Chapter 1

General Discussion

1.1 Summary of principal results

The work described within this thesis has centered around how prediction is used in sensory processes such as object recognition and prolonged object learning. The work is heavily motivated by the LeabraTI (TI: Temporal Integration) framework (Chapter **??**) which leverages the laminocolumnar structure of the neocortex (Mountcastle, 1997; Buxhoeveden & Casanova, 2002; Horton & Adams, 2005) to learn to predict temporally structured sensory inputs. Predictive learning in the LeabraTI framework is made possible by temporally interleaving predictions and sensory processing across the same populations of neurons so that powerful error-driven learning mechanisms (O'Reilly & Munakata, 2000; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012) can be used to compute a prediction error that can be learned against to minimize the difference between predictions and sensory events over time.

LeabraTI relies on a 10 Hz prediction-sensation period as its core "clock cycle", suggested to correspond to the widely studied alpha rhythm observable across posterior cortex using scalp EEG (Palva & Palva, 2007; Hanslmayr, Gross, Klimesch, & Shapiro, 2011; VanRullen, Busch, Drewes, & Dubois, 2011). Chapter **??** investigated the role of the alpha rhythm in prediction by using an entrainment paradigm (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008; Calderone, Lakatos, Butler, & Castellanos, in press) in which stimuli were presented rhythmically at 10 Hz so that predictions and sensory information could be interleaved regularly at the optimal rate proposed by LeabraTI. The experiment made use of three-dimensional objects that required integration over multiple sequential views to extract their three-dimensional structure. Thus, relatively rapid predictive learning mechanisms that operate over subsequent 100 ms periods could be leveraged to optimally encode the the objects. The spatial coherence between views and temporal onset of each view were independently manipulated to determine their effect on stimulus encoding quality and the putative role of the alpha rhythm in predictive processing.

The results of the Chapter **??** experiment indicated that spatial coherence and predictable temporal onset of each stimulus in an entraining sequence enhanced discriminability of a subsequently presented probe stimulus. Oscillatory analyses indicated strong bilateral alpha power and phase coherence modulation as a function of stimulus predictability. Specifically, spatial predictability of entraining stimuli suppressed alpha power with a lower degree of phase alignment relative to unpredictable stimuli. Temporally predictable entraining stimuli had the opposite effect, with increased alpha power and phase alignment, indicating successful entrainment. Importantly, phase alignment due to temporal predictability remained elevated compared to temporally unpredictable stimuli during a 200 ms blank period between the entraining sequence and probe, indicating that the effects of temporal predictability could persist without exogenous entrainment. In addition to these bilateral main effects, right hemisphere sites exhibited synergistic effects of combined spatial and temporal probe predictability on EEG amplitude and 10 Hz phase coherence approximately 200 ms after probe onset.

Overall, the results of the Chapter **??** experiment support the basic claims put forward by the LeabraTI framework. The predictable 10 Hz presentation rate of the entraining sequence improved encoding of the target object, enhancing discriminability for the subsequent probe stimulus. This finding was accompanied by increased alpha phase alignment that remained elevated until the onset of the probe, which was necessary for ensuring that the probe event was processed precisely when the brain was expecting sensory information and not when it was generating a prediction.

Given this basic support for the LeabraTI framework, the Chapter **??** experiment was designed to investigate the role of prolonged predictive learning of dynamic stimuli. Previous work has suggested that a temporal association rule similar to the one central to LeabraTI might be leveraged for constructing stable representations of spatially coherent visual inputs (Stringer, Perry, Rolls, & Proske, 2006; Wallis & Baddeley, 1997; Isik, Leibo, & Poggio, 2012) and indeed, a line of experiments by Di Carlo and colleagues demonstrated that predictable object transformation sequences build transformation invariance crucial for robust object recognition (Cox, Meier, Oertelt, & DiCarlo, 2005; Li & DiCarlo, 2008, 2010; Li & Dicarlo, 2012). Given these results, one might hypothesize that combined spatiotemporal predictability would be optimal for prolonged learning of three-dimensional objects.

The Chapter **??** experiment used a subset of the object stimuli from the previous chapter's experiment along with an explicit training period during which observers studied the objects while they were rotated through their views. The study period was followed by a series of test trials that required same-different judgements about static probe stimuli. Somewhat surprisingly, the results of the experiment were an almost complete reversal of the previous chapter's experiment. Discriminability was lowest when stimuli were learned in a combined spatiotemporally predictable context and highest when learned in a completely unpredictable context. Furthermore, there was some indication that the principal differences between predictability conditions during training were driven primarily by degenerate viewing angles caused by three-dimensional foreshortening in the objects used (Balas & Sinha, 2009). In three out of four of the objects used in the experiment, accuracy was lower for degenerate views learned in a spatiotemporally predictable context compared to a completely unpredictable one.

Chapter **??** revisited the LeabraTI framework and described a neural network model that implemented the columnar substructure necessary for predictive learning. The model was trained to recognize the same three-dimensional objects used in the Chapter **??** and **??** experiments with the goal of being able to reproduce the conflicting behavioral results of the experiments. LeabraTI predicts that spatially predictable sequences presented at a regular temporal interval should elicit a synergistic effect on behavioral measures due to the multiple prediction-sensation cycles that successfully integrate visuospatial information at optimal temporal intervals (see also Doherty, Rao, Mesulam, & Nobre, 2005; Rohenkohl, Gould, Pessoa, & Nobre, 2014). Such a synergistic effect was demonstrated in the Chapter **??** EEG results, but was simply additive for behavioral measures. Still, the model provided a reasonable account of these data. Furthermore, the model was able to produce the reversal effect observed in Chapter **??** by increasing the scale of a single projection of synaptic weights as a simple proxy for prolonged learning.

The model further indicated that the synaptic weight scaling that accompanies prolonged learning promoted viewpoint invariance that has been suggested to be formed by spatiotemporal associations (Stringer et al., 2006; Wallis & Baddeley, 1997; Isik et al., 2012; Wallis & Bulthoff, 2001; Wallis, Backus, Langer, Huebner, & Bulthoff, 2009). However, this invariance "trickled down" to lower-level retinotopic feature representations. This was problematic for the objects used, since some of them suffered from extreme foreshortening, causing severely degenerate views. This caused confusion between objects, potentially accounting for the overall reversal observed between the Chapter **??** and **??** experiments.

1.2 Open questions

1.2.1 Does the brain predict at 10 Hz or 5 Hz?

The results of the Chapter **??** experiment indicated that delta-theta band oscillations centered around 5 Hz indexed predictability leading into the probe judgement, in addition to the 10 Hz effects of primary interest. Specifically, 5 Hz power was suppressed due to spatial predictability during the 200 ms blank period between the entraining sequence and probe. 5 Hz power was also suppressed due to temporal predictability of the entraining sequence but then increased at the onset of the probe and remained elevated for over 250 ms after its presentation. 5 Hz phase angle alignment was elevated due to temporal predictability throughout the probe judgement and furthermore and exhibited synergistic enhancement in alignment when spatial prediction was also possible. Altogether, these results suggest delta-theta band oscillations, might also play a role in predictive processing in addition to alpha oscillations as suggested by the LeabraTI framework.

There are two potential explanations for these delta-theta band predictability effects. First, the 5 Hz effects could simply be due to their being s subharmonic frequency of the fundamental effects of predictability observed at 10 Hz. Steady state visual evoked potentials (SSVEP) from exogenous rhythmic stimulation are known to cause power increases at harmonic frequencies, in addition to the fundamental frequency of stimulation. It is unclear, however, whether increases in harmonic power are simply due to sharing the same zero crossing as the fundamental frequency waveform (and thus contributing power) or whether they actually might serve a functionally distinct role from the fundamental frequency (e.g., Herrmann, 2001; Kim, Grabowecky, Paller, & Suzuki, 2011).

The second potential explanation of the observed delta-theta band effects is that sensory predictions occur at both 10 Hz and 5 Hz. Two recent reviews have suggested that sensory prediction is at least partially subserved by delta-theta oscillations (Arnal & Giraud, 2012; Giraud & Poeppel, 2012). These theories were formulated to explain prediction during speech recognition, with a focus on multiple windows of integration that are necessary for robust comprehension of speech's hierarchy of time varying features (e.g., phonemes, syllables, words). Delta-theta band predictions support integration over periods of 150 ms or longer which can be used to predict the overall speech envelope (Aiken & Picton, 2008) or phrasal structure of sentences. These theories are supported by recent empirical evidence of the importance of delta-theta band oscillations in the speech (Arnal, Wyart, & Giraud, 2011) and general audition domains (Stefanics et al., 2010). Others have noted, however, that auditory processing does not exhibit the ubiquitous perceptual discretization that visual processing does (VanRullen, Zoefel, & Ilhan, 2014), which is the fundamental process by which LeabraTI interleaves predictions with sensory events. This does not necessarily mean that audition is not a predictive process. It could suggest, however, that auditory prediction happens at a more abstract level after feature extraction, consistent with the slower overall prediction rate, opposed to the sensory level for visual prediction and as suggested by LeabraTI.

In the Chapter **??** experiment, 5 Hz power and phase coherence were sensitive to the spatial predictability of the entraining sequence, and so it seems unlikely that they were simply a subharmonic side effect of 10 Hz rhythmic stimulation. Thus, the most reasonable explanation of the 5 Hz predictability effects observed in the Chapter **??** experiment results is that they reflected a slower, more abstract visual prediction process such as anticipation of the appearance of the probe and whether it might depict the same object as the entraining sequence or a distractor.

1.2.2 Are “paper clip” objects somehow special?

The “paper clip” objects used throughout the current work have a long history of use in studies of three-dimensional object recognition in human observers (Bulthoff & Edelman, 1992; Edelman & Bulthoff, 1992; Sinha & Poggio, 1996) as well as monkey physiology studies (Logothetis, Pauls, Bulthoff, & Poggio, 1994; Logothetis, Pauls, & Poggio, 1995). The objects are easy to generate systematically and thus can be combined with a staircase procedure to titrate difficulty or can be generated *en masse* to find the parameters that elicit maximal responses during neural recordings. Various effects with the objects have been replicated using computational models of object recognition with identical stimuli (Riesenhuber & Poggio, 1999) and geometric properties of the objects are known to capture a large amount of variability in behavior (Balas & Sinha, 2009). Thus, it can be reasonably concluded that paper clip objects are a useful class of stimuli for studying three-dimensional object recognition.

However, other work brings under question the ecological validity of paper clip objects. The objects are constructed from thin line segments separated by empty space and thus self-occlusion of features is less of a problem than for three-dimensional volumetric objects with surfaces. This might imply that viewpoint invariance is not actually necessary to represent the full three-dimensional structure of paper clip objects, since the majority of features can be extracted from a single estatic view. Accordingly, studies comparing three-dimensional objects composed of line segment with volumetric objects found that the line segment objects were not represented in a viewpoint invariant manner (Farah, Rochlin, & Klein, 1994; Pizlo & Stevenson, 1999).

Thus, it is is possible that a spatiotemporally predictable training context is simply not optimal for learning to represent paper clip objects. Temporal association mechanisms (Stringer et al., 2006; Wallis & Baddeley, 1997; Isik et al., 2012; Wallis & Bulthoff, 2001; Wallis et al., 2009) might bias the development of viewpoint invariance by forming associations between canonical and degenerate views that promote viewpoint robustness by accounting for feature variability, but lower overall accuracy levels and slow reaction times. Another way of stating this idea is in terms of the nature and complexity of the representation at each stage of visual processing. At early stages of vision, objects are represented in terms of spatially localized oriented edges (Hubel & Wiesel, 1962), which are an optimal feature for encoding paper clip objects. At later stages, such as intermediate visual areas and inferior temporal (IT) cortex, neurons encode stimuli in terms of surfaces and complex volumetric features (Cox et al., 2013; Hayworth & Biederman, 2006; Kourtzi & Connor, 2010). Thus, it may be the case that the spatiotemporal associations between entire views learned by IT neurons (Sakai & Miyashita, 1991; Meyer & Olson, 2011; Cox et al., 2005; Li & DiCarlo, 2008, 2010; Li & Dicarlo, 2012) are actually surface-based encodings, which cannot effectively be used to represent paper clip objects since they are composed of line segments separated by empty space. This idea might also account for the conflicting effects in the literature regarding whether sequence predictability is actually advantageous for object recognition, as some investigations used line drawing stimuli (Lawson, Humphreys, & Watson, 1994) whereas others used surfaced volumetric stimuli (Harman & Humphrey, 1999) (see Chapter **??** Discussion).

1.2.3 Is synaptic weight scaling a good proxy for prolonged learning?

The model described in Chapter **??** made use of a relatively simple proxy for prolonged learning in order to account Chapter **??** experiment results. This proxy was necessary because it was somewhat inappropriate to use the LeabraTI algorithm for learning under spatially and temporally unpredictable contexts, especially considering the interpretation of the Chapter **??** experiment results was that such predictive learning mechanisms were not invoked in those learning contexts. The proxy for learning involved training the model with spatiotemporally predictable input sequences using the LeabraTI algorithm and then increasing the scale of the synaptic weights on one of the principal projections before presenting spatially and temporally unpredictable input sequences with synaptic plasticity disabled. Synaptic weight scaling was suggested to be a reasonable first order approximation of prolonged learning since it is one of the many effects of learning, especially when considering the long timescale self-organizing mechanisms presumed by Leabra that reinforce the most active units (O'Reilly & Munakata, 2000; O'Reilly et al., 2012).

An alternative to synaptic weight scaling that would also be more biologically plausible would be to explicitly model other visual areas that do not require LeabraTI's predictive learning. For example, IT cortex has been suggested to have a different alpha rhythm properties than earlier visual areas (i.e., V1, V2, and V4) both in terms of its physiological generators and its behavioral correlates (Bollimunta, Chen, Schroeder, & Ding, 2008). Under this view, IT neurons don't actively predict their own inputs, but do code the spatiotemporal associations (Cox et al., 2005; Li & DiCarlo, 2008, 2010; Li & Dicarlo, 2012) of lower-level neurons that actively perform prediction. Thus, modeling IT cortex would not necessarily require predictive learning and could be used to represent and operate on the three-dimensional object stimuli with more standard learning mechanisms.

To implement such a heterogenous model would require the following more complex approach: First, layers that learn via the LeabraTI algorithm would need to be pre-trained on a set of unrelated input sequences (e.g., natural images) to establish how inputs are capable of predictably transforming from moment-to-moment (100 ms periods in LeabraTI). This pre-training would discover features that capture the principal variance across input transformations but are still relatively task-independent. After pre-training, the model would need to be expanded to include IT cortex and Output layers, which would then be trained on a more specific object recognition task using the standard Leabra algorithm. Such a multistage training regime has been successfully used to train other deep neural network architectures without suffering from error signal dilution (Hinton &

Salakhutdinov, 2006) and has showed promise for combining predictive feature learning with other task-specific architectures (O'Reilly, Wyatte, Rohrlich, & Herd, in preparation).

The synaptic weights produced by the pre-training step along with a relatively short amount of task-specific training on the paper clip objects would probably be sufficient to produce the results of the Chapter **??** experiment in which objects were still psychologically novel and likely to be represented by combinations of task-independent features. In the prolonged task-specific training required to produce the results of the Chapter **??** experiment, the combined spatially and temporally predictable training context would maximally activate early and intermediate visual features due to a close match with the environmental statistics of pre-training, leading to *smaller* error signals that are used to train the later IT and Output stages. The completely unpredictable training context, in contrast, would only weakly activate early and intermediate visual features due to the unexpected spatiotemporal irregularity across presentations, causing larger error signals in later stages. Effectively, IT neurons might learn more robust representations from an unpredictable training context since associations between views of the objects are constantly psychologically novel. This more robust representation would probably benefit recognition of the objects from static views without full sequence information, and might also be considered more viewpoint-centric (compared to the viewpoint invariance learned from spatially and temporally predictable training), although this would need to be quantified via the representation similarity methods used in Chapter **??** (Kriegeskorte, Mur, & Bandettini, 2008a; Kriegeskorte et al., 2008b).

1.3 Conclusion

This thesis has advanced a comprehensive theory of neocortical predictive processing of sensory input sequences. The cumulative work has spanned the detailed analysis of the underlying biological neural circuitry in accommodating predictive learning, experimental support for the theory's core testable predictions, and a neural network model that was constrained by biological details yet capable of producing the experimental results. Some of the idiosyncratic details that differentiate between various models of sensory prediction remain to be resolved and refinements on experimental methods and more complex neural models would also be illuminating to this end. Overall though, the current work represents a significant contribution to our understanding of the core computations that make the brain a “prediction machine.”