only indirectly. The model of Mill *et al.* (2013) assumes a very detailed memory of the features and timing of sound events and explicit representation of patterns detected in the incoming sequence. However, both models build and maintain alternative (non-dominant) representations in parallel, consistent with electrophysiological results that suggest that representations of alternative sound organizations are maintained in the brain (Sussman *et al.*, 2014).

A similar lack of consensus regarding the need for explicit temporal predictions is found in the literature on mismatch negativity (MMN). For many years there has been disagreement between those favouring a predictive account of MMN, e.g., (Winkler, 2007), and those favouring an explanation based on memory traces (Näätänen, 1990) or adaptation (May & Tiitinen, 2010). This has given rise to numerous experiments seeking to demonstrate the validity of particular theoretical positions, and to the development of competing computational models. However, while some models and experiments, e.g. (Garrido *et al.*, 2008; Wacongne *et al.*, 2012), clearly favour an explicit predictive account, it has been surprisingly difficult to dismiss the adaptive memory trace MMN model (May & Tiitinen, 2010) that simulates many key MMN results without explicit predictions.

The modelling literature on stimulus specific adaptation (SSA) is also in disagreement (in SSA, deviant responses are investigated at the microscopic level rather than macroscopic level targeted by MMN, using very similar paradigms). On the one hand, a wide range of SSA phenomena, including sensitivity to context and the distinction between novelty and rarity, can be replicated by the model

of Mill *et al.* (2011) without explicit predictions (memory of the recent past, and hence implicit expectations of future activity, are contained the synaptic state of the network, similar to the adaptive trace MMN model). On the other hand, Rubin *et al.* (2016) argue that responses measured in primary auditory cortex in the SSA paradigm should be understood as prediction errors and proposed a model that generates predictions and prediction errors as a probabilistic function of recent context. They derived a large number of candidate models and showed that the representations that best matched the neural data tended to have rather long memory spans but coarse featural resolution; challenging assumptions of both the Mill *et al.* (2013) streaming model and predictive coding explanations regarding the correlates of perception that we will come to later.

Experimental perspectives

Pattern detection

A first step towards establishing a role for prediction in auditory perception is to show that listeners are sensitive to acoustic regularities or patterns in sound sequences as such knowledge would in principle provide a basis for predicting upcoming sounds. As previously noted, demonstrating that predictability influences perception is not sufficient to show that predictions are generated, but it does provide some support in this direction. The literature on MMN provides a great deal of evidence for a role for predictability; e.g., (Winkler & Schröger, 1995). Another source of evidence comes from the experiments of Chait and colleagues (Chait *et al.*, 2012; Jaunmahomed & Chait, 2012; Barascud *et al.*, 2016; Southwell *et al.*, 2017) in which rapid sequences of tones characterised by different pitches are arranged randomly or as repeating patterns. Listeners easily detect transitions between random and patterned sections. Pattern detection is also very rapid; listeners can detect repeating tone patterns as rapidly as an ideal observer, taking only roughly 1.5 repeats of the pattern to do so (Barascud *et al.*, 2016). The detection of pattern termination is also very rapid, and is marked by an offset response. However, while the offset response can be interpreted as a sign of prediction error in response to failed predictions of the repeating pattern, this data pose an explanatory challenge. Contrary to expectations based on the classical predictive coding framework, it was found that magnetoencephalogram (MEG) activation increased as a function of the predictability of the acoustic signal; see figure 3 (Barascud *et al.*, 2016). For similar findings, see (Sohoglu & Chait, 2016; Southwell *et al.*, 2017). This is also the reverse of what was expected based on evidence from MMN and SSA neurophysiology and models. The problems posed for the predictive coding framework are immediately apparent; in the MEG signal, the offset response, interpreted as indicating prediction error (increasing surprise) is superimposed on a sustained response that increases with predictability (decreasing surprise). Of course, not all signals measurable from the brain necessarily represent prediction errors, and the authors suggested that the increase in predictability-related sustained activity might indicate precision-weighting of prediction errors (i.e. predictions made with higher confidence give rise to larger error signals than those made with lower confidence). It may also be a more general example of the so-called repetition positivity (Haenschel *et al.*, 2005; Baldeweg, 2006), which increases with increasing repetitions of the same tone. However, interpretation in terms of (precision-weighted) prediction error is difficult to reconcile with near optimal pattern detection performance; why is there any prediction error at all in response to precisely predictable sequences?

Another issue that comes from this study relates to the saliency of the sequences. Increased magnitude of brain responses is generally interpreted in terms of higher perceptual saliency and, consistent with this idea, it has been shown that task irrelevant regularities attract attention, both in vision (Zhao *et al.*, 2013) and in audition (Levänen & Sams, 1997). Therefore, a natural hypothesis is that more predictable tone sequences should be more salient, and hence more distracting, if task irrelevant. However, this is not the case (Southwell *et al.*, 2017); random sequences are far more potent distractors than regular ones (Jones *et al.*, 1999; Macken *et al.*, 1999), and a stream can be easier to segregate when the other stream is regular than when it is random (Bendixen *et al.*, 2010; Andreou *et al.*, 2011; Rimmele *et al.*, 2012).

Do patterns play a role in auditory streaming?

The role of patterns in auditory perception has also been explored in the context of auditory streaming (Bendixen *et al.*, 2010). The question in this case was not whether patterns could be detected, but whether their presence, and the associated increase in predictability of the stimulus, influenced perceptual organisation. This was addressed by jittering the pitch and loudness of the high and low tones using the triplet version of an auditory streaming sequence and the phenomenon of multistability to assess the influence of patterns on perception. In various conditions, the predictability of the pitch and loudness features was manipulated (see (Bendixen *et al.*, 2010) Figure 1) and their influence on the probability and mean phase durations of reporting different perceptual organisations was measured according to four categories: integration, segregation, the more complex 'both' response, or none of these (Denham *et al.*, 2014). It was found that the presence of repeating patterns (precisely predictable pitch and loudness features), which were detectable only while the segregated state occurred, served to increase the probability of segregation and reduce that of integration, relative to the control condition in which the same features that made up the patterns were randomly ordered (see (Bendixen *et al.*, 2010) Figure 2). This change was brought about through a significant increase in segregated phase durations with no significant effect on integrated phase durations, leading to the suggestion that patterns stabilise but do not trigger segregation; for similar results, see also (Bendixen *et al.*, 2013).

In a follow up experiment, evidence for the influence of predictability in the opposite direction (i.e. favouring the integrated perceptual state) was also found (Bendixen *et al.*, 2014). In this case, featural regularities (pitch, spatial direction, relative onset time) which favoured integration (tones in a triplet shared the same perturbation, while perturbations from one triplet to the next were unpredictable) resulted in a larger proportion of integrated responses, while regularities which favoured segregation (perturbations changed smoothly, but independently, in the high and low streams) resulted in a larger proportion of segregated responses (see (Bendixen *et al.*, 2014) Figures 1 and 2). Moreover, the effect of predictability was rather large in comparison with the effect of differences in the acoustic features generally manipulated in streaming experiments (i.e. difference in pitch between the high and low tones, and presentation rate). However, in this experiment the influence of predictability was found to be symmetrical (i.e. increased integrated phase durations were associated with decreased segregated phase durations, and *vice versa*), contradicting the conclusion from the 2010 experiment; for similar results, see (Szalárdy *et al.*, 2014).

Given the demonstration that the presence of predictable patterns influences perceptual organisation, a follow-up series of experiments (unpublished) was designed to investigate the speed with which patterns might be discovered and exert their influence on perceptual organisation. In these experiments, which used the same pitch and loudness manipulations reported in (Bendixen *et al.*, 2010), a wider range of patterns was used and in some conditions, 30 second patterned and random segments were interleaved; see Figure 1. However, in contrast to the previous studies, described above, no consistent effect of predictability on auditory perception was found; i.e., the presence of within stream patterns did not always result in an increase of segregation. Rather, slow fluctuations in the probability of segregation, which are not aligned with the temporal schedule of the predictability manipulation, can be observed.

In sum, evidence from these experiments regarding the influence of predictability on auditory perceptual organisation is not unequivocal, raising the possibility that the larger context within which the sequences are encountered may modulate the effects of predictability on auditory stream segregation.

In contrast, evidence obtained using a different stimulus paradigm is compatible with the notion that predictable sequences do help listeners to form more effective representations of the auditory environment. Sohoglu and Chait (2016) presented listeners with complex soundscapes, comprising multiple parallel sequences of identical tone pips, differing from each other in pitch, tone duration and inter-tone interval. The temporal schedule of all tone sequences was either isochronous (regular context) or jittered (random context). Listeners were faster and more accurate in detecting the emergence of an additional tone sequence within the regular than in the random context, suggesting that the regular context allowed them to quickly identify tones that did not conform to any of the previously encountered sequences. These results were then extended in a separate series of experiments (Aman *et al.*, in press) to different numbers of parallel sequences, appearance and termination of a target sequence, and non-isochronous regularities. Further, the results suggest that listeners don't need to be aware of the presence of the regularities for utilizing their advantages in detecting changes in the sequences. Lacking the relevant measures, this study does not provide information regarding whether the predictability of the sequences that make up the context aids in their segregation. However, it does provide strong evidence that regularities are utilized in the context of auditory scene analysis.

Overall, the picture emerging about the role of predictable patterns in sound processing is that 1) the human auditory system is sensitive to patterns and statistical regularities in sequences of sounds, 2) detecting regularities does not require attention to be focused on the sounds (Sussman, 2007) and listeners are not necessarily aware of the detected regularities, e.g., (van Zuijen et al., 2006; Paavilainen et al., 2007; Aman et al., in press), and 3) the utilization of regularities varies with context and experimental details in ways that have yet to be fully understood. Taken together, what these studies show is that the influence of predictability on perceptual organisation is not mandatory. While an influence of higher order statistics might be expected from investigations of hierarchical structures in language (Federmeier, 2007; Fitch & Martins, 2014) and music (Rohrmeier & Koelsch, 2012), it is not clear to what extent novel hierarchical structures (here, the presence of recurrent embedded patterns) influence auditory perceptual organisation in general. Perhaps an

answer to the puzzle raised by these studies may lie in the contradictory findings that predictable sequences attract attention while also being easier to suppress; the presence of both tendencies may allow the influence of predictability to be easily modulated according to intrinsic preferences, attentional set and task demands.

Perceptual change without prediction error?

An explanation for perceptual bistability has been proposed within the predictive coding framework, e.g., the model of binocular rivalry by Dayan *et al.* (1995), and the theoretical analysis by Hohwy *et al.* (2008). In these accounts, it is suggested that there is no direct competition between alternative interpretations of the stimulus or between low-level features, rather bistability results from the presence of a residual error that is not explained away by the currently dominant percept. This evidence for the presence of "something else" in the scene causes instability that leads to perceptual switching, and if the scene does not change then the instability and perceptual switching persist. However, while this explanation works for the case of binocular rivalry, it cannot account for perceptual bistability in the auditory streaming paradigm. The problem is that the percept known as integration accounts for everything in the scene. Therefore, there is no residual error to drive switching. So, the perceptual flexibility, that is evidenced by the demonstrable exploration of viable alternative interpretations in the face of an unchanging scene, requires a different explanation at least in the case of auditory streaming.

In conclusion, neither the theoretical nor the computational modelling nor the experimental literature is conclusive with regard to the need for explicit predictions or persistent detailed memory representations to explain the decomposition of the acoustic scene into auditory object representations.

Some general challenges to the predictive coding framework

What is meant by prediction?

A precise definition of what is meant by prediction is needed. In the literature, the term has different meanings which are often conflated; e.g. correlation (a relationship between two variables such as one part of an object being said to 'predict' another simultaneously present part) versus inference (a perceptual decision regarding likely explanation or cause). Although the net result may be similar, computationally the two cases are different: the former involves testing relationships between subsets of the available information, whereas the latter involves extrapolating from the available information to something for which there is not (currently) full support. From the modelling literature, as discussed above, it is often unclear whether there is necessarily an explicit prediction signal, or whether the expectations implicit in the state of the system should also be considered to constitute predictions.

Another related question touches on what is predicted, specifically whether predictions convey information about expected timing. In auditory perception, this is particularly important as events occur (sometimes rather precisely) in time, e.g. consider the rhythmic patterns that characterise music. Indeed, this is the focal point of Jones' (1976) rhythmic attention theory. One answer to this question may come from the work of Costa-Faidella *et al.* (2011) who showed that the auditory N1 and P2 event-related responses were differently affected by the temporal regularity of repeating tones; N1 only exhibited repetition suppression when the tones were isochronous, while P2 showed repetition suppression for regular and randomly timed sequences, suggesting that information about the content (what) and timing (when) of predictions may be separately represented in audition. But if this is the case, what are the prerequisites for the two representations to be utilized in a conjoined or separate manner? Answering this question may lead to understanding the equivocal pattern of results obtained regarding the role of patterns in auditory stream segregation: perhaps when predictions refer to both content and timing, a pattern effect emerges, but when only the content is predicted, the presence of patterns is not effective.

What is the correlate of perception?

Is perceptual experience somehow related to the multiscale array of predictions produced by the hierarchy of generative models, or should it be understood more in terms of feedforward prediction errors that might also convey some sense of the perceptual saliency of the stimulus? If perception equates to the latter, then there is a problem in the limit as perfectly predicted events should become inaudible (or invisible). This is clearly not the case; while listening to a repeating two tone sequence for 10 minutes may be very boring, it remains clearly audible. It is also difficult to reconcile perception with the continued existence of prediction errors in response to very simple, precisely predictable sequences with near optimal pattern detection performance (Barascud *et al.*, 2016).

There are problems too when we try to equate predictions with perception, as is assumed in the free energy formulation of predictive coding; e.g. see (Friston *et al.*, 2012). Subjectively, we are aware of very fine nuances in the sounds we hear; small variations in pitch can change the sense of an utterance (e.g. questions versus statements; meaning in tonal languages), and the fine structure of sounds (even meaningless noises) is recognisable and memorable; e.g., see (Kaernbach, 2004; Agus *et al.*, 2010). This suggests that predictions need at some point to be very detailed indeed if they are to account for perceptual experience. However, the analysis of neural responses in primary auditory cortex by Rubin *et al.* (2016) showed that the generative models that best explained neural activity had rather coarse representations. Further, if the predictions of the generative models represent perceptual experience, then one might also expect the same perceptual fidelity from mental imagery; while this may be the case for some people, for the majority mental imagery is clearly less detailed than veridical experience.

The proposed hierarchical generative structure doesn't offer any solutions to this problem; e.g. simply moving the correlate of perception up the hierarchy raises questions about readout; the 'Cartesian theatre' problem (Dennett, 1991). Proposals from outside of the predictive coding literature regarding the neural correlates of awareness include "ignition" (Dehaene & Changeux, 2011), and "information integration" (Tononi & Koch, 2015). However, neither of these maps easily onto the predictive coding hierarchical generative architecture, and large scale brain connectivity

has been shown to exhibit small world properties (Sporns & Honey, 2006) rather than a strictly hierarchical organisation.

Related to this issue is also the question of what determines the contents of perception. In the predictive coding framework, it is proposed that the hypothesis with highest posterior probability determines perceptual content (Hohwy *et al.*, 2008), where posterior probability is calculated as likelihood (how well the hypothesis predicts the input) multiplied by the prior probability of the hypothesis. However, this is not really computable within the predictive coding framework as currently formulated, e.g. see (Friston *et al.*, 2012); evaluating the likelihood of a hypothesis at one level of the perceptual hierarchy changes activity at the lower level, thus only one hypothesis can be evaluated. How then is the hypothesis with highest posterior probability determined? This becomes even more complicated if one leaves the laboratory and enters a natural environment in which multiple objects are simultaneously present; what determines which hypotheses can co-exist and which compete for dominance? In the model of Mill *et al.* (2013) it was proposed that competition and co-existence should be mediated through prediction (representations that predict the same event compete, while representations that predict different events do not). In the predictive coding framework, even if the posterior probability of multiple models were to be computable, it is not clear how compatibility (versus competition) between hypotheses might be determined.

Perhaps in the end predictive coding (at least in its current state of development) does not model (subjective) perceptual experience but rather provides a functionalist model of perceptual decision making. In other words, predictive coding may provide a principled account of the mechanisms needed to react to the world, and in its more recent (sensorimotor, social communication) advances, even to interact with the world. In this sense, being enslaved to the need to make faithful predictions may not matter so much. As Brunswik (1956) suggests, an organism happily sacrifices perceptual precision in perceptual decisions, as it is far more important to quickly detect the presence of a dangerous predator than to form a detailed representation of its exact shape and colouring. Studies showing that human perceptual decisions can be modelled on the basis of Bayesian inference rules, e.g. (Barascud *et al.*, 2016) and that similar rules apply to action decisions (Friston *et al.*, 2011), provide some of the best examples of the utility of the predictive coding theory.

However, even if we restrict our considerations to perceptual decision making, further difficult questions of relevance to auditory scene analysis remain, including the following: How are the generative models created? What determines which generative models are active at any given point in time? How is dominance determined? How is 'effective' competition between alternative models mediated? Where (at what level of the hierarchy) does this 'competition' occur? If competition occurs throughout the hierarchy, how is consistency between levels ensured?

Summary

Predictive coding can be viewed as a milestone in the study of perception, because it is the first general theory that aims to explain psychological and neural aspects of perception by the same principles (thus answering both to behavioural and neurophysiological evidence), expressing them using mathematical formulae and computational models, thereby allowing direct falsification. As a

result, predictive coding is arguably the currently dominant theory of perception. In writing this review our purpose was not to dismiss predictive coding, rather we have tried to explore specific questions raised by considering some phenomena of auditory scene analysis from a predictive coding point of view. We think that at this point, predictive coding should be regarded as a conceptually attractive framework which is far from being sufficiently specified for formal proof or falsification. Our aim has been to point out issues that require consideration and which, if addressed, would allow for more stringent tests of and refinements to the theory.

It is clear from this review that there is a lack of consensus in the field regarding the role of predictions in auditory perception; a) from a basic theoretical perspective there is no agreement on the need for object representations to be predictive; b) models of sequence processing (sequential auditory stream segregation, MMN, and SSA) differ in their implementation specifically with regard to the need for explicit versus implicit predictions; and c) behavioural experiments demonstrate sensitivity to patterns, but evidence regarding their role in decomposing the auditory scene is inconclusive. While there are many studies and models, which have been interpreted in terms of predictability, there are few studies that provide direct evidence for prediction, and what we have tried to show is that the extant evidence is somewhat equivocal.

The above mentioned experimental and modelling issues raise three general questions, answering which would provide better theoretical clarity to the predictive coding framework. The first question asks what precisely is meant by prediction. The second concerns the issues of what governs which generative models make the predictions and how their (in)compatibility is determined. The third important question is to more fully consider what within the predictive framework is proposed to correlate with perceptual experience. However, if predictive coding is restricted to explaining perceptual decision making, rather than trying to account for perception *per se*, then we are once again faced with the enormous problem of explaining where subjective perceptual experience comes from and trying to figure out some escape from the resulting dualism.

Acknowledgements
IW was supported by the Hungarian Academy of Sciences (Lendület Project LP-36/2012). SD thanks Prof Georg Klump for the invitation to participate in 'The Active Auditory System' meeting, Bad Honneff, Feb 2017, which provided the impetus for writing this paper.

Abbreviations

EEG: electroencephalogram

MEG: magnetoencephalogram

MMN: mismatch negativity

SSA: stimulus specific adaptation

Competing interests
None

Author contributions

SD and IW jointly contributed to writing the paper.

References

Agus, T.R., Thorpe, S.J. & Pressnitzer, D. (2010) Rapid formation of robust auditory memories: insights from noise. *Neuron*, **66**, 610-618.

Aman, L., Andreou, L.-V. & Chait, M. (in press) Sensitivity to statistical structure facilitates perceptual analysis of complex auditory scenes. *eLife*.

Andreou, L.V., Kashino, M. & Chait, M. (2011) The role of temporal regularity in auditory segregation. *Hearing research*, **280**, 228-235.

Anstis, S. & Saida, S. (1985) Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.*, **11**, 257-271.

Atal, B.S. & Hanaver, S.L. (1971) Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, **50**, 637-655.

Baldeweg, T. (2006) Repetition effects to sounds: evidence for predictive coding in the auditory system. *Trends in cognitive sciences*, **10**, 93-94.

Barascud, N., Pearce, M.T., Griffiths, T.D., Friston, K.J. & Chait, M. (2016) Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, E616-625.

Barniv, D. & Nelken, I. (2015) Auditory Streaming as an Online Classification Process with Evidence Accumulation. *PloS one*, **10**, e0144788.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P. & Friston, K.J. (2012) Canonical microcircuits for predictive coding. *Neuron*, **76**, 695-711.

Bendixen, A. (2014) Predictability effects in auditory scene analysis: a review. *Frontiers in neuroscience*, **8**, 60.

Bendixen, A., Bőhm, T.M., Szalárdy, O., Mill, R., Denham, S.L. & Winkler, I. (2013) Different roles of proximity and predictability in auditory stream segregation. *Learning & Perception*, **5**, 37-54.

Bendixen, A., Denham, S.L., Gyimesi, K. & Winkler, I. (2010) Regular patterns stabilize auditory streams. *The Journal of the Acoustical Society of America*, **128**, 3658-3666.

Bendixen, A., Denham, S.L. & Winkler, I. (2014) Feature predictability flexibly supports auditory stream segregation or integration. *Acta Acustica united with Acustica*, **100**, 888-899.

Bendixen, A., Schroger, E. & Winkler, I. (2009) I heard that coming: event-related potential evidence for stimulus-driven prediction in the auditory system. *Journal of neuroscience*, **29**, 8447-8451.

Bregman, A.S. (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Bradford Books, MIT Press, Cambridge, Mass.

Brunswik, E. (1956) *Perception and the representative design of psychological experiments*. University of California Press, Berkeley, CA.

Chait, M., Ruff, C.C., Griffiths, T.D. & McAlpine, D. (2012) Cortical responses to changes in acoustic regularity are differentially modulated by attentional load. *NeuroImage*, **59**, 1932-1941.

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

Dayan, P., Hinton, G.E., Neal, R.M. & Zemel, R.S. (1995) The Helmholtz machine. *Neur. Comp.* , **7**, 889–904.

Dehaene, S. & Changeux, J.P. (2011) Experimental and theoretical approaches to conscious processing. *Neuron*, **70**, 200-227.

Demany, L. & Semal, C. (2002) Limits of rhythm perception. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, **55**, 643-657.

Denham, S., Bohm, T.M., Bendixen, A., Szalardy, O., Kocsis, Z., Mill, R. & Winkler, I. (2014) Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. *Frontiers in neuroscience*, **8**, 25.

Denham, S.L. & Winkler, I. (2006) The role of predictive models in the formation of auditory streams. *Journal of physiology, Paris*, **100**, 154-170.

Dennett, D.C. (1991) *Consciousness Explained*. Little, Brown & Co. , USA.

Devergie, A., Grimault, N., Tillmann, B. & Berthommier, F. (2010) Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America*, **128**, EL1-7.

Duncan, J. (1984) Selective attention and the organization of visual information. *J Exp Psychol Hum Percept Perform*, **113**, 510-517.

Federmeier, K.D. (2007) Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*, **44**, 491-505.

Fitch, W.T. & Martins, M.D. (2014) Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, **1316**, 87-104.

French-St George, M. & Bregman, A.S. (1989) Role of predictability of sequence in auditory stream segregation. *Perception & psychophysics*, **46**, 384-386.

Friston, K. (2005) A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **360**, 815-836.

Friston, K., Mattout, J. & Kilner, J. (2011) Action understanding and active inference. *Biological cybernetics*, **104**, 137-160.

Friston, K., Thornton, C. & Clark, A. (2012) Free-energy minimization and the dark-room problem. *Frontiers in psychology*, **3**, 130.

Friston, K.J. & Frith, C.D. (2015) Active inference, communication and hermeneutics. *Cortex; a journal devoted to the study of the nervous system and behavior*, **68**, 129-143.

Garrido, M.I., Friston, K.J., Kiebel, S.J., Stephan, K.E., Baldeweg, T. & Kilner, J.M. (2008) The functional anatomy of the MMN: a DCM study of the roving paradigm. *NeuroImage*, **42**, 936-944.

Gibson, J.J. (1979) *The ecological approach to visual perception*. Houghton Mifflin, Boston, MA.

Gregory, R.L. (1980) Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London*, **Series B**, 181–197.

Griffiths, T.D. & Warren, J.D. (2004) What is an auditory object? *Nature reviews. Neuroscience*, **5**, 887-892.

Haenschel, C., Vernon, D.J., Dwivedi, P., Gruzelier, J.H. & Baldeweg, T. (2005) Event-related brain potential correlates of human auditory sensory memory-trace formation. *Journal of neuroscience*, **25**, 10494-10501.

Heilbron, M. & Chait, M. (2017) Great expectations: Is there evidence for predictive coding inb auditory cortex? *Journal of Physiology - Paris*, **in press**.

Hillyard, S.A. & Picton, T.W. (1978) On and off components in the auditory evoked potential. *Perception & psychophysics*, **24**, 391-398.

Hohwy, J. (2007) Functional integration and the mind. *Synthese*, **159**, 315-328.

Hohwy, J., Roepstorff, A. & Friston, K. (2008) Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, **108**, 687-701.

Horvath, J., Sussman, E., Winkler, I. & Schroger, E. (2011) Preventing distraction: assessing stimulus-specific and general effects of the predictive cueing of deviant auditory events. *Biological psychology*, **87**, 35-48.

Hubel, D.H. & Wiesel, T.N. (1968) Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, **195**, 215-243.

Huron, D.B. (2006) *Sweet anticipation: Music and the psychology of expectation*. MIT Press.

Jaunmahomed, Z. & Chait, M. (2012) The timing of change detection and change perception in complex acoustic scenes. *Frontiers in psychology*, **3**, 396.

Jones, D.M., Alford, D., Bridges, A., Tremblay, S. & Macken, W.J. (1999) Organizational factors in selective attention: the interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **25**, 464–473.

Jones, M.R. (1976) Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychological review*, **83**, 323-355.

Jones, M.R. & Boltz, M. (1989) Dynamic attending and responses to time. *Psychological review*, **96**, 459-491.

Joutsiniemi, S.-L. & Hari, R. (1989) Omissions of auditory stimuli may activate frontal cortex. *European Journal of Neuroscience*, **1**, 524–528.

Kaernbach, C. (2004) The memory of noise. *Experimental psychology*, **51**, 240-248.

Kamide, Y., Scheepers, C. & Altmann, G.T. (2003) Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *Journal of psycholinguistic research*, **32**, 37-55.

Kiebel, S.J., Daunizeau, J. & Friston, K.J. (2009) Perception and hierarchical dynamics. *Frontiers in neuroinformatics*, **3**, 20.

Kilner, J.M., Friston, K.J. & Frith, C.D. (2007) Predictive coding: an account of the mirror neuron system. *Cognitive processing*, **8**, 159-166.

Kogo, N. & Trengove, C. (2015) Is predictive coding theory articulated enough to be testable? *Frontiers in computational neuroscience*, **9**, 111.

Köhler, W. (1947) *Gestalt psychology*. Liveright, New York.

Kraemer, D.J., Macrae, C.N., Green, A.E. & Kelley, W.M. (2005) Musical imagery: sound of silence activates auditory cortex. *Nature*, **434**, 158.

Kral, A. (2013) Auditory critical periods: a review from system's perspective. *Neuroscience*, **247**, 117-133.

Kubovy, M. & van Valkenburg, D. (2001) Auditory and visual objects. *Cognition*, **80**, 97-126.

Kutas, M., DeLong, K.A., Smaith, N.J. (2011) A look at what lies ahead:Prediction and predictability in language processing. In Bar, M. (ed) *Predictions in the Brain: Using Our Past to Generate a Future*. Oxford University Press.

Kutas, M. & Hillyard, S.A. (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature*, **307**, 161-163.

Levänen, S. & Sams, M. (1997) Disrupting human auditory change detection: Chopin is superior to white noise. *Psychophysiology*, **34**, 258-265.

Macken, W.J., Mosdell, N. & Jones, D.M. (1999) Explaining the irrelevant sound effect: temporal distinctiveness or changing state? . *Journal of Experimental Psychology: Learning, Memory and Cognition*, **25**, 810– 814.

May, P.J. & Tiitinen, H. (2010) Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology*, **47**, 66-122.

Meyer, L.B. (1956) *Emotion and Meaning in Music*. University of Chicago Press, Chicago.

Meyer, L.B. (1967) *Music, the Arts and Ideas: Patterns and Predictions in Twentieth-century Music*. University of Chicago Press, Chicago.

Mill, R., Coath, M., Wennekers, T. & Denham, S.L. (2011) A neurocomputational model of stimulus-specific adaptation to oddball and Markov sequences. *PLoS computational biology*, **7**, e1002117.

Mill, R.W., Bohm, T.M., Bendixen, A., Winkler, I. & Denham, S.L. (2013) Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS computational biology*, **9**, e1002925.

Mumford, D. (1992) On the computational architecture of the neocortex II. The role of cortico-cortical loops. *Biol. Cybern.*, **66**, 241-251.

Näätänen, R. (1990) The role of attention in auditory information processing as revealed by event related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences*, **13**, 201 288.

Näätänen, R., Kujala, T. & Winkler, I. (2011) Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses. *Psychophysiology*, **48**, 4-22.

Paavilainen, P., Arajärvi, P. & Takegata, R. (2007) Preattentive detection of nonsalient contingencies between auditory features. *Neuroreport*, **18**, 159–163.

Pressnitzer, D. & Hupe, J.M. (2006) Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current biology : CB*, **16**, 1351-1357.

Rabiner, L. & Gold, B. (1975) *Theory and Application of Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.

Rao, R. & Ballard, D. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* , **2**, 79-87.

Rimmele, J., Schroger, E. & Bendixen, A. (2012) Age-related changes in the use of regular patterns for auditory scene analysis. *Hearing research*, **289**, 98-107.

Rinne, T., Antila, S. & Winkler, I. (2001) Mismatch negativity is unaffected by top-down predictive information. *Neuroreport*, **12**, 2209-2213.

Rogers, W.L. & Bregman, A.S. (1993) An experimental evaluation of three theories of auditory stream segregation. *Perception & psychophysics*, **53**, 179-189.

Rohrmeier, M.A. & Koelsch, S. (2012) Predictive information processing in music cognition. A critical review. *International journal of psychophysiology*, **83**, 164-175.

Rubin, J., Ulanovsky, N., Nelken, I. & Tishby, N. (2016) The Representation of Prediction Error in Auditory Cortex. *PLoS computational biology*, **12**, e1005058.

Schwartz, J.L., Grimault, N., Hupe, J.M., Moore, B.C. & Pressnitzer, D. (2012) Multistability in perception: binding sensory modalities, an overview. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 896-905.

Sohoglu, E. & Chait, M. (2016) Detecting and representing predictable structure during auditory scene analysis. *eLife*, **5**.

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K. & Chait, M. (2017) Is predictability salient? A study of attentional capture by auditory patterns. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **372**.

Sporns, O. & Honey, C.J. (2006) Small worlds inside big brains. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 19219-19220.

Spratling, M.W. (2008) Predictive coding as a model of biased competition in visual attention. *Vision research*, **48**, 1391-1408.

Spratling, M.W. (2017a) A Hierarchical Predictive Coding Model of Object Recognition in Natural Images. *Cognitive computation*, **9**, 151-167.

Spratling, M.W. (2017b) A review of predictive coding algorithms. *Brain and cognition*, **112**, 92-97.

Sussman, E., Winkler, I., Huotilainen, M., Ritter, W. & Näätänen, R. (2002) Top-down effects can modify the initially stimulus-driven auditory organization. *Brain research. Cognitive brain research*, **13**, 393-405.

Sussman, E.S. (2007) A new view on the MMN and attention debate: The role of context in processing auditory events. *Journal of Psychophysiology*, **21**, 164-175.

Sussman, E.S., Bregman, A.S. & Lee, W.W. (2014) Effects of task-switching on neural representations of ambiguous sound input. *Neuropsychologia*, **64**, 218-229.

Szalárdy, O., Bendixen, A., Bőhm, T.M., Davies, L.A., Denham, S.L. & Winkler, I. (2014) The effects of rhythm and melody on auditory stream segregation. *Journal of the Acoustical Society of America*, **135**, 1392–1405.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K. & Griffiths, T.D. (2011) Brain bases for auditory stimulus-driven figure-ground segregation. *Journal of neuroscience*, **31**, 164-171.

Tononi, G. & Koch, C. (2015) Consciousness: here, there and everywhere? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **370**.

Ulanovsky, N., Las, L. & Nelken, I. (2003) Processing of low-probability sounds by cortical neurons. *Nature neuroscience*, **6**, 391-398.

van Noorden, L.P.A.S. (1975) Temporal coherence in the perception of tone sequences. Technical University Eindhoven.

van Zuijen, T.L., Simoens, V.L., Paavilainen, P., Näätänen, R. & Tervaniemi, M. (2006) Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: an event-related brain potential study. *Journal of cognitive neuroscience*, **18**, 1292-1303.

Wacongne, C., Changeux, J.P. & Dehaene, S. (2012) A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of neuroscience*, **32**, 3665-3678.

Wang, D. & Brown, G. (eds) (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

Wightman, F.L. & Jenison, R. (1995) Auditory spatial layout. In Epstein, W., Rogers, S.J. (eds) *Perception of space and motion*. Academic Press, San Diego, CA, pp. 365-400.

Winkler, I. (2007) Interpreting the mismatch negativity. *Journal of Psychophysiology*, **21** 147-163.

Winkler, I. & Czigler, I. (2012) Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *International Journal of Psychophysiology*, **83**, 132-143.

Winkler, I., Denham, S.L. & Nelken, I. (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in cognitive sciences*, **13**, 532-540.

Winkler, I. & Schröger, E. (1995) Neural representation for the temporal structure of sound patterns. *Neuroreport*, **6**, 690-694.

Winkler, I. & Schröger, E. (2015) Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and language*, **148**, 1-22.

Yabe, H., Tervaniemi, M., Sinkkonen, J., Huotilainen, M., Ilmoniemi, R.J. & Näätänen, R. (1998) Temporal window of integration of auditory information in the human brain. *Psychophysiology*, **35**, 615-619.

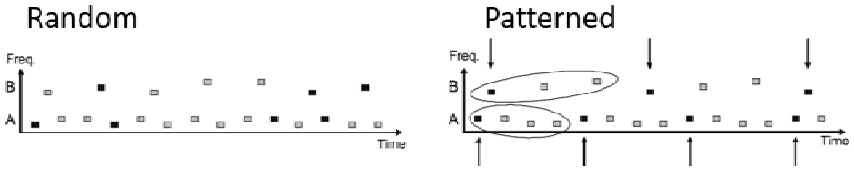
Zhao, J., Al-Aidroos, N. & Turk-Browne, N.B. (2013) Attention is spontaneously biased toward regularities. *Psychological science*, **24**, 667-677.

Figures

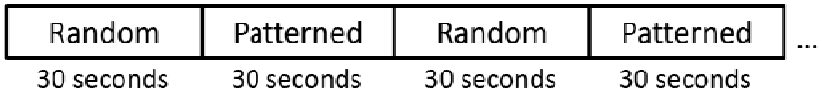
Figure Captions

Figure 1. The influence of higher order structure on auditory perceptual organisation in response to the triplet version of a 'two-tone' auditory streaming sequence. a) Tone sequences were constructed with nominal streams of high and low tones separated by a mean of 7 semitones. The pitches of the tones within each of the streams were drawn from 5 pitch levels within the range +/- 0.7 semitones of the mean pitch. The mean frequency of the low tones was 400 Hz, and of the high tones, 599 Hz (+ 7 semitones), tone duration was 100ms with onset to onset interval, 150ms. In the random segments, pitches were chosen randomly from the 5 levels with equal probability, while in the patterned segments repeating three and four tone patterns were used, in the high and low pitch streams, respectively. Pattern 1 was the pattern used by (Bendixen *et al.*, 2010). Loudness took one of two levels, normal, and loud (+6 dB), with the probability of loud sounds being 33% and 25% in the high and low pitch streams, respectively. b) 30 second random and patterned segments were concatenated to make up a 180 second trial. The same pattern was used throughout a single trial. c) The probability of reporting segregation across 30 participants is plotted as a function of time. The mean probability of segregation for each 30s segment is indicated by a horizontal bar for each random (dotted) and patterned (dash-dotted) segment.

a)



b)



c)