This is a preprint of an article published in Hippocampus 2006 16(3):329-343.
http://www.interscience.Wiley.com/

A functional hypothesis for adult hippocampal neurogenesis: Avoidance of catastrophic interference in the dentate gyrus

Laurenz Wiskott[1], Malte J. Rasch[1*], and Gerd Kempermann[23]

[1]Institute for Theoretical Biology
Humboldt-University Berlin, Germany
http://itb.biologie.hu-berlin.de/

[2]Max Delbrück Center for Molecular Medicine
Berlin-Buch, Germany
[3]Dept. of Experimental Neurology, Charité
University Medicine Berlin, Germany

Abstract

The dentate gyrus is part of the hippocampal memory system and special in that it generates new neurons throughout life. Here we discuss the question of what the functional role of these new neurons might be. Our hypothesis is that they help the dentate gyrus to avoid the problem of catastrophic interference when adapting to new environments. We assume that old neurons are rather stable and preserve an optimal encoding learned for known environments while new neurons are plastic to adapt to those features that are qualitatively new in a new environment. A simple network simulation demonstrates that adding new plastic neurons is indeed a successful strategy for adaptation without catastrophic interference.

Keywords: hippocampus, dentate gyrus, adult neurogenesis, network model, catastrophic interference.

*Now at the Institute for Theoretical Computer Science, Technical University of Graz, Austria.

1 Introduction

The hippocampus is a structure in the mammalian brain that is instrumental for acquiring episodic and declarative memories (Gluck & Myers, 2001). If the hippocampus is lesioned, no new episodes or facts can be stored in long term memory. Skills, however, can still be learned (procedural memory). Because of its importance for learning and memory the hippocampus has been being extensively studied anatomically, physiologically, psychophysically, and by means of computational models.

The principal substructures of the hippocampal formation relevant for our discussion are the entorhinal cortex (EC), the dentate gyrus (DG), the CA3- and CA1-region and the subiculum (see Amaral & Witter, 1989; Hammond, 2001, for an overview). They can be arranged in a loop based on the connectivity like EC $\rightarrow$ DG $\rightarrow$ CA3 $\rightarrow$ CA1 $\rightarrow$ subiculum $\rightarrow$ EC. There are also shortcut connections, e.g. from EC to CA3 or CA1. The entorhinal cortex seems to serve as an interface between the hippocampus and other cortical areas. It sends input to the hippocampus from the superficial layers II/III and receives its output in the deep layers V/VI. The CA3-region has a particularly high degree of recurrent connectivity. Relevant in our context is also that DG has many more neurons than the EC superficial layers (Amaral et al., 1990; Harding et al., 1998; West & Slomianka, 1998) and that it has a relatively low firing rate (Jung & McNaughton, 1993; Gothard et al., 2001) compared to the entorhinal cortex (Barnes et al., 1990; Frank et al., 2000) suggestive for a sparser representation.

One peculiar property of the dentate gyrus is that it generates new neurons throughout life, a phenomenon referred to as adult neurogenesis (Altman & Das, 1965; Gould & Gross, 2002; Kempermann et al., 2004b). The new cells arise from precursor cells and differentiate into granular cells that appear to become fully functional and integrated over a time course of several weeks (van Praag et al., 2002; Jessberger & Kempermann, 2003). However, young granular cells seem to differ from old ones (when they are old, not when they were young) in that they are more plastic, i.e. they show enhanced long-term potentiation (LTP), which has a lower threshold of induction (Schmidt-Hieber et al., 2004) and cannot be inhibited by GABA (Wang et al., 2000; Snyder et al., 2001). It has also been argued that it is reasonable to assume that new cells form synaptic connections more rapidly (Gould & Gross, 2002).

The number of newly added neurons depends on the rate of proliferation, i.e. cell generation, and the probability of cell survival, since most of the newly generated cells actually die again after a short period of time (apoptosis) (Eriksson et al., 1998; Gould et al., 2001; Kempermann et al., 2003). If they survive the initial phase, they can last for a long time, e.g. at least 1 year in mice (Kempermann et al., 2003), 8 months in rats (Altman & Das, 1965), and 2 years in humans (Eriksson et al., 1998). Cell proliferation decreases with age (Altman & Das, 1965; Seki & Arai, 1995; Kempermann et al., 1998b) and can also be reduced by stressful experiences (Gould et al., 1997, 1998; Tanapat et al., 2001). Both might be explained by the dependency of the proliferation on hormones (glucocorticoids); other hormones (ovarian steroids / estrogen) stimulate the proliferation (see Gould & Gross, 2002, for an overview). Other factors that increase neurogenesis are dietary restriction (rat) (Lee et al., 2000), voluntary physical activity (mouse) (van Praag et al., 1999a,b; Kronenberg et al., 2003), enriched environments (mouse, rat) (Kempermann et al., 1998b; Nilsson et al., 1999; van Praag et al., 1999b), and some hippocampal-dependent learning-tasks (Gould et al., 1999), the former factors effecting more the proliferation the latter ones effecting more the cell survival. Genetics also has a strong influence on neurogenesis (Kempermann et al., 1997a; Kempermann & Gage, 2002). The rate of adult neurogenesis showed strong strain differences and so did the survival of the newborn cells. Depending on the strain the rate of survival was between only 25% and 75% (Kempermann et al.,

1997a). The total number of new cells generated per day by proliferation in the dentate gyrus of young adult rats (9 weeks old) has been estimated to be 9000, only about 60% of which survive the first four weeks (Cameron & McKay, 2001). Already the first report on adult neurogenesis by Altman & Das (1965) had suggested a net growth and two studies from the 1980s had elaborated on this issue (Bayer et al., 1982; Boss et al., 1985). The estimated growth was between 10 and 40%. However, these studies have been criticized because of their method of quantification. Preliminary new data from our group that are based on stereological measurements suggest that the increase could be as high as 30% (unpublished observation). However, strain (and species) differences exist with regard to both the rate of neurogenesis and the total granule cell number (Kempermann & Gage, 2002).

In any case, adult neurogenesis occurs at only a very low rate during most of the life, so that any growth would only become noticeable relatively early in life. Adult hippocampal neurogenesis decreases rather sharply in young adulthood and is maintained at a low level throughout most of the remaining life (Kronenberg et al., 2005). In 20 month-old mice we found that only 11 cells were generated over a period of 12 days (Kempermann et al., 2002). With a different method and in rats, Cameron and colleagues estimated a somewhat higher rate that was still very low in relation to the total granule cell count (Cameron & McKay, 1999). We and others have published rates of neurogenesis that seemed much higher (Kempermann et al., 1997a; Cameron & McKay, 2001), but this always applied to rather young adult animals. We have also found in young mice that experience of a complex environment has a measurable effect on the total granule cell number and did not just result in an increased turnover (Kempermann et al., 1997b). The very low rate of neurogenesis after young adulthood makes it unlikely that quantitatively a substantial turnover of dentate gyrus granule cells could occur. The majority of new neurons generated during early adulthood have a survival time of at least a year (Kempermann et al., 2003), but it is not yet clear, whether new granule cells produced in older age have a similar long-term survival. This does not rule out that some neuronal turnover might occur, but both the measurable increase in total cell number at least early in adulthood and the long-term survival of the new granule cells suggest to us that at growth is at least predominant. Nevertheless, some researchers favor the idea that there might not be a net growth at all but rather a turnover. This issue cannot be fully settled at present but we interpret the available data as supporting a net growth.

Like behavioral and environmental factors have an effect on the generation and survival of new neurons, there is also some evidence that neurogenesis affects performance in some hippocampal-dependent learning-tasks, such as trace conditioning (Shors et al., 2001, 2002) or the water maze task (Kempermann & Gage, 2002; Snyder et al., 2005). Reduced neurogenesis might contribute to major depression (Santarelli et al., 2003; Kempermann & Kronenberg, 2003).

What exactly is the functional role of these new neurons that are generated in the dentate gyrus throughout life? What makes the dentate gyrus so special? This question is not only interesting in itself but also offers a new approach for investigating the hippocampus as a whole.

We follow the view that the dentate gyrus performs an encoding operation and assume that this encoding has to adapt to a changing environment the animal lives in. Our hypothesis is that the new neurons are necessary to avoid the negative side effects of this adaptation, referred to as catastrophic interference. In this paper we lay out our arguments in detail and present a simple network model that illustrates our hypothesis and shows that adding new neurons is indeed a reasonable strategy for avoiding catastrophic interference.

In the following section we will introduce relevant concepts from the theory of neural networks. The basic model of hippocampal function we have adopted is described in Section 3. Within this basic model we present our hypothesis in Section 4. The computational model illustrating our hypothesis is defined in Section 5, which also includes the simulation results. We conclude with a discussion.

The basic ideas presented here have earlier been sketched in (Kempermann & Wiskott, 2004). The model simulations have been described in detail in (Rasch, 2003). A mathematical description of the model can also be found in (Wiskott et al., 2004).

2 Artificial neural networks and catastrophic interference

2.1 Feedforward and Hopfield networks

Like physiological neural networks, artificial neural networks consist of a number of relatively simple computational units that are mutually connected (see Hertz et al., 1991, for an introduction). Each unit integrates the activity it receives from other units via the connections, performs some simple stereotype operation on it, and sends its own activity to other units it is connected to. The connections have weights that determine how much one unit contributes to the activity of another unit and thereby what kind of computation the whole network performs. It is characteristic for neural networks that the computation is distributed over many weights, which has advantages and disadvantages. There are a number of learning rules that adapt the weights such that the network performs a certain desired computation, such as pattern recoding or memory storage and retrieval. Feedforward networks (see Bishop, 1995, for an introduction) and Hopfield networks (see Amit, 1989, for an introduction) are two classes of artificial neural networks that are commonly used for modeling hippocampal function.

Feedforward networks have an input layer of units, possibly several intermediate (so called hidden) layers, and an output layer. As the name indicates connections only go from earlier layers to later layers. Thus the network has no particular dynamics but simply computes an output based on an input. It is therefore equivalent to a mathematical function with vectors as an input and vectors as an output. Feedforward networks are usually trained in a supervised fashion, which means that the network gets told explicitly what output it should generate for a given input. Thus the training patterns are actually input-output pairs, for each input there is a desired output and the training procedure is designed such that the difference between desired output and actual output is minimized.

Hopfield networks are very different. Every computational unit serves as an input as well as an output unit and, in its simplest form, can only assume the values 0 or 1. All units are reciprocally connected, so that the network has a highly recurrent connectivity. Since each unit has an influence on every other unit, such a network can have quite a complex dynamics. However, it is typically designed such that after some time the network activity settles in a stable state (a certain combination of zeros and ones) and stays there. Training a Hopfield network means choosing the connection weights such that the stable states the network can settle in correspond to a given set of desired patterns. These are the stored patterns. If one initializes the activity of the network with a corrupted version of a stored pattern, the dynamics of the network will typically arrive at the stable state that corresponds to the uncorrupted pattern. This effect is known as auto-association and is the main advantage of a Hopfield network.

2.2 Catastrophic interference

Due to the distributed nature of the computation performed in a neural network, learning a new pattern typically requires to change all weights. This poses a problem. Assume the weights have been trained for input pattern **a**, so that the network behaves as desired. Then training pattern **b** is given and all weights change to be optimal for input **b**. This will typically degrade the performance on pattern **a**. If a third pattern **c** is learned, the performance on **a** will degrade even further and

very quickly pattern **a** will be forgotten completely. This effect can be quite dramatic and is referred to as *catastrophic interference* (see McClelland et al., 1995; French, 1999, for an overview).

In feedforward networks the problem of catastrophic interference is typically solved by interleaved training, which means that patterns **a**, **b**, and **c** in the example above are presented repeatedly in an interleaved fashion, like **abcabcabcabc** or more irregularly **abbcbaaccbacccaabacbb**, with a relatively small learning rate, which makes the weights change only little per pattern presentation. The patterns are therefore effectively learned simultaneously and no pattern dominates another one. Interleaved learning, however, requires that all training patterns are available all the time. A sequence like **aaaabbbbcccc** would not work.

One way to avoid catastrophic interference in a feedforward network even if the training sequence is of the form **aaaabbbbcccc** is to use a dual network (e.g. Ans & Rousset, 1997; French, 1997). Such an architecture takes advantage of the fact that some networks can spontaneously reproduce patterns that they have learned earlier. In a dual network two such networks are coupled and (partially) train each other, thereby retraining old patterns while learning new ones and thus avoiding catastrophic interference.

Hopfield networks solve the problem of catastrophic interference very differently. They can be easily setup in such a way that orthogonal patterns, i.e. patterns that are uncorrelated, do not interfere with each other at all. Thus if a set of patterns has to be stored in a Hopfield network, it is good to first encode them such that the transformed patterns are orthogonal to each other. They can then be stored without any interference. Of course, one has to also memorize the encoding and in addition the encoding must be invertible, otherwise one could not retrieve patterns from the memory anymore.

One way to make patterns orthogonal is to generate a very sparse representation of them (Bentz et al., 1989). This means that they are encoded such that only few units are significantly active at a time. If only very few units are active, there is a good chance that two different patterns have no active units in common and are therefore orthogonal. This technique might require increasing the number of units for representing the patterns, though. To compensate for this increase one can compress the patterns first, so that irrelevant dimensions are eliminated and only relevant dimensions are represented.

Even with a sparse representation of the patterns, if there are more and more patterns stored in the Hopfield network, at some point the capacity will be exhausted, and new patterns will interfere with old ones. This effect again can be quite dramatic, so that all in a sudden none of the patterns can be retrieved anymore, even those that have been stored recently. Measures can be taken to avoid also this type of catastrophic interference, at least to some extent. For example with the learning of each new pattern all previous weights could be decreased a little bit, so that old patterns are gradually forgotten and make room for new patterns (Howard et al., 1987, cited in Amit, 1989).

When discussing the problem of catastrophic interference, which results from the distributed representation in neural networks, one should not forget the advantages that distributed representations provide. For instance, they are much more robust with respect to damage like the loss of units or connections, a property known as *graceful degradation*. More importantly, feed forward networks can generalize. Only a limited number of training patterns can be learned exactly. As the number of training patterns increases and the network cannot memorize all of them individually, the network has to find an input-output function that is a compromise and provides a reasonably good performance on all training patterns. In finding this compromise the network discovers and employs regularities in the data which results in the nice effect that the network will generate reasonable responses also for new patterns never seen before. This property is referred to as *generalization*.

The problem of catastrophic interference is somewhat related to the *stability-plasticity dilemma* (Carpenter & Grossberg, 1987, cited in Gerstner & Kistler, 2002). The latter refers to the problem of setting the right learning rate. If the rate is too low the network cannot learn new patterns; if the rate is too high the network can learn new patterns well, but old patterns are forgotten quickly, catastrophic interference occurs.

Catastrophic interference is very distinct from the *bias-variance dilemma* (Geman et al., 1992), which refers to the problem of choosing the right network complexity (does not apply to Hopfield networks). If the network is too simple (large bias case) it will not even be able to represent the training patterns well and will produce a large error; if it is too complex (large variance case) it can memorize all training patterns individually and does not need to discover regularities, thus it will not generalize well.

3 A model of hippocampal function

In our considerations we largely adopt a standard model of hippocampal function as proposed and discussed previously (e.g. Treves & Rolls, 1994; McClelland et al., 1995) and summarized in this section. The next section will introduce our hypothesis that adult neurogenesis can prevent catastrophic interference. The model illustrating our hypothesis is described in detail in Section 5.

We have seen above that neural networks tend to have the problem of catastrophic interference and that measures have to be taken to avoid it. It is clear that the cortex is neither simply a feedforward network nor a Hopfield network. However, conceptually it shares important properties with artificial neural networks and will most likely have the same problem of catastrophic interference. If in particular we have in mind its ability to learn from examples by exploiting their regularities and thereby generalizing to new situations, we see that it would be useful to store examples temporarily and permit the cortex to perform some variant of interleaved training. The function as a temporary storage of patterns is usually assigned to the hippocampus with the entorhinal cortex serving as an interface between cortex and hippocampus. (There is actually a debate whether the hippocampus is temporarily or permanently necessary for episodic memories (Nadel & Moscovitch, 1997). However, this is not essential for our discussion here since we focus on en- and decoding and not on the storage of memories in hippocampus.)

Simply storing input patterns can be done in a Hopfield network. Characteristic for a Hopfield network is its highly recurrent connectivity. This fits well to the recurrent connectivity in the CA3 region of the hippocampus, and it is mainly for this reason that CA3 is often considered the actual site of storage in the hippocampus.

We have seen above that a Hopfield network has two versions of the catastrophic-interference problem, one because stored patterns are too similar, and one because too many patterns are stored. The first can be solved by recoding the input patterns such that they become pairwise orthogonal, at least approximately, which can be done by generating a sparse representation. This role is usually assigned to the dentate gyrus for two reasons. Firstly, it provides a strong input to CA3 and secondly it has a very sparse activity. The role of the dentate gyrus in generating a sparse representation has been theoretically studied in detail by O'Reilly & McClelland (1994). In order to use the CA3 memory efficiently, it is assumed that dentate gyrus must also learn to select the important dimensions of the patterns and discard others that are redundant, thereby performing a lossy compression.

There are different views on whether CA3 also suffers from the second version of the problem of catastrophic interference. Theoretically, it is simply a question of capacity. Either one assumes that CA3 has sufficient capacity for lifelong storage of patterns, then there is no interference problem

due to an exhausted capacity, or the capacity is not sufficient, then the measures described above have to be taken, so that old patterns are gradually forgotten to make room for new patterns. This issue does not effect our ideas about the functional role of neurogenesis much, so we will not discuss it further.

If one wants to retrieve the encoded patterns in CA3, one also needs a decoding network. This role is commonly assigned to CA1 (and the subiculum), mainly for the reason that there are strong connections from CA3 via CA1 and subiculum back to the entorhinal cortex. We too will refer to CA1 as the decoding network, but one should keep in mind that the exact location of decoding is not so clear.

In summary we have the following picture. The hippocampus serves as a temporary (or even permanent) storage that can store new patterns quickly. CA3 serves as the actual storage site and works like a Hopfield network. Dentate gyrus performs an encoding of the patterns to make them suitable for storage. This includes compression and sparsification. CA1 and subiculum perform a decoding of patterns retrieved from CA3. The entorhinal cortex serves as an interface between hippocampus and cortex.

So far we have discussed the model for a fixed environment, where only CA3 is plastic to store new patterns. In the following section we discuss how an additional catastrophic-interference problem arises if the model adapts to a changing environment and how adult neurogenesis might contribute to its solution.

4 Why new neurons?

What might be the role of new neurons in the dentate gyrus in the overall framework given above? Why are they generated only in the dentate gyrus and not in CA3 or CA1? Why is it sufficient to have a very low rate of adult neurogenesis over most of the lifetime? Why should adult neurogenesis be regulated by activity or richness of the environment?

4.1 Avoiding catastrophic interference in the dentate gyrus with new neurons

In the framework presented above, the dentate gyrus has the role of encoding input patterns coming from entorhinal cortex such that they can be efficiently stored in CA3. This includes sparsification and compression. The optimal encoding depends on the distribution of the input patterns or, in other words, on the environment the animal lives in. We assume that when the environment changes, the encoding should change, too, to guarantee an optimal encoding and efficient storage of new patterns.

However, as the dentate gyrus adapts to a new environment, the encoding learned in the preceding environment might get lost quickly due to catastrophic interference. A pattern that has been encoded and stored previously can no longer be recognized and retrieved, because the encoding has changed and the representation of the same pattern now looks very different as it arrives in CA3 (Treves & Rolls (1994) have suggested to perform the recognition and retrieval via the direct perforant-path connections from EC to CA3, which might reduce this problem but cause a more subtle version of catastrophic interference, since the perforant path could presumably accommodate only a very limited number of different DG-encodings). Note that this problem of catastrophic interference is distinct from the ones discussed in Section 3. Here the problem arises within the dentate gyrus due to an adaptation to a new environment while in Section 3 it was a problem within CA3 that arose if similar or too many patterns had to be stored.

The solution of interleaved training to avoid catastrophic interference cannot be applied, because

the animal does not have arbitrary access to the different stimuli to generate an interleaved stimulus sequence nor does the hippocampus have an intermediate buffer, which in any case would have the same problem of catastrophic interference again. Thus, a different strategy is needed.

French (1991) has shown that catastrophic interference in a feedforward network trained with the back-propagation algorithm can be significantly reduced by using a sparse representation in the hidden layer. This fits well to the sparse firing patterns found in the dentate gyrus and seems to be one possible strategy. However, this solution comes at the price that generalization is reduced because of the localizing effect of sparsification.

Here we want to consider an alternative or complementary hypothesis, namely that the problem of catastrophic interference in the dentate gyrus can be avoided by extending the encoding by adding new neurons instead of changing it by modifying existing synaptic weights; old neurons and their synaptic weights are fixed and preserve the old code while new neurons are added and provide those features that are novel. In Section 5 it will be shown with a computational model that this is indeed a possible strategy to avoid catastrophic interference and stabilize the encoding.

4.2 Why no new neurons in CA3 and CA1?

If we assume that the dentate gyrus has the problem of catastrophic interference and can solve it by adding new neurons, wouldn't neurogenesis also be helpful in CA3 and CA1?

If we take a Hopfield network as a good model of CA3, then it is clear it does not have the problem of catastrophic interference in the same way as the dentate gyrus. Adding a few new neurons in a Hopfield network would neither help much in making patterns orthogonal nor would it increase the capacity significantly. Therefore neurogenesis does not seem a good strategy in CA3.

The situation is different in CA1. Within the framework considered here, dentate gyrus and CA1 are complementary networks. One performs the encoding; the other performs the decoding approximately inverting the encoding. Due to the close relationship between en- and decoding, one can expect that CA1 has the problem of catastrophic interference to the same extent as the dentate gyrus. However, for exactly this reason it might be sufficient to solve the problem in the dentate gyrus alone. If the encoding is stabilized by neurogenesis, the optimal decoding learned by CA1 might be stabilized as well. If that is true, no new neurons are necessary in CA1. Following this argument, we assume that while dentate gyrus employs new neurons to avoid catastrophic interference, CA1 simply always adapts such that it optimally inverts the encoding performed by dentate gyrus.

In the model simulation in Section 5 we will see that this argument is not quite as straight forward as it appears here, because the optimal decoding depends not only on the encoding but also on the pattern statistics. But we will also see that it is basically valid in the case of neurogenesis. On the other hand, if one really wanted to employ new neurons in CA1 as well, it might be difficult to coordinate the learning in DG and CA1, for instance the new CA1 might need to 'know' which DG-neurons are old (fixed) and which ones are new (plastic) in order to take advantage of the new neurons in CA1, something that does not seem feasible.

4.3 How few new cells might make a big difference

In a naive view one could argue that new neurons have to be added whenever new stimuli occur and have to be stored. Thus the number of neurons should increase linearly with the number of stored patterns. Even if there is at present no definite evidence how many new neurons are actually generated in adulthood and in how far the net effect of adult neurogenesis is a turnover,

a cumulative growth, or a combination of both, there can be no doubt that the levels of adult neurogenesis after young-adulthood are very low. For our present study we have assumed that there is net growth and no turnover, even if the growth is very small for most part of the life. However, most parts of our model are not dependent on this assumption. What is essential is the experimental evidence that adult neurogenesis occurs at high rates early in life and then strongly declines to very low levels throughout the remaining life (Kronenberg et al., 2005) and that new neurons have a long survival time. To understand how few new cells might actually make a difference two things are important to consider: firstly, the location of adult neurogenesis in an anatomical bottleneck situation, where large amount of information has to pass through a small network (the hippocampus is very small compared to the cortical layers where the input comes from), and secondly, the distinction between *new* and *novel* stimuli. A new stimulus, in our view, is a stimulus that has not occurred before but is made up of features that are all known, much like a new word might occur in a text, not seen before but made up of letters that are all known. A novel stimulus, in contrast, is not only new but also contains features not encountered before, much like a Chinese character would be novel to people used to Latin characters. We assume that a new stimulus can be represented without difficulty with the existing code while a novel stimulus actually requires an extension of the code to represent the novel features. It seems reasonable to assume that new stimuli occur with almost constant rate throughout life while novel stimuli occur primarily early in life or if the environment changes significantly. Thus a very low rate of adult neurogenesis might be sufficient to account for novel stimuli occurring after the early phase of life.

This argumentation, of course, suffers from the difficulty to define what *new* and *novel* actually means. In our simulations new stimuli are new input vectors that lie within the region populated by previous stimuli, while novel stimuli are farther away and require the representation of a new dimension.

4.4 How is adult neurogenesis regulated?

New neurons are being generated in the dentate gyrus continually. Some of these neurons are integrated into the DG-network and become functional, but most of them are not used and are removed through apoptosis (cell death). Both, the generation of new neurons and their integration, are regulated and depend on behavior.

Neurogenesis is thus primarily regulated by two key mechanisms: firstly, control of the proliferation of precursor cells leading to an expansion of the pool of immature new neurons with the potential to become integrated, and secondly, control of the selective recruitment of new neurons or otherwise apoptotic elimination (Kempermann et al., 2004a). Many stimuli affect the proliferation phase of adult neurogenesis, most notably physical exercise (van Praag et al., 1999b) and seizures (Parent et al., 1997) but including many others (for a review see Lie et al., 2004). With respect to hippocampal function these stimuli are considered more or less non-specific. Exposure to an enriched environment (Kempermann et al., 1997b) or in some instances the application of specific learning stimuli (Gould et al., 1999; Dobrossy et al., 2003; Leuner et al., 2004) resulted in an increase in survival and neuronal recruitment. Such stimuli would thus be considered to be more specific. In some sense, environmental enrichment is a less reductionistic and more natural setting that incorporates specific and non-specific aspects, although the specific, survival-promoting stimuli seem to prevail. However we found that depending on the genetic background, environmental enrichment might not only exert a survival-promoting but also a pro-proliferative effect on adult neurogenesis (Kempermann et al., 1998a).

One problem of this regulation are the different time scales. Since the generation and integration of new neurons takes weeks, the animal cannot start generating new neurons when they are needed.

Instead, it must have some of them in stock all the time; the greater the chance of encountering novel stimuli, the more new neurons need to be held in reserve.

The different time scales of neurogenesis and behavior might explain two observations. Firstly, if we assume that new neurons have a limited life-span if they are not integrated in the network, then new neurons must be generated all the time and most of them are removed again just to have some available quickly when needed. Secondly, in order not to waste too many resources on the generation of new neurons, the rate of neurogenesis should depend on the likelihood of novel stimuli. However, since the generation of new neurons is such a slow process, the likelihood of novel stimuli has to be predicted on a fairly long time-scale of weeks. Thus, unspecific indicators such as the level of an animal's activity seem to be reasonably good predictors.

5 A simple computational model

In this section we present a very simple computational model to illustrate the main point of our hypothesis, namely that catastrophic interference in an encoding network can be avoided by keeping old units stable and adding new units, which are plastic. The model is not meant to be biologically plausible but to elucidate the mechanisms by which adult neurogenesis can prevent catastrophic interference in a network with a distributed representation. Compared to the general model sketched in Section 3, we make the following major simplifications:

- To make the model as simple as possible, so that it can be well understood and treated also analytically, we have disregarded sparsification, which we know might also play an important role in preventing catastrophic interference (see Sec. 4.1). We suggest neurogenesis as a complementary mechanism. Further modeling work will have to build on this and investigate the interactions between sparsification and neurogenesis. As a consequence of this simplification we use a small hidden layer (with sparsification the hidden layer would be larger but without necessarily increasing the effective dimensionality of the representation).
- Since we are not concerned with the problem of how CA3 itself stores and recalls patterns, we do not model its recurrent dynamics explicitly. Instead, for each network we compute the retrieval error analytically under the assumption of perfect storage and recall within CA3, i.e. we assume that the retrieval error only results from compression and a mismatch between en- and decoding during storage and retrieval.
- We make the simplifying assumption that the encoding performed by the dentate gyrus is done in the EC-to-DG connections and that the DG-to-CA3 connections can be approximated by a linear one-to-one connectivity. This allows us to model DG and CA3 with just one layer, since the activities in CA3 would be a copy of the DG-activities in any case.

Making these assumptions we arrive at a simple linear autoencoder network as described in detail below.

5.1 Methods

Network architecture

The model considered here is a linear autoencoder network; see Figure 1. Such a network consists of an input layer, a smaller hidden layer, and an output layer, which has as many units as the input layer. We identify the input layer with layer II of the entorhinal cortex (EC-II), the hidden

layer with the dentate gyrus (DG), and the output layer with layers V/VI of the entorhinal cortex (EC-V/VI). For storage and retrieval experiments we take the hidden layer also as a representation of CA3, implicitly making the simplifying assumption of a one-to-one connectivity between DG and CA3. The mapping realized by the connectivity from the hidden to the output layer is interpreted as the decoding performed by CA1 and subiculum.

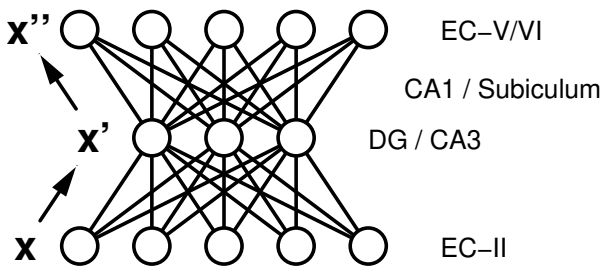


Figure 1: Architecture of the linear autoencoder network. The input and output layer represent layer II and layers V/VI of the entorhinal cortex, respectively. The hidden layer represents the dentate gyrus and in the retrieval experiments also CA3. CA1 and subiculum are not represented explicitly but summarized in the hidden-to-output-layer connections. The whole network is linear. In the simulations input and output layer each had 60 units and the hidden layer had 15 units plus 5 units due to neurogenesis.

The number of units in each layer is denoted by lowercase letters and a triangle ($\triangleright$) denotes the direction of connectivity from left to right. Thus $n \triangleright m \triangleright n$ indicates a network with n input units, m hidden units, and n output units. As mentioned above, an autoencoder network has always the same number of units in the input and output layer. The hidden layer may undergo neurogenesis in which case the network may change from $n \triangleright m \triangleright n$ to $n \triangleright (m + l) \triangleright n$, where l is the number of added units. This basic notation will be extended below to include information about the training procedure.

Network computation

The input layer represents the input patterns, denoted by a vector $\mathbf{x}$, and does not perform any computation. The hidden layer performs an encoding and the output layer performs a decoding. Both layers are linear, and so is the whole network, since a combination of linear functions is also linear. The representation of input vector $\mathbf{x}$ in the hidden layer (dentate gyrus) will be referred to as $\mathbf{x}'$; the reconstruction in entorhinal cortex will be referred to as $\mathbf{x}''$.

The name 'autoencoder network' indicates that the task of the network is simply to reproduce its input in the output layer. However, since the hidden layer has fewer units than the input and output layer, the network has a bottleneck, which enforces a dimensionality reduction and therefore a compression from the input to the output layer. Compression is one of the hypothesized functions of the dentate gyrus that is captured by the model. We do not consider sparsification, since this is a nonlinear operation, which is not within the scope of the model. Such a model may seem ridiculously simple, but we will see that it provides some interesting insights into the possible role of adult neurogenesis. The simplicity is actually a great advantage, because it permits an analytical treatment of the network (Wiskott et al., 2004).

Figure 2 (a) illustrates optimal en- and decoding for a given data distribution. It is known from principal component analysis that the encoding subspace as well as the decoding subspace should ideally lie in the direction of maximal variance (with identical orientation, of course).

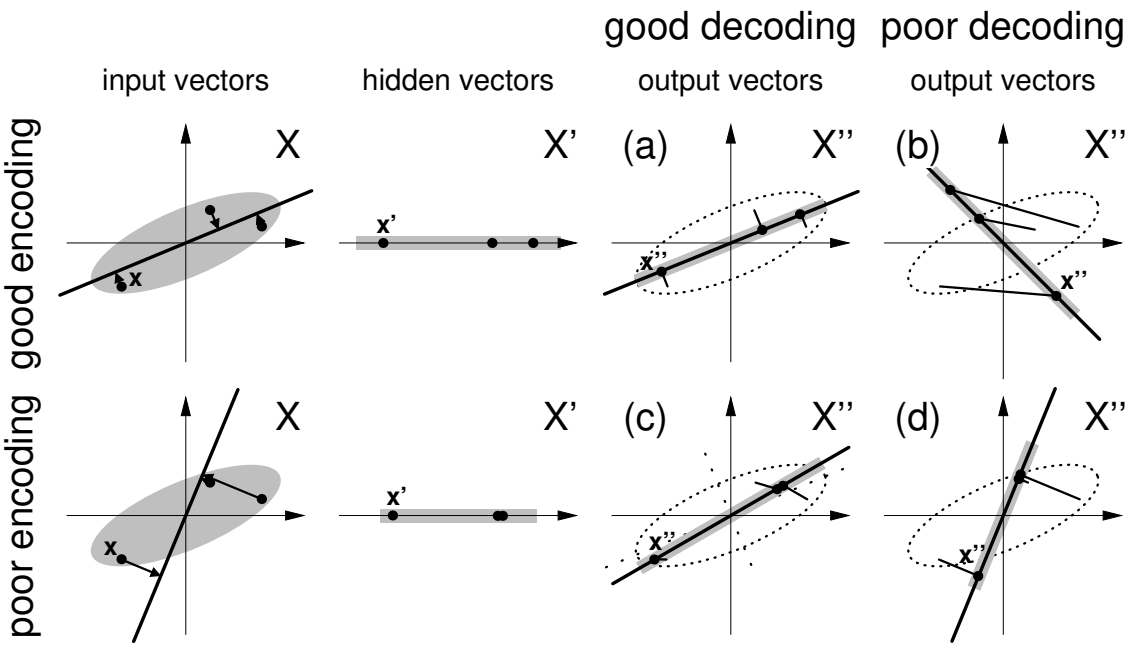


Figure 2: Different en- and decodings. Illustration of the computation performed by an autoencoder network with two input and output units and just one unit in the hidden layer, i.e. with the architecture $2 \triangleright 1 \triangleright 2$. **Column 1:** The stimuli are vectors $\mathbf{x}$ in a high-dimensional vector-space X. Here X is only two-dimensional. Vectors can be identified with their end-points, which are here visualized as big dots. The gray ellipses indicate the distribution of input vectors and three of them are shown explicitly (dots). During encoding, the two-dimensional distribution has to be reduced to a one-dimensional distribution, since there is only one unit in the hidden layer. In the linear case this corresponds to a projection (indicated by the arrows) onto a line (neglecting any scaling for simplicity and without loss of generality). **Top row:** In this example the line lies in the direction of maximal variance of the distribution, which is the optimal choice in order to minimize the reconstruction error. **Bottom row:** In this example the line has a random orientation corresponding to a poor encoding. **Column 2:** In the hidden layer, the stimuli are now represented by vectors $\mathbf{x}'$ in a lower-dimensional vector-space X', thus some information is lost. Here X' is only one-dimensional. **Columns 3 and 4:** In order to obtain a reconstruction of the original high-dimensional input-vectors, the vectors $\mathbf{x}'$ of the space X' are embedded in a high-dimensional vector-space X'', which can be identified with the original space X. The direction of the embedding is referred to as decoding direction. The recoding error made by the transformation from $\mathbf{x}$ to $\mathbf{x}''$ is indicated by the thin solid lines. **(a)** In case of optimal encoding, the optimal decoding-direction is again in the direction of maximal variance of the data distribution. **(b)** A random decoding-direction leads to a large error. **(c)** The optimal decoding-direction in case of poor encoding lies somewhere between the original encoding-direction and the direction of maximal variance. **(d)** Embedding the vectors in the direction of the original encoding may actually lead to a relatively large error.

If the encoding was not optimal, neither the original direction of encoding (Fig. 2 d) nor the direction of maximal variance are optimal for decoding. A direction in between these two cases is actually optimal including some scaling (Fig. 2 c). If the data distribution is more spherical with similar variance in all directions then the optimal decoding should lie close to the direction of the encoding; if the data distribution is more elongated then the optimal decoding should lie close to the direction of maximal variance (Rasch, 2003; Wiskott et al., 2004).

Finally, as Figure 2 (b) illustrates, if the encoding is optimal, but the decoding is random, the recoding error is large. It is often actually greater than if one would not decode anything, i.e. set all output vectors to zero. Thus, no decoding is better than random decoding.

Input data

We assume the animal lives in two different environments, which we denote by uppercase letters A and B. Within these environments the animal encounters different stimulus situations (or events), which give rise to different patterns of activity in EC-II representing the stimulus situation. Such patterns will be denoted by vectors $\mathbf{a}$ and $\mathbf{b}$ (instead of $\mathbf{x}$) for the environments A and B, respectively. The distribution of stimulus situations depends, of course, on the environment. We model this by drawing the input vectors $\mathbf{a}$ and $\mathbf{b}$ from two different multidimensional Gaussian distributions (we used 60 dimensions in the simulations). The distributions in the different environments differ only in their orientation in space. Thus, the distribution in environment B is a rotated version of the distribution in environment A. Mathematically this means that the eigenvalues of the covariance matrices are identical. Figure 3 illustrates data distributions of the two different environments and shows the eigenvalues of the actual distributions used in the simulation experiments.

Gaussian distributions are, of course, far from the true data distribution one would expect in the entorhinal cortex. But since the network is linear and the error function is the mean squared error (1, 2), the model is insensitive to higher-order moments. More realistically structured data distributions would lead to identical results as long as the eigenvalues of the covariance matrix are the same, and an exponential decay of eigenvalues, as used here, is rather generic.

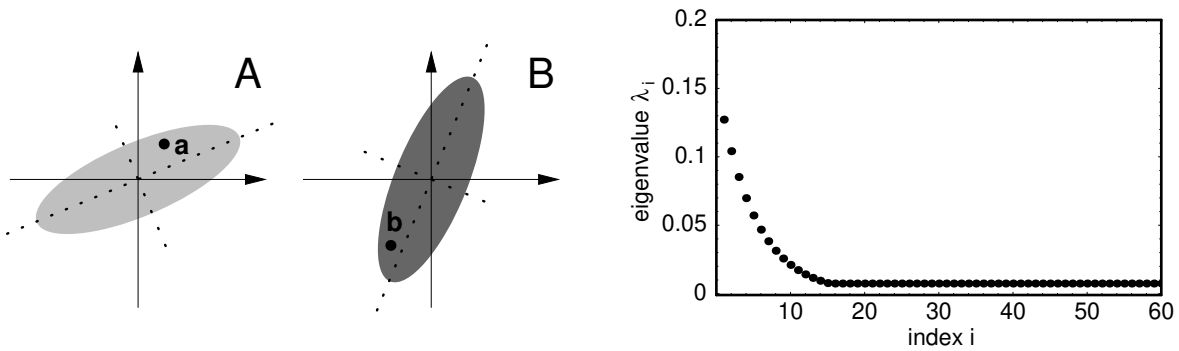


Figure 3: Distribution of input vectors a and b in environments A and B. Left: Schematic drawing of the two distributions. They differ only in their orientation in space but not in any other statistical characteristics. Dotted lines indicate the principal axes. **Right:** Eigenvalues of the covariance matrices of the input-vector distributions in environments A and B. They correspond to the variances along the principal axes of the distributions. The graph illustrates that out of the 60 principal components the first 15 carry most of the variance. The other 45 are interpreted as noise.

Network adaptation

Since the input layer only serves as a representation of the input and does not perform any computation, it is only the hidden layer and the output layer that can adapt to a particular environment. This will be denoted by subscripts to the indices of the number of units per layer. We assume that units in the hidden layer always adapt jointly with the output layer. Units in the output layer, however, can adapt with the connections from the input to the hidden units fixed. $n \triangleright m_A \triangleright n_A$, for instance, indicates that the hidden layer as well as the output layer are optimally adapted to environment A. $n \triangleright (m_A + l_B) \triangleright n_B$ indicates that an animal fully adapted to environment A has moved to environment B, where the hidden layer got l new units, which adapted to B along with the complete output layer. Note that the adaptation of the output layer to environment B depends on the hidden layer, so that the networks $n \triangleright m_A \triangleright n_B$ and $n \triangleright m_B \triangleright n_B$ actually have different output layers. In the results presented here we always impose the constraint that weight vectors of different hidden units are normalized to one and are orthogonal to each other, i.e. maximally different. Results for non-orthogonal weight vectors can be found in (Wiskott et al., 2004).

Storage and retrieval

Although our model does not include an explicit CA3 layer, we can still address the question of how well patterns can be stored and retrieved under different scenarios. We assume the role of a CA3 layer would only be to store patterns coming from the dentate gyrus and later recall them. Thus patterns in CA3 would be identical to those coming from the dentate gyrus, which is the reason why we do not need to represent CA3 by an extra layer. Storage with a network $n \triangleright m_A \triangleright n_A$ adapted to environment A and later retrieval with a network $n \triangleright m_B \triangleright n_B$ adapted to B is thus modeled by storing the vector $\mathbf{a}'$ encoded by the first network, indicated by $\mathbf{a}'_\square$, and then decoding the stored vector by the second network, indicated by $\mathbf{a}''_\square$. In case of neurogenesis, the number of units in the hidden layer changes. We take that into account by simply adding an appropriate number of zeros to the stored vector $\mathbf{a}'_\square$.

Performance measures

To assess the performance of the autoencoder network we determine two kinds of reconstruction errors. Firstly, we are interested in how well the patterns in EC-II are reconstructed in EC-V/VI and how large the error is due to the recoding. Thus, we compare the input vector $\mathbf{a}$ with the output vector $\mathbf{a}''$ and call the mean squared difference the recoding error. Secondly, we are interested in how well patterns stored in CA3 can be reconstructed when they are retrieved (spontaneously or by a cue). The corresponding error is called the retrieval error. To get reliable measures we average over the whole distributions (analytically) and, if applicable, over 5000 randomly rotated data distributions B (numerically) both indicated by angle-brackets $\langle\cdot\rangle$.

$$\text{recoding error:} \quad E := \left\langle |\mathbf{a} - \mathbf{a}''|^2 \right\rangle \tag{1}$$

$$\text{retrieval error:} \quad E_\square := \left\langle |\mathbf{a} - \mathbf{a}''_\square|^2 \right\rangle \tag{2}$$

We will also consider the recoding error for patterns $\mathbf{b}$.

5.2 Results

Assume our model animal first lives in environment A, then it moves to a new environment B, and finally it returns to the old environment A. For best performance in this setting the autoencoder

network should adapt somehow to the different pattern statistics of the two environments. In this section we consider three different adaptation strategies. In all three cases the decoding part of the network (hidden-to-output-layer connections) always fully adapts to the current environment to minimize the recoding error. The encoding part (input-to-hidden-layer connections), however, may have one of the following three strategies:

(a) **No DG-adaptation:** In this strategy (Tab. 1 left) we assume there is something like a critical period in which the animal adapts to environment A. After that, no adaptation takes place in the encoder part of the network anymore, so that the encoder does not change when the animal moves to environment B. The hidden layer has 20 units.

(b) **Neurogenesis:** In this strategy (Tab. 1 middle) the network starts with a few units in the hidden layer and new units are added as required. Only the most recent hidden units adapt to a new environment to account for those feature dimensions that are not yet well represented. Here we start with 15 hidden units that adapt to environment A and add five new units after the animal has moved to environment B.

(c) **Full adaptation:** In this strategy (Tab. 1 right) all neurons always fully adapt to the current environment, first A then B. The hidden layer has 20 units.

To assess the performance of the system we consider five different errors. The first two are determined for the network adapted to environment A and the other three for the network adapted to environment B.

(A) **Network adapted to environment A.**

(i) **Recoding error A:** When adapted to environment A the recoding error of patterns $\mathbf{a}$, i.e. $\left\langle |\mathbf{a} - \mathbf{a}''|^2 \right\rangle$, should obviously be small. The retrieval error for patterns $\mathbf{a}$ should also be small, but since the network has not changed yet, the retrieval error is the same as the recoding error and is not considered separately.

(ii) **Recoding error B:** When the animal has just moved to environment B it did not have the time yet to adapt to the new environment but ideally the recoding error for patterns $\mathbf{b}$, i.e. $\left\langle |\mathbf{b} - \mathbf{b}''|^2 \right\rangle$, should already be small right from the beginning, e.g. because of the experience gathered earlier in other environments. However, we will find that this is generally not the case in our simulations.

(B) **Network adapted to environment B.**

(iii) **Recoding error B:** Of course, performance improves with adaptation, so we are also interested in the recoding error for patterns $\mathbf{b}$, i.e. $\left\langle |\mathbf{b} - \mathbf{b}''|^2 \right\rangle$, after the network has adapted to environment B.

(iv) **Retrieval error A:** Next we ask how well the animal can retrieve memories that it has stored earlier in environment A. This is the retrieval error $\left\langle |\mathbf{a} - \mathbf{a}''_{\square}|^2 \right\rangle$.

(v) **Recoding error A:** Finally we want to look at the performance of the animal when it returns to environment A but does not have the time to newly adapt to it. This is measured by the recoding error $\left\langle |\mathbf{a} - \mathbf{a}''|^2 \right\rangle$.

Note that in (i) and (v) as well as in (ii) and (iii) the same recoding errors are computed, but that the values will differ because the network has changed due to adaptation.

Table 1 lists the five different errors for the three different adaptation strategies described above. As one would expect, if a network is fully adapted to environment A, recoding error A (i) is small and recoding error B (ii) is large. Furthermore, performance is obviously better if the network has more units in the hidden layer; compare (a, c) with (b). In fact the error would go down to zero if the network had as many hidden units as input or output units, because then no dimensions would be lost in the hidden layer.

With adaptation to environment B, the performance on patters $\mathbf{b}$ improves, of course; compare recoding error B before (ii) and after (iii) adaptation. The more complete the adaptation the better the performance, thus no adaptation (a) is worst and full adaptation (c) is best; neurogenesis (b) lies in between. In the no-adaptation case (a) the encoding learned for environment A is actually random with respect to the distribution of patterns $\mathbf{b}$, because the distributions in the two environments are randomly rotated relative to each other. Thus this case (a-iii) corresponds to the situation of random encoding and optimal decoding illustrated in Figure 2 (c).

So far the results were as one would expect. The interesting question now is how the networks adapted to environment B perform on patterns $\mathbf{a}$; see (iv, v). With no DG-adaptation (a) performance degrades significantly due to catastrophic interference within the decoding part of the network; compare (iv, v) with (i). Recoding and retrieval error are identical, because the encoding part of the network has not changed. Note that in this case decoding changes even though encoding is fixed, because the pattern statistics of the input has changed. Thus stabilizing the encoding does not necessarily stabilize the optimal decoding as was argued in Section 4. But see below and Figure 4 for why the argument basically holds in the neurogenesis case.

With full adaptation (c), the effect of catastrophic interference is even more severe. The recoding error A (v) increases to the level of the recoding error B (ii) of the network adapted to environment A, which is clear for symmetry reasons. The retrieval error A even increases beyond the value of 1, indicating that the reconstructed output is not only not helpful but even misleading. The reason for this is that the representation of the hidden layer has changed completely, so that the stored patterns $\mathbf{a}'_{\square}$ cannot be interpreted anymore. The network effectively hallucinates random output patterns, which leads to the large retrieval error. This is equivalent to the case of optimal encoding and random decoding discussed in Figure 2 (b). This dramatic effect of catastrophic interference is partly due to the fact that we find the optimal weights analytically, so that the new weights are chosen independently of the old ones. With an incremental learning rule one would expect that the old weights bias the solution for the new ones, so that interference would be somewhat reduced.

With neurogenesis (b), the network can largely avoid the effect of catastrophic interference and achieves a performance that is better than that of the two other strategies.

Figure 4 illustrates catastrophic interference and how neurogenesis can avoid it. If the final number of units is limited, the idea is that instead of wasting the units early on to represent also the low-variance directions of distribution A, one starts with fewer units and adds new units later to represent high-variance directions of distribution B without the effect of catastrophic interference.

6 Discussion

6.1 Hypothesis and model

In this paper we have addressed the question of what the functional role of adult neurogenesis in the dentate gyrus is. As a basis for our discussion we have largely adopted one of the standard models of hippocampal function (Sec. 3) based on concepts from neural network theory (Sec. 2). In this model the dentate gyrus serves as an encoder, the CA3-region as a memory, and the

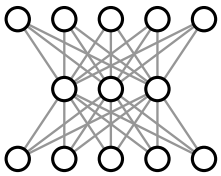
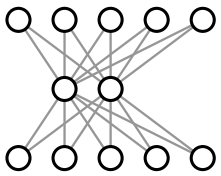
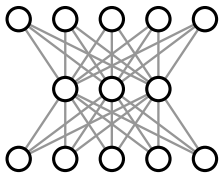
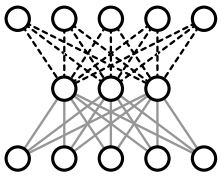
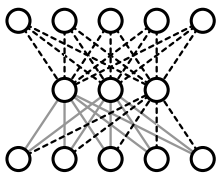
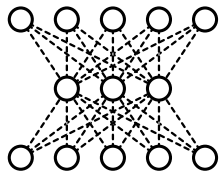
		(a) no DG-adaptation	(b) neurogenesis	(c) full adaptation
(A)	network adapted to environment A	 $n \triangleright (m+l)_A \triangleright n_A$	 $n \triangleright m_A \triangleright n_A$	 $n \triangleright (m+l)_A \triangleright n_A$
(i)	recoding error A $\langle \lvert \mathbf{a} - \mathbf{a}'' \rvert^2 \rangle$	0.30	0.33	0.30
(ii)	recoding error B $\langle \lvert \mathbf{b} - \mathbf{b}'' \rvert^2 \rangle$	0.67	0.75	0.67
(B)	network adapted to environment B	 $n \triangleright (m+l)_A \triangleright n_B$	 $n \triangleright (m_A + l_B) \triangleright n_B$	 $n \triangleright (m+l)_B \triangleright n_B$
(iii)	recoding error B $\langle \lvert \mathbf{b} - \mathbf{b}'' \rvert^2 \rangle$	0.44	0.36	0.30
(iv)	retrieval error A $\langle \lvert \mathbf{a} - \mathbf{a}''_{\square} \rvert^2 \rangle$	0.61	0.45	1.69
(v)	recoding error A $\langle \lvert \mathbf{a} - \mathbf{a}'' \rvert^2 \rangle$	0.61	0.41	0.67

Table 1: Five different errors (i–v) for three different adaptation strategies (a–c); see text for details. In the network icons connections adapted to environment A and B are drawn as solid (grey) and dashed (black) lines, respectively. An error value of 0 would indicate no error, i.e. perfect recoding or retrieval, and a value of 1 would be obtained if a network would always return a zero vector, i.e. if it would not do anything. Thus a value greater 1 indicates that the network produces misleading outputs. If not computed completely analytically, each error is an average over 5000 simulation runs with randomly rotated data distributions; standard deviations were between 0 and 0.05. Rows (iv, v) show that neurogenesis (b) indeed avoids the problem of catastrophic interference with still reasonable performance in cases (i, iii). None of the networks performs well in case (ii). If in the neurogenesis case (b) the weight vectors are not constrained to be orthogonal but would always be along the principal axes, only the retrieval error A would change from 0.45 to 0.56, the other errors (iii, v) would remain the same.

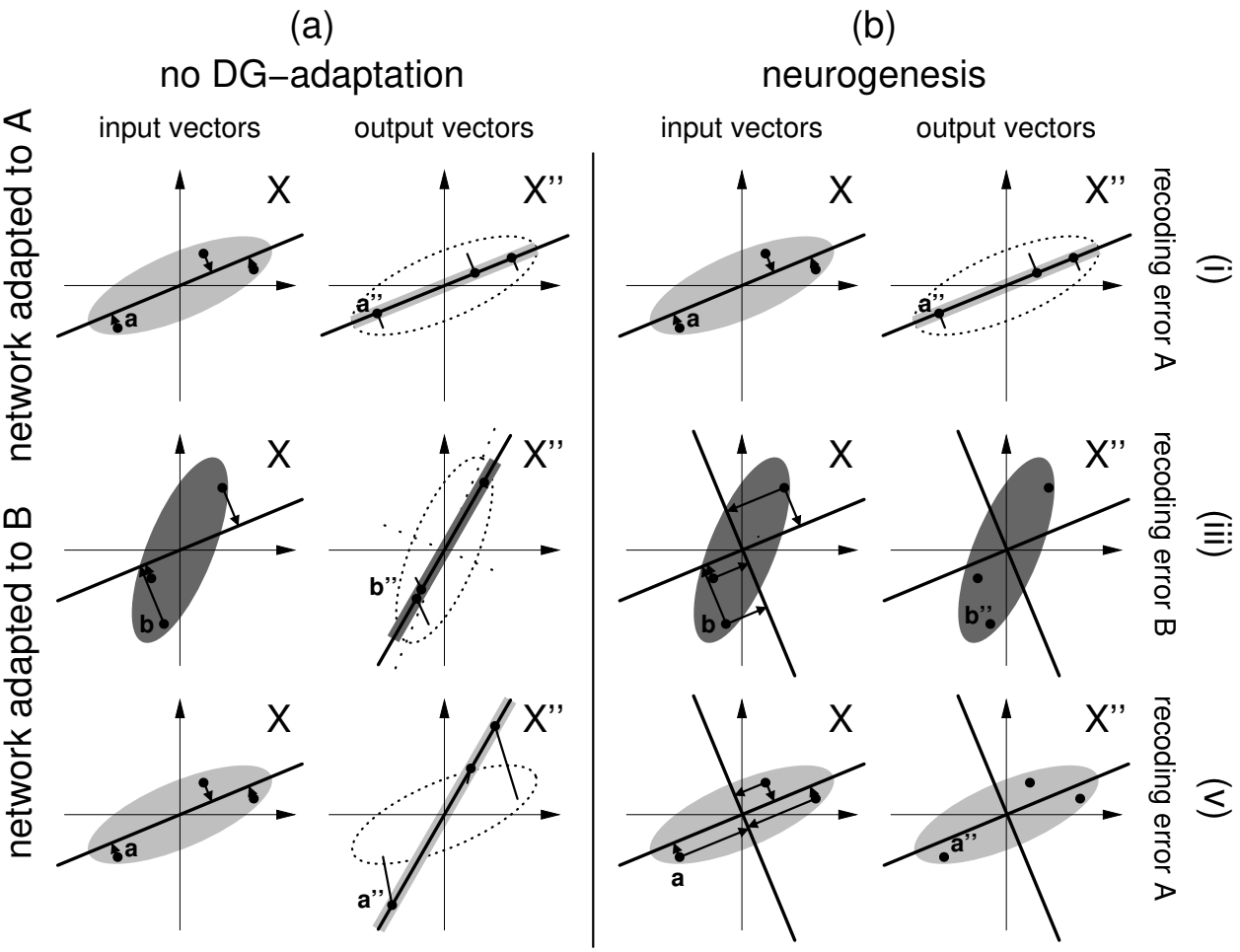


Figure 4: Effect of catastrophic interference in the no-DG-adaptation scenario and how the problem is avoided by neurogenesis. See next page for figure caption.

Figure 4: Effect of catastrophic interference in the no-DG-adaptation scenario and how the problem is avoided by neurogenesis. In contrast to previous figures, the 2D-plots represent a suitable two-dimensional projection of a high-dimensional space of input (and output) vectors. Otherwise, the figure uses the same conventions as Figure 2. In the top row the two networks are adapted to environment A and recoding error A is considered; in the middle and bottom row they are adapted to B and the recoding errors A and B, respectively, are considered. **No DG-adaptation, Panel (a-i):** In the network adapted to environment A the encoding as well as the decoding is optimal for input vectors **a** taken from distribution A. The recoding error A is correspondingly low (right). **Panel (a-iii):** If the encoding is fixed and no new units can be added (no DG-adaptation), the network can only optimize the decoding to environment B (right). This corresponds to the case of random encoding and optimal decoding already illustrated in Figure 2 (c). The network reduces the recoding error by misusing the encoding dimension (left) optimized for A to represent the high-variance direction of distribution B (right). **Panel (a-v):** Now that the decoding is optimized for B the decoding of vectors **a** is poor (right) and the recoding error A correspondingly large. **Neurogenesis, Panel (b-i):** In these 2D-plots the initial network in the neurogenesis scenario looks identical to that in the no-DG-adaptation scenario. However, note that the number of hidden units is smaller, so that there are some dimensions (with low variance in distribution A) not shown here that are less well represented than in the no-DG-adaptation scenario. **Panel (b-iii):** The old hidden units are fixed, like in the no-DG-adaptation scenario, but the network can add new hidden units (only one is shown here), which adapt to environment B. Thus the encoding gets extended by dimensions that are not yet represented but have high variance in distribution B (left). This reduces (in this figure eliminates) the misuse of encoding dimensions optimized for distribution A (right) and also improves the recoding of vectors **b**. **Panel (b-v):** The good decoding of vectors **a** is largely preserved (right). In this figure the recoding actually looks perfect, since we now have two hidden units representing the two plotted dimensions. Other dimensions (with lower variance) not plotted still suffer from the misuse of encoding dimensions and contribute to the recoding error, but to a much lesser degree than in the no-DG-adaptation scenario.

CA1/subiculum-system as a decoder. We assume that the dentate gyrus, like most neural systems, suffers from the problem of catastrophic interference when adapting to new environments. Our basic hypothesis for the role of adult neurogenesis is that new neurons help the dentate gyrus avoiding the problem of catastrophic interference by keeping the old neurons, which are adapted to earlier environments, fixed and adding new neurons, which are more plastic and can code for those aspects that are qualitatively new in the current environment.

We have used a linear auto-encoder network-model to illustrate and quantify the advantage new neurons might offer according to our hypothesis. This model is very simple and obviously neglects many prominent features of the biological system. Nevertheless, it demonstrates and explains several aspects of our hypothesis and yields some non-trivial insights.

Let us summarize some of the key features of our hypothesis and the network model.

- Our hypothesis requires that old neurons are relatively stable while new neurons are plastic. This is consistent with physiological findings (see Sec. 1).
- The results summarized in Table 1 show that adult neurogenesis can indeed reduce the effect of catastrophic interference significantly; see rows (iv) and (v) and compare the neurogenesis strategy (b) with the other two (a, c).
- The results of the no-adaptation strategy show that even if the encoding optimized for environment A is kept fixed does the performance on patterns **a** degrade drastically if the decoding is optimized for environment B. This is a somewhat surprising result and a consequence of the misuse of high-variance A-dimensions for representing **b**-patterns, as illustrated in Figure 4. This misuse is reduced significantly by adding new neurons.
- In the neurogenesis scenario in order to achieve a small recoding error a large number of new hidden units is necessary initially (in environment A) but fewer new units suffice later (in B). This can be easily shown mathematically and can also be seen in the simulation results. Compare entry (b-i) with entry (b-iii) in Table 1; the former required 15 new units, the latter only 5 new units, but both errors are relatively small. Thus in the model the effect of neurogenesis accumulates and with increasing 'age' the need for new units decreases, which is consistent with neuroanatomical observations.
- According to our hypothesis CA1 always fully adapts to optimize the decoding for the current environment. We have made this assumption mainly because we did not see a feasible way of coordinating learning in dentate gyrus and CA1 in a more specific way. However, the simulations at least partially justify this assumption, because in the neurogenesis case performance is good with full CA1 adaptation, see Table 1. The strong plasticity observed in CA1 physiologically also seems to be consistent with our assumption.
- According to our hypothesis new neurons are only added when the pattern statistics changes, i.e. it depends on the environment. However, the generation of new neurons takes too much time to initiate neurogenesis exactly when needed. Instead, there must always be new neurons available that could be used quickly. This is consistent with the fact that most new neurons are not integrated but die after some time.

6.2 Comparison with other models

There are a number of artificial neural networks that employ new units (artificial neurons) as part of their learning strategy. A prominent example in the field of vector quantization is the adaptive resonance theory (ART) (Carpenter & Grossberg, 1987; Carpenter et al., 1991).

In the ART-network new units are added as new patterns have to be learned that cannot be well represented by the existing units. This is similar in spirit to the model we propose, but the representation is very different. In the ART-network a pattern is represented by a single unit while in our model a pattern is represented with a population code.

Self-organizing maps (SOM) differ from vector quantization networks in that also the topological neighborhood structure of the units is being represented. SOM-networks that employ new units have been extensively studied by Fritzke (1994). In his growing cell structures new units can be added to represent new patterns, much like in the ART-network, or to refine the representation. Again, patterns are represented by single units and not populations of units.

Vector quantization and self-organizing maps are examples of unsupervised learning tasks. For feedforward networks supervised learning algorithms are of particular interest. Also for this class of networks do algorithms exist that successively add new units to improve the performance (see Bishop, 1995, sec. 9.5, for an overview). The model most closely related to ours is probably the cascade-correlation learning architecture (Fahlman & Lebiere, 1990). This architecture is a feedforward network with initially no hidden unit. After having trained this reduced network, hidden units are added one by one with repeated retraining of the new units and keeping old units fixed. Like our model this network learns a distributed representation. The main difference is that in the cascade-correlation network new units are added to improve the performance on identical training data while we have considered the problem of extending a representation to new patterns without catastrophic interference, a problem not addressed by Fahlman & Lebiere (1990).

There are also a few more biologically motivated models. Cecchi et al. (2001) have proposed a model for neurogenesis in the olfactory bulb. They used laterally connected inhibitory units to perform an orthogonalization of odor representations. Neurogenesis was assumed to occur at a constant rate while survival of an inhibitory unit was taken to depend on activity and was highest for units establishing an inhibitory interaction between correlated units. This way correlations between output units were reduced and the network converged to a 1-of-N code. In contrast to our model, this model reduces interference for a given set of patterns and does not address the issue of adaptation to a changing input statistics.

Deisseroth et al. (2004) and Chambers et al. (2004) have independently investigated the effect of neurogenesis in a nonlinear three-layer network trained to learn hetero-associations (input and output patterns are different). The authors make the assumption that there is a turnover of units in the hidden layer, i.e. some units die and are replaced by the same number of new units. Such a network tends to forget old memories more quickly but is able to store new memories without the effect of catastrophic interference. This is a convincing role for neurogenesis if one assumes neural turnover. It would be interesting, however, to see how it compares to simple weight decay as another means to achieve the same goal (Howard et al., 1987, cited in Amit, 1989).

The most detailed functional model of neurogenesis currently is probably the one presented by Becker (2005). This model is based on the assumption that the encoding and storage of patterns is done via the dentate gyrus while the retrieval is mediated by the perforant path connections to CA3. A neural turnover in the dentate gyrus is used to change the encoding over time to represent similar patterns differently and thereby avoid their interference. This is an interesting proposal in direct contrast to our hypothesis that the encoding of the dentate gyrus should be stabilized.

Finally, it should be mentioned that Lehmann et al. (2005) have taken a very different perspective and have considered the role of neurogenesis for the stability of the network dynamics due to homeostatic processes. They do not consider functional implications.

6.3 Future perspectives

The model presented here is obviously only a very first step. It illustrates our hypothesis about the functional role of adult neurogenesis and has the advantage of being simple, which provides insights that might not be so easily accessible with a more complex model. However, it neglects a number of important aspects of the hippocampal formation and needs to be extended in several ways. Let us list only a few important extensions:

- Sparsification has been disregarded in our computational model but is thought to play an important role in orthogonalizing patterns for storage in CA3 and it might also help to prevent catastrophic interference in the dentate gyrus. A refined model should have a larger hidden layer for the dentate gyrus and use a sparse code, more consistent with the neuroanatomy. It will be important to investigate the interactions between sparseness and neurogenesis as two strategies for avoiding catastrophic interference.
- CA3 not only receives input via the dentate gyrus but also directly from entorhinal cortex. It is an open question how these two very different pathways complement each other. One hypothesis is that they serve different modes of operation such as storage and retrieval (Treves & Rolls, 1992; Becker, 2005); another hypothesis states that they help learning to predict events (Borisyuk et al., 1999). We will have to reconsider such hypotheses with respect to neurogenesis and extend the model correspondingly with an explicit layer for CA3.
- Another interesting direction is to consider how neurogenesis is regulated and how this depends on the functional requirements.

The most obvious experimental prediction to be made from our model is that after consecutive exposure to two different novel environments A and B, each of which should have a measurable effect on adult neurogenesis, return to environment A should lack this regulatory effect of adult neurogenesis because it is not novel any more. We have begun pilot studies for this experiment to identify sufficiently different environments to yield enough statistical power to detect a difference in the number of recruited new granule cells after consecutive exposure. After experimental blockade of adult neurogenesis, e.g. by means of a pharmacological, irradiative, or preferably transgenic approach, our model would predict that the ability to identify novel stimuli or environments as novel would be reduced and the computational expenses to deal with novel stimuli would be increased. These might be measurable.

Acknowledgments

This work has been supported by grants from the Volkswagen Foundation to LW and GK. We thank two anonymous reviewers for helpful comments.

References

Altman, J. and Das, G. D. (1965). Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *The Journal of Comparative Neurology*, 124:319–335.

Amaral, D. G., Ishizuka, N., and Claiborne, B. J. (1990). Neurons, numbers and the hippocampal network. *Progress in Brain Research*, 83:1–11.

Amaral, D. G. and Witter, M. P. (1989). The three-dimensional organization of the hippocampal formation: A review of anatomical data. *Neuroscience*, 31(3):571–591.

Amit, D. J. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, Cambridge, UK.

Ans, B. and Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie*, 320(12):989–997.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J., Leonard, B. W., and Lin, L. H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research*, 83:287–300.

Bayer, S. A., Yackel, J. W., and Puri, P. S. (1982). Neurons in the rat dentate gyrus granular layer substantially increase during juvenile and adult life. *Science*, 216(4548):890–892.

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, 15(6):722–738.

Bentz, H. J., Hagstroem, M., and Palm, G. (1989). Information storage and effective data retrieval in sparse matrices. *Neural Networks*, 2:289–293.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Borisyuk, R., Denham, M., Denham, S., and Hoppensteadt, F. (1999). Computational models of predictive and memory-related functions of the hippocampus. *Reviews in the Neurosciences*, 10(3-4):213–232.

Boss, B. D., Peterson, G. M., and Cowan, W. M. (1985). On the number of neurons in the dentate gyrus of the rat. *Brain Research*, 338(1):144–150.

Cameron, H. A. and McKay, R. D. (1999). Restoring production of hippocampal neurons in old age. *Nature Neuroscience*, 2:894–897.

Cameron, H. A. and McKay, R. D. (2001). Adult neurogenesis produces a large pool of new granule cells in the dentate gyrus. *The Journal of Comparative Neurology*, 435(4):406–417.

Carpenter, G. A. and Grossberg, S. (1987). ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930.

Carpenter, G. A., Grossberg, S., and Rosen, D. B. (1991). ART 2-A; an adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4(4):493–504.

Cecchi, G. A., Petreanu, L. T., Alvarez-Buylla, A., and Magnasco, M. O. (2001). Unsupervised learning and adaptation in a model of adult neurogenesis. *The Journal of Computational Neuroscience*, 11(2):175–182.

Chambers, R. A., Potenza, M. N., Hoffman, R. E., and Miranker, W. (2004). Simulated apoptosis/neurogenesis regulates learning and memory capabilities of adaptive neural networks. *Neuropsychopharmacology*, 29(4):747–758.

Deisseroth, K., Singla, S., Toda, H., Monje, M., Palmer, T. D., and Malenka, R. C. (2004). Excitation-neurogenesis coupling in adult neural stem/progenitor cells. *Neuron*, 42(4):535–552.

Dobrossy, M. D., Drapeau, E., Aurousseau, C., Le Moal, M., Piazza, P. V., and Abrous, D. N. (2003). Differential effects of learning on neurogenesis: learning increases or decreases the number of newly born cells depending on their birth date. *Molecular Psychiatry*, 8:974–982.

Eriksson, P. S., Perfilieva, E., Bjork-Eriksson, T., Alborn, A. M., Nordborg, C., Peterson, D. A., and Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine*, 4(11):1313–1317.

Fahlman, S. E. and Lebiere, C. (1990). The cascade-correlation learning architecture. In Touretzky, D., editor, *Advances in Neural Information Processing Systems 2 (NIPS'1989)*, pages 524–532. Morgan-Kaufmann.

Frank, L. M., Brown, E. N., and Wilson, M. A. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, 27(1):169–178.

French, R. M. (1991). Using semi-distributed representationgs to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th Annual Cognitive Science Society Conference, Hillsdale, NJ:LEA*, pages 173–178.

French, R. M. (1997). Pseudo-recurrent connectionist networks: An approach to the "sensitivity-stability" dilemma. *Connection Science*, 9:353–379.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Fritzke, B. (1994). Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58.

Gerstner, W. and Kistler, W. (2002). *Spiking Neuron Models*. Cambridge University Press, Cambridge, UK.

Gluck, M. A. and Myers, C. E. (2001). *Gateway to memory*. MIT Press, Cambridge, MA, USA.

Gothard, K. M., Hoffman, K. L., Battaglia, F. P., and McNaughton, B. L. (2001). Dentate gyrus and CA1 ensemble activity during spatial reference frame shifts in the presence and absence of visual input. *The Journal of Neuroscience*, 21(18):7284–7292.

Gould, E., Beylin, A., Tanapat, P., Reeves, A. J., and Shors, T. J. (1999). Learning enhances adult neurogenesis in the hippocampal formation. *Nature Neuroscience*, 2(3):260–265.

Gould, E. and Gross, C. G. (2002). Neurogenesis in adult mammals: some progress and problems. *The Journal of Neuroscience*, 22(3):619–623.

Gould, E., McEwen, B. S., Tanapat, P., Galea, L. A., and Fuchs, E. (1997). Neurogenesis in the dentate gyrus of the adult tree shrew is regulated by psychosocial stress and NMDA receptor activation. *The Journal of Neuroscience*, 17(7):2492–2498.

Gould, E., Tanapat, P., McEwen, B. S., Flugge, G., and Fuchs, E. (1998). Proliferation of granule cell precursors in the dentate gyrus of adult monkeys is diminished by stress. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6):3168–3171.

Gould, E., Vail, N., Wagers, M., and Gross, C. G. (2001). Adult-generated hippocampal and neocortical neurons in macaques have a transient existence. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10910–10917.

Hammond, C. (2001). *Cellular and molecular neurobiology*. Academic Press, San Diego.

Harding, A. J., Halliday, G. M., and Kril, J. J. (1998). Variation in hippocampal neuron number with age and brain volume. *Cerebral Cortex*, 8(8):710–718.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Howard, R. E., Schwartz, D., Denker, J. S., Epworth, R. W., Graf, H. P., Hubbard, W. E., Jackel, L. D., Straughn, B. L., and Tennant, D. M. (1987). An associative memory based on an electronic neural network architecture. *IEEE Trans. ED*, 34:1553.

Jessberger, S. and Kempermann, G. (2003). Adult-born hippocampal neurons mature into activity-dependent responsiveness. *European Journal of Neuroscience*, 18(10):2707–2712.

Jung, M. W. and McNaughton, B. L. (1993). Spatial selectivity of unit activity in the hippocampal granular layer. *Hippocampus*, 3(2):165–182.

Kempermann, G., Brandon, E. P., and Gage, F. H. (1998a). Environmental stimulation of 129/SvJ mice results in increased cell proliferation and neurogenesis in the adult dentate gyrus. *Current Biology*, 8:939–942.

Kempermann, G. and Gage, F. H. (2002). Genetic determinants of adult hippocampal neurogenesis correlate with acquisition, but not probe trial performance, in the water maze task. *European Journal of Neuroscience*, 16(1):129–136.

Kempermann, G., Gast, D., and Gage, F. H. (2002). Neuroplasticity in old age: sustained fivefold induction of hippocampal neurogenesis by long-term environmental enrichment. *Annals of Neurology*, 52(2):135–143.

Kempermann, G., Gast, D., Kronenberg, G., Yamaguchi, M., and Gage, F. H. (2003). Early determination and long-term persistence of adult-generated new neurons in the hippocampus of mice. *Development*, 130(2):391–399.

Kempermann, G., Jessberger, S., Steiner, B., and Kronenberg, G. (2004a). Milestones of neuronal development in the adult hippocampus. *Trends in Neuroscience*, 27:447–452.

Kempermann, G. and Kronenberg, G. (2003). Depressed new neurons? — Adult hippocampal neurogenesis and a cellular plasticity hypothesis of major depression. *Biological Psychiatry*, 54(5):499–503.

Kempermann, G., Kuhn, H. G., and Gage, F. H. (1997a). Genetic influence on neurogenesis in the dentate gyrus of adult mice. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19):10409–10414.

Kempermann, G., Kuhn, H. G., and Gage, F. H. (1997b). More hippocampal neurons in adult mice living in an enriched environment. *Nature*, 386(6624):493–495.

Kempermann, G., Kuhn, H. G., and Gage, F. H. (1998b). Experience-induced neurogenesis in the senescent dentate gyrus. *The Journal of Neuroscience*, 18(9):3206–3212.

Kempermann, G. and Wiskott, L. (2004). What is the functional role of new neurons in the adult dentate gyrus? In Gage, F. H., Björklund, A., Prochiantz, A., and Christen, Y., editors, *Proc. Stem Cells in the Nervous System: Functional and Clinical Implications, Paris, January 20, 2003*, Research and Perspectives in Neurosciences (Fondation Ipsen), pages 57–65, Berlin. Springer.

Kempermann, G., Wiskott, L., and Gage, F. H. (2004b). Functional significance of adult neurogenesis. *Curr. Opin. Neurobiol.*, 14:186–191.

Kronenberg, G., Bick-Sander, A., Bunk, E., Wolf, C., Ehninger, D., and Kempermann, G. (2005). Physical exercise prevents age-related decline in precursor cell activity in the mouse dentate gyrus. *Neurobiology of Aging*.

Kronenberg, G., Reuter, K., Steiner, B., Brandt, M. D., Jessberger, S., Yamaguchi, M., and Kempermann, G. (2003). Subpopulations of proliferating cells of the adult hippocampus respond differently to physiologic neurogenic stimuli. *The Journal of Comparative Neurology*, 467(4):455–463.

Lee, J. H., Duan, W., Long, J. M., Ingram, D. K., and Mattson, M. P. (2000). Dietary restriction increases the number of newly generated neural cells, and induces BDNF expression, in the dentate gyrus of rats. *Journal of Molecular Neuroscience*, 15(2):99–108.

Lehmann, K., Butz, M., and Teuchert-Noodt, G. (2005). Offer and demand: proliferation and survival of neurons in the dentate gyrus. *European Journal of Neuroscience*, 21(12):3205–3216.

Leuner, B., Mendolia-Loffredo, S., Kozorovitskiy, Y., Samburg, D., Gould, E., and Shors, T. J. (2004). Learning enhances the survival of new neurons beyond the time when the hippocampus is required for memory. *The Journal of Neuroscience*, 24:7477–7481.

Lie, D. C., Song, H., Colamarino, S. A., Ming, G. L., and Gage, F. H. (2004). Neurogenesis in the adult brain: new strategies for central nervous system diseases. *Annual Review of Pharmacology and Toxicology*, 44:399–421.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

Nadel, L. and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*, 7(2):217–227.

Nilsson, M., Perfilieva, E., Johansson, U., Orwar, O., and Eriksson, P. S. (1999). Enriched environment increases neurogenesis in the adult rat dentate gyrus and improves spatial memory. *Journal of Neurobiology*, 39(4):569–578.

O'Reilly, R. C. and McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus*, 4(6):661–682.

Parent, J. M., Yu, T. W., Leibowitz, R. T., Geschwind, D. H., Sloviter, R. S., and Lowenstein, D. H. (1997). Dentate granule cell neurogenesis is increased by seizures and contributes to aberrant network reorganization in the adult rat hippocampus. *The Journal of Neuroscience*, 17(10):3727–3738.

Rasch, M. (2003). Modellierung adulter Neurogenese im Hippocampus [Modeling adult neurogenesis in the hippocampus]. Diploma thesis, Institute for Biology, Humboldt University Berlin, D-10115 Berlin, Germany.

Santarelli, L., Saxe, M., Gross, C. G., Surget, A., Battaglia, F. P., Dulawa, S., Weisstaub, N., Lee, J. H., Duman, R. S., Arancio, O., Belzung, C., and Hen, R. (2003). Requirement of hippocampal neurogenesis for the behavioral effects of antidepressants. *Science*, 301(5634):805–809.

Schmidt-Hieber, C., Jonas, P., and Bischofberger, J. (2004). Enhanced synaptic plasticity in newly generated granule cells of the adult hippocampus. *Nature*, 429(6988):184–187.

Seki, T. and Arai, Y. (1995). Age-related production of new granule cells in the adult dentate gyrus. *NeuroReport*, 6(18):2479–2482.

Shors, T. J., Miesegaes, G., Beylin, A., Zhao, M., Rydel, T. A., and Gould, E. (2001). Neurogenesis in the adult is involved in the formation of trace memories. *Nature*, 410:372–376.

Shors, T. J., Townsend, D. A., Zhao, M., Kozorovitskiy, Y., and Gould, E. (2002). Neurogenesis may relate to some but not all types of hippocampal-dependent learning. *Hippocampus*, 12(5):578–584.

Snyder, J. S., Hong, N. S., McDonald, R. J., and Wojtowicz, J. M. (2005). A role for adult neurogenesis in spatial long-term memory. *Neuroscience*, 130(4):843–852.

Snyder, J. S., Kee, N., and Wojtowicz, J. M. (2001). Effects of adult neurogenesis on synaptic plasticity in the rat dentate gyrus. *Journal of Neurophysiology*, 85(6):2423–2431.

Tanapat, P., Hastings, N. B., Rydel, T. A., Galea, L. A., and Gould, E. (2001). Exposure to fox odor inhibits cell proliferation in the hippocampus of adult rats via an adrenal hormone-dependent mechanism. *The Journal of Comparative Neurology*, 437(4):496–504.

Treves, A. and Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, 2(2):189–199.

Treves, A. and Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391.

van Praag, H., Christie, B. R., Sejnowski, T. J., and Gage, F. H. (1999a). Running enhances neurogenesis, learning, and long-term potentiation in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 96(23):13427–13431.

van Praag, H., Kempermann, G., and Gage, F. H. (1999b). Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nature Neuroscience*, 2(3):266–270.

van Praag, H., Schinder, A. F., Christie, B. R., Toni, N., Palmer, T. D., and Gage, F. H. (2002). Functional neurogenesis in the adult hippocampus. *Nature*, 415(6875):1030–1034.

Wang, S.-S., Scott, B. W., and Wojtowicz, J. M. (2000). Heterogenous properties of dentate granule neurons in the adult rat. *Journal of Neurobiology*, 42(2):248–257.

West, M. J. and Slomianka, L. (1998). Total number of neurons in the layers of the human entorhinal cortex. (and corrigendum). *Hippocampus*, 8:69–82 (and 426).

Wiskott, L., Rasch, M., and Kempermann, G. (2004). What is the functional role of adult neurogenesis in the hippocampus? Cognitive Sciences EPrint Archive (CogPrints) 4012, http://cogprints.org/4012/.