scientific reports

Check for updates

OPEN

Neural network based formation of cognitive maps of semantic spaces and the putative emergence of abstract concepts

Paul Stoewer[1,2], Achim Schilling[1,3], Andreas Maier[2] & Patrick Krauss[1,2,3,4]✉

How do we make sense of the input from our sensory organs, and put the perceived information into context of our past experiences? The hippocampal-entorhinal complex plays a major role in the organization of memory and thought. The formation of and navigation in cognitive maps of arbitrary mental spaces via place and grid cells can serve as a representation of memories and experiences and their relations to each other. The multi-scale successor representation is proposed to be the mathematical principle underlying place and grid cell computations. Here, we present a neural network, which learns a cognitive map of a semantic space based on 32 different animal species encoded as feature vectors. The neural network successfully learns the similarities between different animal species, and constructs a cognitive map of 'animal space' based on the principle of successor representations with an accuracy of around 30% which is near to the theoretical maximum regarding the fact that all animal species have more than one possible successor, i.e. nearest neighbor in feature space. Furthermore, a hierarchical structure, i.e. different scales of cognitive maps, can be modeled based on multi-scale successor representations. We find that, in fine-grained cognitive maps, the animal vectors are evenly distributed in feature space. In contrast, in coarse-grained maps, animal vectors are highly clustered according to their biological class, i.e. amphibians, mammals and insects. This could be a putative mechanism enabling the emergence of new, abstract semantic concepts. Finally, even completely new or incomplete input can be represented by interpolation of the representations from the cognitive map with remarkable high accuracy of up to 95%. We conclude that the successor representation can serve as a weighted pointer to past memories and experiences, and may therefore be a crucial building block to include prior knowledge, and to derive context knowledge from novel input. Thus, our model provides a new tool to complement contemporary deep learning approaches on the road towards artificial general intelligence.

The hippocampal-entorhinal complex supports spatial navigation and forms cognitive maps of the environment[1]. However, recent research suggests that formation of and navigation on cognitive maps are not limited to physical space, but extend to more abstract conceptual, visual or even social spaces[2–4]. A simplified processing framework for the complex can be described as following: highly processed information from our sensory organs are fed into the hippocampal complex where the perceived information is put into context, i.e. associated with past experiences[5]. Grid[6] and place[7] cells enable map like codes, and research suggests that they form cognitive maps[8,9], thereby contributing to process memories, emotions and navigation[10](cf. Fig. 1).

Furthermore, it is known that the hippocampus plays a crucial role for episodic and declarative memory[11,12]. However, whether memories are directly stored in the hippocampus, and how they are retrieved through the hippocampus, is depending on different theories still under discussion. Therefore, the exact role of the hippocampus in the domain of memory is still not fully understood[13]. According to the *multiple trace theory*[14], memories are not directly stored in the hippocampus. Instead, memory content is stored in the cerebral cortex, and the hippocampus forms representations of memory traces which can serve as pointers to retrieve memory content from the cerebral cortex.

[1]Cognitive Computational Neuroscience Group, University Erlangen-Nuremberg, Erlangen, Germany. [2]Pattern Recognition Lab, University Erlangen-Nuremberg, Erlangen, Germany. [3]Neuroscience Lab, University Hospital Erlangen, Erlangen, Germany. [4]Linguistics Lab, University Erlangen-Nuremberg, Erlangen, Germany. ✉email: patrick.krauss@fau.de

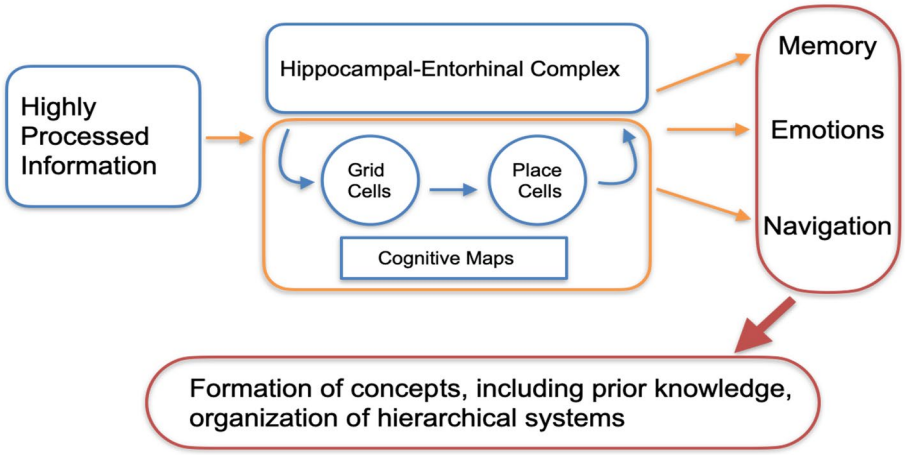

Figure 1. Simplified sketch model of the hippocampal-entorhinal complex. Highly processed information is fed into the system and becomes associated with existing memories and past experiences. Place and grid cells enable the formation of map-like codes, and finally cognitive maps. The hippocampal-entorhinal complex also supports navigation, emotions, the formation of concepts, inclusion of prior knowledge, and the organization of hierarchical representations.

Furthermore, memory can be represented at different scales along the hippocampal longitude axis, like e.g. varying spatial resolutions[15]. In the context of spatial navigation the different scales serve to navigate with different horizons[16]. In the context of abstract conceptual spaces, different scales might correspond to different degrees of abstraction[17]. In general, multi-scale cognitive maps enable flexible planning, generalization and detailed representation of information[18].

Various different computational models try to explain the versatility of the hippocampal-entorhinal complex. One of these candidate models successfully reproduces the firing patterns of place and grid cells in a large number of different experimental scenarios, indicating that the hippocampus works like a predictive map based on multi-scale successor representations (SR)[19–21].

In a previous study, we introduced a neural network based implementation of this frame-work, and demonstrated its applicability to several spatial navigation and non-spatial linguistic tasks[22]. Here, we further extended our model as shown in Fig. 1. In particular, we build a neural network which learns the SR for a non-spatial navigation task based on input feature vectors representing different animal species. To the best of our knowledge, our approach combines, for the first time, the memory trace theory with the cognitive map theory within a neural network framework.

Methods

Successor representation. The developed cognitive map is based on the principle of the successor representation (SR). As proposed by Stachenfeld and coworkers the SR can model the firing patterns of the place cells in the hippocampus[20]. The SR was originally designed to build a representation of all possible future rewards $V(s)$ that may be achieved from each state s within the state space over time[23]. The future reward matrix $V(s)$ can be calculated for every state in the environment, whereas the parameter t indicates the number of time steps in the future that are taken into account, and $R(s_t)$ is the reward for state s at time t. The discount factor $\gamma[0, 1]$ reduces the relevance of states s_t that are further in the future relative to the respective initial state s_0 (cf. Eq. 1).

$$V(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)|s_0 = s\right] \tag{1}$$

Here, $E[\,]$ denotes the expectation value.

The future reward matrix $V(s)$ can be re-factorized using the SR matrix M, which can be computed from the state transition probability matrix T of successive states (cf. Eq. 2). In case of supervised learning, the environments used for our model operate without specific rewards for each state. For the calculation of these SR we set $R(s_t) = 1$ for every state.

$$V(s) = \sum_{s'} M(s, s')R(s')M = \sum_{t=0}^{\infty} \gamma^t T^t \tag{2}$$

Animal data set. The construction of the cognitive map is based on a data set which quantifies seven different semantic features of 32 animal species (Table 1). The corresponding test data set is shown in Table 2.

The data matrix represents the memory matrix $M(m)$, which our cognitive map is based on. Therefore every animal represents a past memory, and reflects a state in our model. To use the matrix for our supervised learning

Name	Height (cm)	Weight (kg)	Number legs	Danger (subjective)	Reproduction (2 = Birth, 1= Eggs)	Fur (2 = No, 1 = Yes)	Lungs (2 = No, 1 = Yes)
Elephant	350	6000	4	60	2	2	1
Tiger	100	100	4	100	2	1	1
Lion	120	175	4	100	2	1	1
Dog	70	30	4	20	2	1	1
Rabbit	40	2	4	0	2	1	1
Bear	200	500	4	60	2	1	1
Cow	120	500	4	20	2	1	1
Deer	70	20	4	0	2	1	1
Cat	30	4	4	5	2	1	1
Beaver	60	25	4	5	2	1	1
Giraffe	500	1200	4	40	2	1	1
Ape	70	40	4	30	2	1	1
Horse	120	250	4	10	2	1	1
Camel	125	400	4	10	2	1	1
Goat	70	60	4	5	2	1	1
Sheep	60	20	4	5	2	1	1
Pig	60	200	4	5	2	2	1
Hamster	5	0.2	4	0	2	1	1
Dolphine	200	60	0	10	2	2	1
Raccoon	50	15	4	5	2	1	1
Red Pander	30	5	4	5	2	1	1
Ant	0.1	0.00001	6	1	1	2	2
Bee	1	0.0001	6	5	1	2	2
Cockroach	5	0.005	6	0	1	2	2
Goliathus	8	5	6	0	1	2	2
Giant weta	10	0.035	6	0	1	2	2
Heteropteryx	15	0.05	6	0	1	2	2
Cane toad	15	1	4	0	1	2	1
Fire Salamander	17	0.035	4	0	1	2	1
Frog	4	0.01	4	0	1	2	1
Olm	20	0.02	4	0	1	2	1
Tree Frog	4	0.005	4	0	1	2	1

Table 1. Training data set used to create the cognitive room. It consists of 32 different animal species, which belong to three different taxonomic classes: mammals, insects and amphibians. Each animal is characterized by seven semantic features: Height, weight, number of legs, its danger level, the reproduction system, if it has fur and if it has lungs.

Name	Height (cm)	Weight (kg)	Number legs	Danger (subjective)	Reproduction (2 = Birth, 1 = Eggs)	Fur (2 = No, 1 = Yes)	Lungs (2 = No, 1 = Yes)
Jaguar	70	70	4	90	2	1	1
Donkey	100	200	4	10	2	1	1
Wild boar	70	180	4	20	2	1	1
Melontha	2.5	0.001	6	0	1	2	2
Dragonfly	6	0.0003	6	0	1	2	2
Wasp	1.5	0.00008	6	20	1	2	2

Table 2. Test data used to evaluate the interpolation capabilities of the trained neural network. It consists of 6 different animal species, which belong to three different taxonomic classes: mammals, insects and amphibians. Again, each animal is characterized by seven semantic features: Height, weight, number of legs, its danger level, the reproduction system, if it has fur and if it has lungs.

approach, we need to sample successor labels for every state, which reflect the similarity between animal species. We choose to use the Euclidean distance to calculated the transition probabilities for our state space. Therefore animal species sharing similar semantic features have a higher state transition probability.

$$T(s, s') = \frac{1}{||m_s - m_{s'}||} \tag{3}$$

For the generation of the training and test data set, a random starting state is chosen and also a random probability ranging from 0 to 1 is sampled. The input feature vector for the chosen input state is altered by a random range of 0–15% to make the training more robust to novel inputs. Based on the sampled probability and the cumulative density function of the defined successor representation matrix, a valid successor state is randomly drawn as label. 10 % of the generated samples are not used for training, but are instead preserved as validation data set.

Note that, our data set represents no typical classification data set with a single ground truth label for each input pattern. Instead, our data set represents a simplified version of an animal taxonomy, where each animal is represented as a feature vector and the outputs to be learned by the neural network are the most similar animal species to the input animal species. These correspond to the nearest neighbors in the feature vector space. Since there are on average three nearest neighbors, the theoretical maximum of the accuracy is approximately 0.3 instead of 1.0.

Neural network architectures and training parameters. We set up a three-layered feed forward neural network. The network consists of 7 input neurons and has 7 neurons in the hidden layer. Both use a ReLU activation function. The output layer consists of 32 neurons with a softmax activation function (cf. Fig. 2). The networks learns the transition probabilities of the environment, i.e. in our case the memory space. Smaller number of neurons in the hidden layer did not influence the results in previous experiments[22]. We trained three networks for different discount factors of the successor representation, with $\gamma = (0.3, 0.7, 1.0)$ and $t = 10$. Note that, larger discount factors correspond to a larger time horizon, i.e. taking into account more future steps. The networks were trained for 500 epochs, with a batch size of 50, 50,000 training samples, using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy as loss function.

Transition probability and successor representation matrix. After the training process, the networks can predict all probabilities for the successor states for any given input feature vector. Concatenating the predictions of all known animal states leads to the successor representation matrix of the cognitive room. The output of the network is a vector shaped like a row of the respective environment's SR matrix and can therefore directly be used to fill the SR matrix, respectively.

Interpolating unknown features. We propose that the successor representation can be used as a pointer to stored memories. In our case we have the saved memories of 32 animal species in the memory trace matrix which we use for training the network. If incomplete information is fed into the network (unknown values set to -1 in the input feature vector), it still outputs predictions for the possible transition probabilities.

$$m_{interpolated} = SR_{prediction} * M(m_s) \tag{4}$$

Thus, we can use the prediction from the network, and perform a matrix multiplication with our known memory matrix in order to derive an interpolated feature vector for the incomplete or unknown input (cf. Fig. 4).

Multi-dimensional scaling. A frequently used method to generate low-dimensional embeddings of high-dimensional data is t-distributed stochastic neighbor embedding (t-SNE)[24]. However, in t-SNE the resulting low-dimensional projections can be highly dependent on the detailed parameter settings[25], sensitive to noise, and may not preserve, but rather often scramble the global structure in data[26,27]. In contrast to that, multi-Dimensional-Scaling (MDS)[28–31] is an efficient embedding technique to visualize high-dimensional point clouds by projecting them onto a 2-dimensional plane. Furthermore, MDS has the decisive advantage that it is parameter-free and all mutual distances of the points are preserved, thereby conserving both the global and local structure of the underlying data.

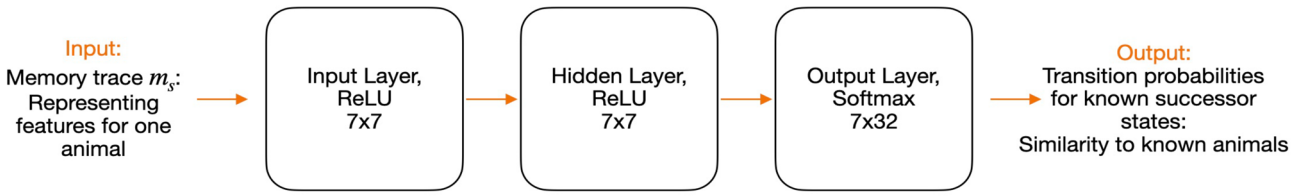

Figure 2. Architecture of the trained neural network. The network receives a memory trace of animal features as input. The size of the input and hidden layer is equal to the number of features in the input. The output layer is a softmax layer with 32 neurons, matching the number of memory traces in the training memory matrix. The output of the network is the probability of the similarity of the input to the entries of the memory matrix used during training.

When interpreting patterns as points in high-dimensional space and dissimilarities between patterns as distances between corresponding points, MDS is an elegant method to visualize high-dimensional data. By color-coding each projected data point of a data set according to its label, the representation of the data can be visualized as a set of point clusters. For instance, MDS has already been applied to visualize for instance word class distributions of different linguistic corpora[32], hidden layer representations (embeddings) of artificial neural networks[33,34], structure and dynamics of recurrent neural networks[35–38], or brain activity patterns assessed during e.g. pure tone or speech perception[32,39], or even during sleep[40,41]. In all these cases the apparent compactness and mutual overlap of the point clusters permits a qualitative assessment of how well the different classes separate.

Code implementation. The models were coded in Python 3.10. The neural networks were design using the Keras[42] library with TensorFlow[43]. Mathematical operations were performed with numpy[44] and scikit-learn[45] libraries. Visualizations were realised with matplotlib[46].

Results

Learning structures by observing states and their successors. The models were trained to learn the underlying structure of the data set. In particular, we trained three different neural networks using different discount factors ($\gamma = (0.3, 0.7, 1.0), t = 10$). The resulting successor representation matrices for each parameter setting are very similar to the ground truth (Fig. 3), and the corresponding root-mean-squared errors (RMSE) are extremely low: 0.02034 for $\gamma = 0.3$ (Fig. 3A), 0.01496 for $\gamma/0.7$ (Fig. 3B), and 0.00854 for $\gamma = 1.0$ (Fig. 3C).

The accuracy for the model with the discount factor $\gamma = 0.3$ increased quickly during the first 100 epochs, and then slowly continued to increase until the end of training at epoch 500 where the highest accuracy of 60% was achieved (Fig. 4A).

In contrast, the training procedure quickly reached a saturation of the accuracy for the two models with discount factors $\gamma = 0.7$ and $\gamma = 1.0$ after around 200 epochs, with maximum training and validation accuracies of approximately 30% or 35% respectively (Fig. 4B,C).

Scaling of cognitive maps depends on discount factor of successor representation. The discount factor of the SR is proposed to enable scaling of the cognitive maps, and thus to represent hierarchical structures, similar to the different mesh sizes of grid cells along the longitudinal axis of the hippocampus and the entorhinal cortex[20]. Actually, memory representations, such as the internal representation of space, systematically vary in scale along the hippocampal long axis[15]. This scaling has been suggested to be used for targeted navigation with different horizons[16] or even for encoding information from smaller episodes or single objects to more complex concepts[17].

In order to visualize the learned SR underlying the cognitive maps, we calculate MDS pojections from the SR matrices (Fig. 5). Furthermore, as an estimate for the map scaling, we calculate the general discrimination value (GDV, cf. "Methods") for each map.

We find that the resulting scaling of the cognitive maps depends on the discount factor of the underlying SR matrix, and that the GDV correlates with the discount factor. A small discount factor of $\gamma = 0.3$ results in a fine-grained and detailed cognitive map where each object is clearly separated from the others, and similar objects, i.e. animal species, are closer together (Fig. 5A). With a GDV of -0.322, the clustering is relatively low compared to the other maps. This cognitive map resembles so called self-organizing maps introduced by Kohonen[47], and might correspond to word fields proposed in linguistics[48].

A discount factor $\gamma = 0.7$ results in an intermediate scale cognitive map with a GDV of -0.355 (Fig. 5B).

Finally, a discount factor of $\gamma = 1.0$ results in the most course-grained cognitive map. Here, individual animal species are no longer clearly separated from each other, but are forming instead representational clusters that correspond to taxonomic animal classes, i.e. mammals, insects and amphibians (Fig. 5c). Consequently, these map has the lowest GDV of -0.403, indicating the best clustering. This type of representation generalizing from individual objects might correspond to the emergence of cognitive categories, as suggested e.g. in prototype semantics[49].

Feature inference for incomplete feature vectors. The neural network which learned the structure of the input data successfully can now be used to interact with unseen data. The prediction of the trained neural network can be used as weighted pointer to the memorized objects (animal species) in the training data set. The vector of a previous unseen animal, the *jaguar*, is fed into the network for prediction (Fig. 6). Three features (danger, fur, lungs) are missing, i.e. are set to -1. The binary features are predicted well independent from the discount factor. The 'danger' feature is inferred best for the smallest discount factor $\gamma = 0.3$. Note that, also the not missing parameters are changed by the prediction. In general, larger discount factors better infer more general features, whereas smaller discount factors better infer more specific features.

We further evaluated the model with our interpolation test data set (cf. Fig. 2). We trained ten models with the parameters $\gamma = 1.0$ and $t = 10$. In Fig. 7 the distances of the predictions of different features in comparison to the ground truth is summarized. The percentage is based on the maximum distance of the according feature. The evaluation is plotted for the feature vectors, with up to 6 missing entries for every prediction. The distance of the prediction to ground truth with no missing entries is in general low ranging from around 5–25% (corresponding to 95–75% accuracy), indicating high similarity. However, dissimilarity increases to 40% in case of 6 missing features. The distance is however different for each feature. While the semantic feature 'number of legs' is predicted well, the height of the animal is predicted with less accuracy. Furthermore, the variance differs for different models. Especially the badly predicted predicted features like 'height' and 'weight' the variance is quite large.

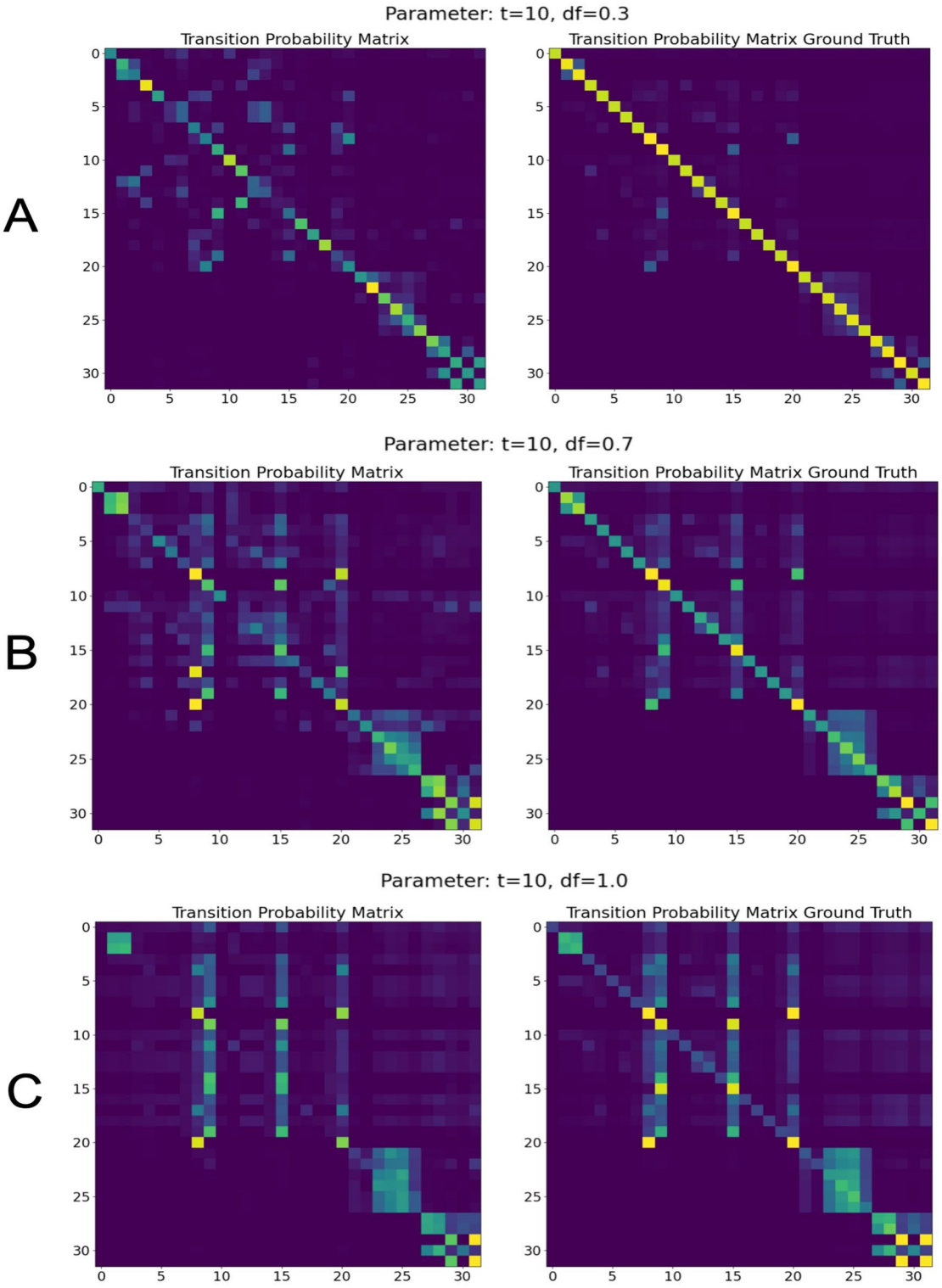

Figure 3. Learned successor representation (SR) matrices and corresponding ground truths. Learned SR matrices (left column) are very similar to their corresponding ground truth SR matrices (right column). (**A**) For a discount factor of $\gamma = 0.3$, the root-mean-squared error (RMSE) between learned and ground truth SR matrix is 0.02034. (**B**) For $\gamma = 0.7$, the RSME is 0.01496. (**C**) For $\gamma = 1.0$, the RMSE is 0.00854. Note that, x-axes denote index of starting state, y-axes denote index of successor state.

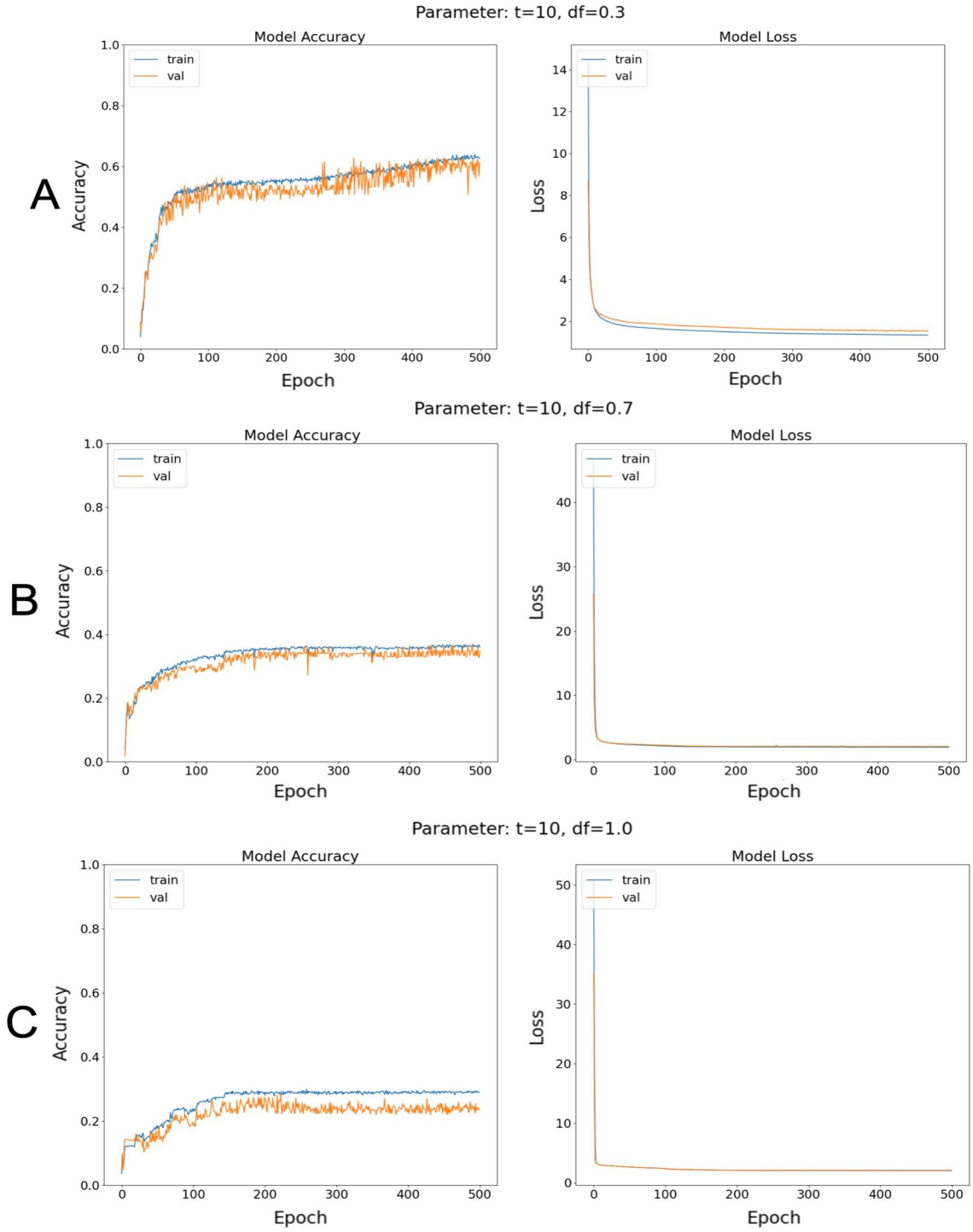

Figure 4. Accuracies and loss for different models. Training accuracies (blue) and validation accuracies (orange) during training are shown in the left column. The corresponding loss is shown in the right column. (**A**) For a discount factor of $\gamma = 0.3$, the highest accuracy of 60% was achieved. (**B**) For $\gamma = 0.7$, the accuracy saturates after 200 epochs at 30%. (**C**) For $\gamma = 1.0$, the accuracy saturates after 200 epochs at 35%.

Discussion

In this study we have demonstrated that arbitrary feature spaces can be learned efficiently on the basis of successor representations with neural networks. In particular, the networks learn a representation of the semantic feature space of animal species as a cognitive map. The network achieves an accuracy of around 30% which is near to the theoretical maximum regarding the fact that all animal species have more than one possible successor, i.e. nearest neighbor in feature space. Our approach therefore combines the concepts of feature space and

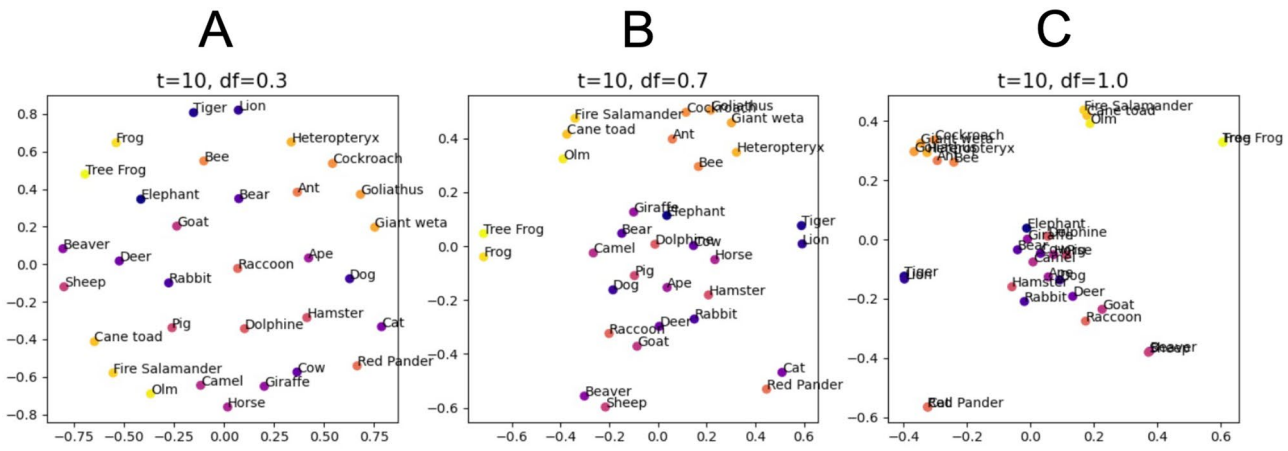

Figure 5. Different scalings of cognitive maps. Shown are MDS projections of SR matrices with different discount factors γ. (**A**) For a low discount factor $\gamma = 0.3$, the resulting map is most fine-grained and detailed with little clustering (GDV: −0.322). (**B**) A medium discount factor of $\gamma = 0.7$ results in an intermediate scale with more clustering (GDV: −0.355) compared to (**A**). (**C**) The largest discount factor $\gamma = 1.0$ results in the most coarse-grained map. Here, individual animal species are no longer distinguishable, but instead form separated, dense clusters possibly enabling the emergence of more abstract concepts in subsequent processing stages, i.e. the taxonomic animal classes mammals (blue, purple), insects (orange), and amphibians (yellow) (GDV: −0.403). Note that, a GDV of −1.0 indicates perfect clustering, whereas a GDV of 0.0 indicates no clustering at all. Note that, the axes in the MDS plots are in arbitrary units and have no particular meaning other than illustrating the relative positions, i.e. similarities, of all objects.

Height	Weight	#Legs	Danger	Rep.	Fur	Lungs
70	70	4	90(-1)	2	1(-1)	1(-1)

γ	Height	Weight	#Legs	Danger	Rep.	Fur	Lungs
0.3	109.35	161.53	4.001	80.02	1.99	1.01	1.002
0.7	62.01	48.14	4.006	4.54	1.94	1.06	1.01
1.0	56.35	21.51	4.006	4.97	1.97	1.002	1.003

Figure 6. Interpolation of the test data set feature vector 'Jaguar'. Three semantic features (dangerous, having a fur and having lungs) are missing, i.e. are replaced by the value −1. The three networks trained with different discount factors $\gamma = (0.3, 0.7, 1.0)$ infer the missing features. Binary semantic features are inferred well in all cases. The 'dangerous' feature is badly predicted for large discount factors $\gamma = (0.7, 1.0)$. In contrast, in case of the lower discount factor $\gamma = 0.3$, it is predicted well.

state space. The emerging representations therefore resemble the proposed general and abstract cognitive maps described by Bellmund et al.[50].

Our model extends our past work, where we reproduced place cell fire patterns in a spatial navigation and a non-spatial linguistic task based on the successor representation[22]. The innovation of our here presented approach is, that we can use the successor representation with arbitrary new input as a weighted pointer to already stored input data, i.e. memories, thereby combining the two hallmarks of hippocampal processing: declarative memory

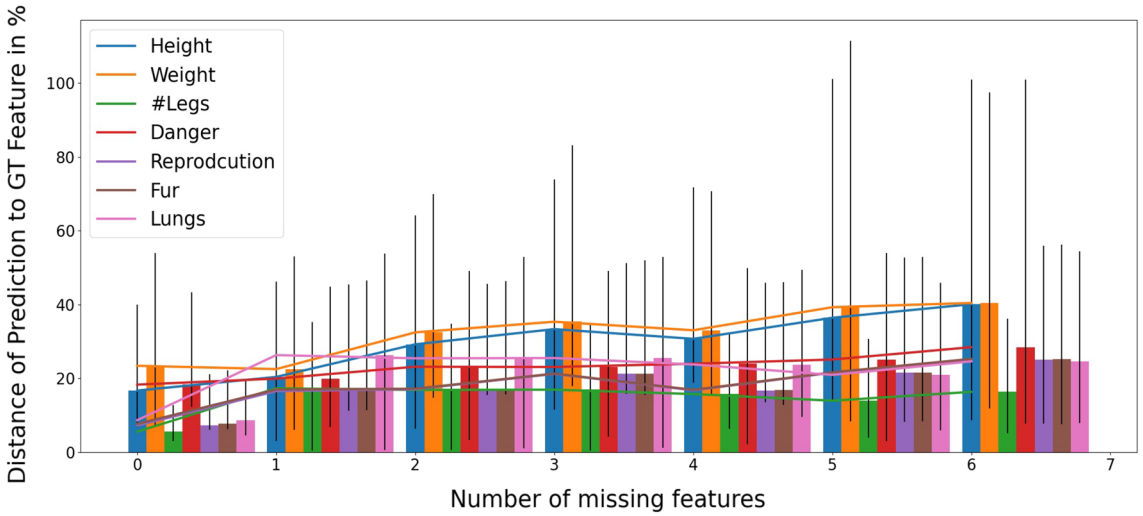

Figure 7. Dissimilarities between interpolated features and ground truth. 10 networks with $\gamma = 1.0$ have been trained. Dissimilarity is low in case of a no or a single missing feature, and increases with number of missing features up to 40% for six missing features. In general, binary semantic features are inferred with better accuracy than non-binary semantic features. The variance of the different networks for the features 'height' and 'weight' are highest.

and navigation. The successor representation might therefore be a tool which can be used to navigate through arbitrary cognitive maps, and find similarities in novel inputs as well as past memories.

We note that, the concept of 'fuzzy cognitive maps'[51] can be seen as a simplified version of our cognitive maps based on successor representations. Whereas fuzzy cognitive maps are signed directed graphs, our cognitive maps are weighted directed graphs, yet without a sign, as the weights represent probabilities. The potential benefit of a combination of both concepts, i.e. introducing negative weights into cognitive maps, could be addressed in a future follow-up study.

We note that, obviously, our neural network based model is still a simplification of the actual neural and mental processes underlying memory and cognitive map formation. For instance, whereas we applied the Euclidean distance as a proxy for the dis-similarity between different objects, existing behavioral data suggest that similarity in psychological space is best represented by a weighted, nonlinear (exponential) decrease as the number of stimulus dimensions increases. This principle was proposed by Shepard[52], and has been demonstrated in several studies on human generalization behavior[53]. Thus, including an exponentially decaying distance function in our model would make it more realistic. However, we do not expect that this would change our overall results, including the successor representations and the cognitive maps that emerge at different scales. Nonetheless, this issue should be considered in future modeling studies of cognitive maps.

We found that, the discount factor γ of the successor representation can be used to model cognitive maps with different scales, which range in our example from separated dense clusters of taxonomic animal classes to individual animal species. The varying grid cell scaling along the long axis of the entorhinal cortex is known to be associated with hierarchical memory content[15]. The discount factor can be used to model this hierarchical structure. In our experiment the hierarchical scale could be used to interpolate novel feature data in different ways. For example, if we want to retrieve general information, a large discount factor resulting in dense clusters, to derive averaged information about the whole cluster, can be used. In contrast, for more detailed information regarding a specific state of the cognitive map, a smaller discount factor is useful.

This ability of our model to represent individual exemplars at different scales, and in particular at a coarse-grained level with high clustering, might enable compression and reduction of complexity for subsequent processing stages. Based on experience, humans learn which features of a given stimuli are more relevant to recognize and identify a certain category membership. Thus, these features are likely to have a disproportionate influence on any conceptual representation. Thus, instead of memorizing individual exemplars, an extension of our model might also account for abstraction, generalization and concept emergence. Actually, 'Generalized Invariance Structure Theory' and 'Generalized Representational Information Theory'[54] suggest that concepts are formed via the detection of relational information between category exemplars encountered in the environment[55–60]. We note that, the presented model does not account for the mentioned features. However, we will address these extensions in a future follow-up study.

Since our approach works with a direct feature vector as input, it still requires highly pre-processed data. A future outlook for this model could be to include a deep neural network for feature extraction as pre-processing. For instance, image analysis is already a well established field for deep neural networks. Our model could be used to replace the last output layer of such networks, which usually performs a classification task, and use the feature space embeddings[61] as input for a subsequent cognitive map. This extended model could enhance learning from simple classification to understanding which features are present in which image. This could potential lead to more context awareness in neural networks.

Analogously, one could also use speech[62] or word vectors[63], or even sentence embeddings [?] as input for our model. By that, our neural network based cognitive maps could serve as a putative extension of contemporary large language models like, e.g. ChatGPT[64,65], or intelligent speech interfaces[66].

As recently suggested, the neuroscience of spatial navigation might be of particular importance for artificial intelligence research[67]. A neural network implementation of hippocampal successor representations, especially, promises advances in both fields. Following the research agenda of Cognitive Computational Neuroscience proposed by Kriegeskorte et al.[68], neuroscience and cognitive science benefit from such models by gaining deeper understanding of brain computations[34,69,70]. Conversely, for artificial intelligence and machine learning, neural network-based multi-scale successor representations to learn and process structural knowledge as an example of neuroscience-inspired artificial intelligence[71–74], might be a further step to overcome the limitations of contemporary deep learning[73–78] and towards human-level artificial general intelligence.

Data availability

The datasets used and/or analysed during the current study are available on github: https://github.com/Pa-Sto/CognitiveRoom.

Received: 28 October 2022; Accepted: 21 February 2023
Published online: 04 March 2023

References

1. O'Keefe, J., & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**(1), 171–175 (1971).
2. Epstein, R. A, Patai, E. Z., Julian, J. B., & Spiers, H. J. The cognitive map in humans: spatial navigation and beyond. *Nat. Neurosci.* **20**(11), 1504–1513 (2017).
3. Park, S. A, Miller, D. S, Boorman, E. D. Inferences on a multidimensional social hierarchy use a grid-like code. *bioRxiv*, 2020–05 (2021).
4. Killian, N. J., & Elizabeth A. B. Grid cells map the visual world. *Nat. Neurosci.* **21**(2) (2018).
5. Opitz, B. Memory function and the hippocampus. *Front. Neurol. Neurosci.* **34**, 51–59 (2014).
6. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**(7052), 801–806 (2005).
7. O'Keefe, J., & Dostrovsky, J. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* (1971).
8. O'keefe, J., & Nadel, L. *The hippocampus as a cognitive map*. Oxford university press, (1978).
9. Moser, E. I., Moser, M.-B. & McNaughton, B. L. Spatial representation in the hippocampal formation: A history. *Nat. Neurosci.* **20**(11), 1448–1464 (2017).
10. Kandel, E. R. editor. *Principles of neural science*. McGraw-Hill, New York, 5th ed edition (2013).
11. Tulving, E. & Markowitsch, H. J. Episodic and declarative memory: Role of the hippocampus. *Hippocampus* **8**(3), 198–204 (1998).
12. Reddy, L. *et al.* Human hippocampal neurons track moments in a sequence of events. *J. Neurosci.* **41**(31), 6714–6725 (2021).
13. Kryukov, V. I. The role of the hippocampus in long-term memory: Is it memory store or comparator?. *J. Integr. Neurosci.* **07**, 117–184 (2008).
14. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
15. Collin, S. H. P., Milivojevic, B. & Doeller, C. F. Memory hierarchies map onto the hippocampal long axis in humans. *Nat. Neurosci.* **18**(11), 1562–1564 (2015).
16. Brunec, I. K., & Momennejad, I. Predictive representations in hippocampal and prefrontal hierarchies. *bioRxiv* 786434 (2019).
17. Milivojevic, B. & Doeller, C. F. Mnemonic networks in the hippocampal formation: From spatial maps to temporal and conceptual codes. *J. Exp. Psychol. Gen.* **142**(4), 1231 (2013).
18. Momennejad, I. Learning structures: Predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* **32**, 155–166 (2020).
19. Stachenfeld, K. L., Botvinick, M. & Gershman, S. J. Design principles of the hippocampal cognitive map. *Adv. Neural. Inf. Process. Syst.* **27**, 2528–2536 (2014).
20. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**(11), 1643 (2017).
21. McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. Flexible modulation of sequence generation in the entorhinal-hippocampal system. *Nat. Neurosci.* **24**(6), 851–862 (2021).
22. Stoewer, P., Schlieker, C., Schilling, A., Metzner, C., Maier, A., & Krauss, P. Neural network based successor representations to form cognitive maps of space and language. *Sci. Rep.* **12**, 11233 (2022).
23. Dayan, P. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Comput.* **5**(4), 613–624 (1993).
24. Maaten, L. V., & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**(11) (2008).
25. Wattenberg, M., Viégas, F. & Johnson, I. How to use t-sne effectively. *Distill* **1**(10), e2 (2016).
26. Vallejos, C. A. Exploring a world of a thousand dimensions. *Nat. Biotechnol.* **37**(12), 1423–1424 (2019).
27. Moon, K. R, van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B, Chen, W.S, Yim, K. E., Antonia van den, H., Matthew, J., Coifman, R. R., *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**(12), 1482–1492 (2019)
28. Torgerson, W. S. Multidimensional scaling: I. theory and method. *Psychometrika* **17**(4), 401–419 (1952).
29. Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**(2), 115–129 (1964).
30. Kruskal, J. B. *Multidimensional scaling*. Number 11. Sage (1978).
31. Cox, M. A. A., Cox, T. F. Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347. Springer (2008).
32. Schilling, A. *et al.* Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods. *Lang. Cogn. Neurosci.* **36**(2), 167–186 (2021).
33. Schilling, A., Maier, A., Gerum, R., Metzner, C. & Krauss, P. Quantifying the separability of data classes in neural networks. *Neural Netw.* **139**, 278–293 (2021).
34. Krauss, P. *et al.* Analysis and visualization of sleep stages based on deep neural networks. *Neurobiol. Sleep Circ. Rhythms* **10**, 100064 (2021).
35. Krauss, P., Zankl, A., Schilling, A., Schulze, H. & Metzner, C. Analysis of structure and dynamics in three-neuron motifs. *Front. Comput. Neurosci.* **13**, 5 (2019).
36. Krauss, P., Prebeck, K., Schilling, A., & Metzner, C. Recurrence resonance' in three-neuron motifs. *Front. Comput. Neurosci.* 64 (2019).

37. Krauss, P. *et al.* Weight statistics controls dynamics in recurrent neural networks. *PLoS ONE* **14**(4), e0214541 (2019).
38. Metzner, C., Krauss, P. Dynamics and information import in recurrent neural networks. *Front. Comput. Neurosci.* 16 (2022).
39. Krauss, P. *et al.* A statistical method for analyzing and comparing spatiotemporal cortical activation patterns. *Sci. Rep.* **8**(1), 1–9 (2018).
40. Krauss, P. *et al.* Analysis of multichannel eeg patterns during human sleep: a novel approach. *Front. Hum. Neurosci.* **12**, 121 (2018).
41. Traxdorf, M., Krauss, P., Schilling, A., Schulze, H. & Tziridis, K. Microstructure of cortical activity during sleep reflects respiratory events and state of daytime vigilance. *Somnologie* **23**(2), 72–79 (2019).
42. François Chollet *et al.* Keras, (2015).
43. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
44. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T.E. Array programming with NumPy. *Nature* **585**(7825), :357–362 (2020).
45. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
46. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**(3), 90–95 (2007).
47. Kohonen, T. The self-organizing map. *Proc. IEEE* **78**(9), 1464–1480 (1990).
48. Aitchison, J. *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons, (2012).
49. Cruse, D. A. Prototype theory and lexical semantics. In *Meanings and Prototypes (RLE Linguistics B: Grammar)*, pp. 392–412. Routledge (2014).
50. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I, Doeller, C. F. Navigating cognition: Spatial codes for human thinking. *Science* **362**(6415) (2018).
51. Kosko, B. Fuzzy cognitive maps. *Int. J. Man Mach. Stud.* **24**(1), 65–75 (1986).
52. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**(4820), 1317–1323 (1987).
53. Nosofsky, R. M. Similarity scaling and cognitive process models. *Annu. Rev. Psychol.* **43**(1), 25–53 (1992).
54. Vigo, R. & Doan, C. A. The structure of choice. *Cogn. Syst. Res.* **36**, 1–14 (2015).
55. Vigo, R., Barcus, M., Zhang, Y., & Doan, C. On the learnability of auditory concepts. *J. Acoust. Soc. Am.* **134**(5), 4064–4064 (2013).
56. Doan, C. A., & Vigo, R. Constructing and deconstructing concepts. *Exp. Psychol.* (2016).
57. Vigo, R., Doan, K.-M.C., Doan, C. A. & Pinegar, S. On the learning difficulty of visual and auditory modal concepts: Evidence for a single processing system. *Cogn. Process.* **19**, 1–16 (2018).
58. Vigo, R., Doan, C. A. & Zeigler, D. E. Context, structure, and informativeness judgments: An extensive empirical investigation. *Mem. Cognit.* **48**, 1089–1111 (2020).
59. Doan, C. A, Vigo, R. A comparative investigation of integral-and separable-dimension stimulus-sorting behavior. *Psychol. Res.* 1–27 (2022).
60. Vigo, R., Doan, C. A., Zhao, L. Classification of three-dimensional integral stimuli: Accounting for a replication and extension of nosofsky and palmeri (1996) with a dual discrimination invariance model. *J. Exp. Psychol. Learn. Mem. Cognit.* (2022).
61. Chen, H., Perozzi, B., Al-Rfou, R., & Skiena, S. A tutorial on network embeddings. arXiv preprint arXiv:1808.02590, (2018).
62. Schneider, S., Baevski, A., Collobert, R., & Auli, M. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862, (2019).
63. Goldberg, Y., & Levy, O. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722, (2014).
64. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.* Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155, (2022).
65. OpenAI. Chatgpt. https://openai.com/blog/chatgpt/ (2022).
66. de Barcelos Silva, A., Miguel Gomes, M., da Costa, C.A., da Rosa Righi, R., Victoria Barbosa, J. L., Pessin, G., Doncker, G.D., & Federizzi, G. Intelligent personal assistants: A systematic literature review. *Expert Syst. Appl.* **147**, 113193 (2020).
67. Bermudez-Contreras, E., Clark, B. J. & Wilber, A. The neuroscience of spatial navigation and the relationship to artificial intelligence. *Front. Comput. Neurosci.* **14**, 63 (2020).
68. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* **21**(9), 1148–1160 (2018).
69. Schilling, A., Gerum, R., Zankl, A., Schulze, H., Metzner, C., & Krauss, P. Intrinsic noise improves speech recognition in a computational model of the auditory pathway. *bioRxiv*, (2020).
70. Krauss, P., Tziridis, K., Schilling, A. & Schulze, H. Cross-modal stochastic resonance as a universal principle to enhance sensory processing. *Front. Neurosci.* **12**, 578 (2018).
71. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**(2), 245–258 (2017).
72. Krauss, P., & Maier, A. Will we ever have conscious machines? *Front. Comput. Neurosci.* 116 (2020).
73. Yang, Z., Schilling, A., Maier, A., & Krauss, P. Neural networks with fixed binary random projections improve accuracy in classifying noisy data. In *Bildverarbeitung für die Medizin 2021*, 211–216 (Springer, 2021).
74. Maier, A., Köstler, H., Heisig, M., Krauss, P., & Hee, S. Known operator learning and hybrid machine learning in medical imaging—a review of the past, the present, and the future. *Prog. Biomed. Engi.* (2022).
75. Krauss, P., Metzner, C., Lange, J., Lang, N. & Fabry, B. Parameter-free binarization and skeletonization of fiber networks from confocal image stacks. *PLoS ONE* **7**(5), e36575 (2012).
76. Marcus, G. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631, (2018).
77. Gerum, R. C. & Schilling, A. Integration of leaky-integrate-and-fire neurons in standard machine learning architectures to generate hybrid networks: A surrogate gradient approach. *Neural. Comput.* **33**(10), 2827–2852 (2021).
78. Maier, A. K. *et al.* Learning with known operators reduces maximum error bounds. *Nat. Mach. Intell.* **1**(8), 373–380 (2019).

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): grants KR 5148/2-1 (project number 436456810), KR 5148/3-1 (project number 510395418) and GRK 2839 (project number 468527017) to PK, and grant SCHI 1482/3-1 (project number 451810794) to AS.

Author contributions

P.S. performed computer simulations and prepared all figures. P.S., A.M. and P.K. designed the study. P.K. and A.M. supervised the study. P.S., A.S., A.M. and P.K. discussed the results and wrote the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. We acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme "Open Access Publication Funding".

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to P.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.