We thank the reviewers for their thoughtful comments.

Proofs/technical content in main paper and SI (R1 and R3): For R1, please note that a full set of technical proofs on capacity and decoding is in the SI of our original submission. As the reviewers can see, the proofs take up a lot of space, and so we decided to give a conceptual description in the paper, with specific pointers to the relevant proof in the SI for each of the results cited. In response to R3, we will add a paragraph in the main paper that outlines key equations/proofs, and makes it easier for the reader to locate the relevant proof in SI for a given result in main paper.

Links to RBMs and to graphical models (R1): Our network has RBM architecture, as noted by R1, if each constraint node is viewed as a unit (however, each constraint node is really a set of conventional neural units with recurrent connections within that set). We prove that the same exponential capacity and robust error correction extend to the stochastic update dynamics of Boltzman Machines (SI S11; pointer in line 274 of the original submission).

Our network can be represented as a factor graph (constraint modules are factors), undirected graphical model (clique potentials are indicator functions on visible neurons in a constraint) and Bayes net (in several different ways). None of these are the Bayes nets used by Mackay et al. to decode ECCs, but our network can be made structurally similar to these other Bayes nets by adding an input layer with fixed noisy inputs and slightly rewriting constraint modules (this is the LDPC Bayes net; turbo codes are not fundamentally different). The big difference is dynamical rather than structural. An expander graph code allows simple, neurally plausible decoding to perform at par with BP. The same simple decoding should apply to codes where the equivalent factor graph has an expander structure (e.g., random LDPC). These expander codes can also be decoded by belief propagation (BP), but it's harder the other way around. In general, codes that can be decoded by BP do not admit a simple neural network decoding by the dynamical/energy-based Hopfield rule: the message from node i to j depends on messages from all neighbors of i except j; this leave-one-out structure is hard for biological neurons, though there are some interesting attempts. Thus, we do not expect general turbo code decoding to be performed by Hopfield dynamics. This is a fascinating research area and it will be interesting, in future work, to determine when the mapping is possible and to analyze concatenated codes as neural networks. We will add a short discussion of the relationship to Bayes nets/belief prop and slightly expand the discussion of RBMs.

Relevance to neuroscience (R3): We plan to follow this paper with another paper describing neuroscience applications. For space and coherence, this paper focuses on the conceptual theory without elaborating on applications. However, in response to these comments, we will add a few paragraphs discussing neuroscience applications, detailed below.

The networks we construct (henceforth HPC nets) can perform a number of canonical neural computations, with natural applications to neocortical-hippocampal interactions. The hippocampus plays an important role in neocortical memory, despite its puzzlingly small size ($\sim 10^7$ hippocampal vs. $\sim 10^{10}$ neocortical neurons in humans). Thus it is an appealing place to look for high-capacity networks. In particular, the HPC net can be used to construct a *robust high-capacity pattern labeler*. Here input patterns in a very high-dimensional space (putative neocortex) are mapped to the exponentially-many stable states of a HPC network (putative hippocampus), which serve as memory labels for the patterns. The connectivity matrix can be constructed in a simple, online, Hebbian way (thus in one-shot) as the outer-product of input and memory patterns. When presented with a noisy version of an input pattern, the memory network robustly retrieves the correct label (and maintains it in the absence of input).

Such a pattern labeler can be used for recognition or familiarity detection, template matching, classification, locality-sensitive hashing and nearest neighbor computations. It could also be used for memory consolidation and the learning of conjunctive representations, filling a gap in general theories of the hippocampal formation by allowing a much smaller hippocampal network to provide and robustly retrieve labels for very high-dimensional input patterns. As illustration, consider recognition memory: our network can rapidly store large numbers of inputs (one-shot learning) and then make robust judgments about familiarity or novelty during testing, compatible with the prodigious recognition memory found in human psychophysics. When presented with a previously learned input pattern, the network dynamics settle to a global energy minimum at the state corresponding to the familiar pattern label. By contrast, when presented with a novel pattern, the network settles to one of the many local minima that populate the spaces between the basins of attraction surrounding the global minima. These local minima are higher-energy states and with their higher number of unsatisfied constraints correspond to a higher level of activation in constraint neurons. A single readout neuron which sums constraint activations can thus signal that the pattern was novel.

Sensitivity to form of data (R4): We expect our results will generalize to different neural responses, since the decoding performance is due primarily to the connectivity structure of the network: with expansion, multiple neurons in the constraint layer receive input from only one corrupted input neuron and can identify the error.

Transition between recovery and failure (R4): The steep transition is typical of both error-correcting codes and random phenomena in high dimensions, where many events happen with probability that is asymptotically 0 or 1 (zero-one laws). Even when decoding fails, there is information in the location of convergence (which we use in the recognition memory model described above).