Modern Hopfield Networks and Attention for Immune Repertoire Classification

Michael Widrich[1], Bernhard Schäfl[1], Milena Pavlović[2,3], Hubert Ramsauer[1], Lukas Gruber[1], Markus Holzleitner[1], Johannes Brandstetter[1], Geir Kjetil Sandve[3], Victor Greiff[2], Sepp Hochreiter[1,4], Günter Klambauer[1]

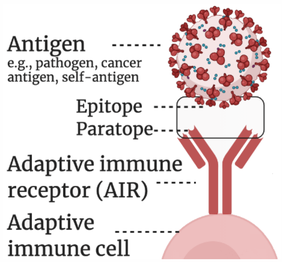
1) ELLIS Unit Linz & LIT AI Lab, Institute for Machine Learning, Johannes Kepler University Linz, Austria
2) Department of Immunology, University of Oslo, Norway
3) Department of Informatics, University of Oslo, Norway
4) Institute of Advanced Research in Artificial Intelligence (IARAI)

Institute for Machine Learning

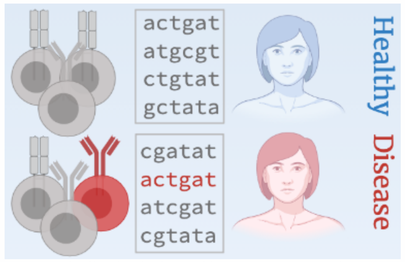
A central mechanism in machine learning is to identify, store, and recognize patterns. How to learn, access, and retrieve such patterns is crucial in Hopfield networks and the more recent transformer architectures. We show that the attention mechanism of transformer architectures is actually the update rule of modern Hopfield networks that can store exponentially many patterns. We exploit this high storage capacity of modern Hopfield networks to solve a challenging multiple instance learning (MIL) problem in computational biology: immune repertoire classification. In immune repertoire classification, a vast number of immune receptors are used to predict the immune status of an individual. This constitutes a MIL problem with an unprecedentedly massive number of instances, two orders of magnitude larger than currently considered problems, and with an extremely low witness rate. Accurate and interpretable machine learning methods solving this problem could pave the way towards new vaccines and therapies, which is currently a very relevant research topic intensified by the COVID-19 crisis. In this work, we present our novel method DeepRC that integrates transformer-like attention, or equivalently modern Hopfield networks, into deep learning architectures for massive MIL such as immune repertoire classification. We demonstrate that DeepRC outperforms all other methods with respect to predictive performance on large-scale experiments including simulated and real-world virus infection data and enables the extraction of sequence motifs that are connected to a given disease class.

Motivation & Biological Background

In this work, we tackle the problem of immune repertoire classification using modern Hopfield networks for extremely massive multiple instance learning. The immune repertoire of an individual consists of about 10^7–10^8 unique immune receptors, with little overlap between individuals [1]. Immune receptors enable the immune system to combat pathogens, such as viruses. This is achieved via binding sites in the receptors that bind to the pathogens. Usually, the presence of only a small fraction of particular receptors determines whether an individual is immune w.r.t. a particular disease [2]. Knowledge about which binding site makes an individual immune against a pathogen would allow for the development of novel treatments [3].



Immune receptor and antigen [4]



Symbolic depiction of immune repertoires; red: receptor sequence with binding site [4]

Immune Repertoire Classification by Massive Multiple Instance Learning

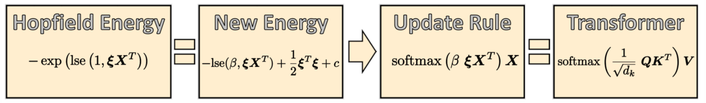
Immune receptors can be extracted from the body, e.g. via blood samples, and represented as amino acid sequences. This results in a very large bag of amino acid sequences. Predicting the immune status of an individual given this bag of sequence instances is essentially a text-book example of a *multiple instance learning* (MIL) problem. Characteristic for this extremely massive MIL problem are (i) large numbers of instances ($\approx 300k$ sequences per bag in our experiments), (ii) labeled bags without instance labels, and (iii) extremely low witness rates, which is the rate of discriminating instances per bag, down to 0.01%, where only subsequences determine the bag label. Solving this problem with an interpretable ML model would allow to infer which (sub)sequences the model used for prediction and possibly identify the biding sites of interest. However, in MIL problems considered by machine learning methods up to now, the number of instances per bag is in the range of hundreds or few thousands. At the same time, the witness rate is already considered low at $1\% - 5\%$ [5].

We employ novel modern Hopfield networks to tackle this challenging massive MIL task, as they allow to store and retrieve exponentially (in the dimension of the association space) many patterns.

Modern Hopfield Networks Have Exponential Storage Capacity

We propose the use of a modern Hopfield network with current state $\boldsymbol{\xi}$ and energy function $\mathrm{E} = -\beta^{-1} \log \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) + \beta^{-1} \log N + \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} + \frac{1}{2} M^2$ as continuous generalizations of binary modern Hopfield-networks [6,7,8,9], where $\boldsymbol{\xi}$ is the *state pattern* and the $\boldsymbol{x}_i$ are the *stored patterns*.

For energy E and state $\boldsymbol{\xi}$, the update rule $\boldsymbol{\xi}^{\mathrm{new}} = f(\boldsymbol{\xi}; \boldsymbol{X}, \beta) = \boldsymbol{X}\boldsymbol{p} = \boldsymbol{X}\ \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})$ is proven to converge globally to stationary points of the energy E, which are local minima or saddle points, using the *concave-convex procedure*. We show that the update rule of modern Hopfield networks is equivalent to the self-attention mechanism of transformers [9,10].



We show that our modern Hopfield networks exhibit exponential storage capacity. For randomly chosen patterns, the number of patterns that can be stored is exponential in the dimension d of the space of the patterns ($\boldsymbol{x}_i \in \mathbb{R}^d$).

Theorem 1 *We assume a failure probability $0 < p \leqslant 1$ and randomly chosen patterns on the sphere with radius $M = K\sqrt{d-1}$. We define $a := \frac{2}{d-1}\left(1 + \ln(2\,\beta\,K^2\,p\,(d-1))\right)$, $b := \frac{2\,K^2\,\beta}{5}$, and $c = \frac{b}{W_0(\exp(a + \ln(b)))}$, where W_0 is the upper branch of the Lambert W function, and ensure $c \geq \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}$. Then with probability $1-p$, the number of random patterns that can be stored is $N \geq \sqrt{p}\, c^{\frac{d-1}{4}}$.*

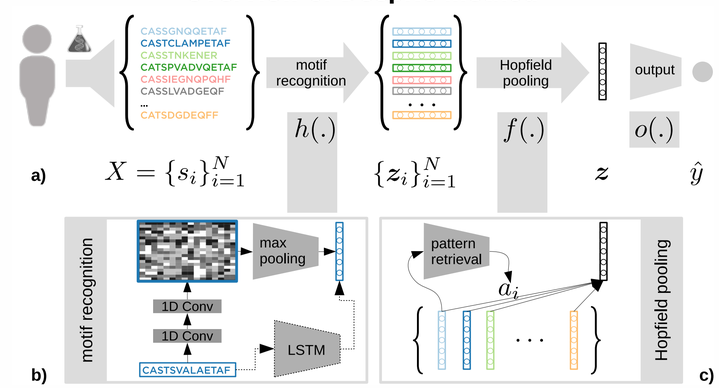
Examples are $c \geq 3.1546$ for $\beta = 1$, $K = 3$, $d = 20$, and $p = 0.001$ ($a + \ln(b) > 1.27$) and $c \geq 1.3718$ for $\beta = 1$ $K = 1$, $d = 75$, and $p = 0.001$ ($a + \ln(b) < -0.94$).

Deep Repertoire Classification

We propose a novel method **Deep R**epertoire **C**lassification (**DeepRC**) for immune repertoire classification with attention-based deep massive multiple instance learning. We consider immune repertoires as input objects, represented as bags of receptor sequence instances.

DeepRC consists of three subnetworks:
(i) An embedding network h to create a fixed-length feature vector $\boldsymbol{z}_i$ from each variable-length sequence instance $\boldsymbol{s}_i$. (ii) The attention network f, which pools the sequence instances into one repertoire representation $\boldsymbol{z}$. We apply a modified modern Hopfield network as pooling mechanism to address the large number of instances. It consists of a fixed and learned state pattern or query vector $\boldsymbol{\xi}$ and stored patterns or keys $\boldsymbol{K} = \mathrm{SNN}(\boldsymbol{Z})$ that are mapped the to the associative space via a self-normalizing neural network (SNN): $\boldsymbol{z} = \mathrm{att}(\boldsymbol{\xi}^T, \boldsymbol{K}, \boldsymbol{Z}; \frac{1}{\sqrt{d_k}}) = \mathrm{softmax}\left(\frac{\boldsymbol{\xi}^T \boldsymbol{K}^T}{\sqrt{d_k}}\right) \boldsymbol{Z}$, where $\boldsymbol{Z} \in \mathbb{R}^{N \times d_v}$ are the sequence-representations stacked row-wise. (iii) The repertoire-representation $\boldsymbol{z}$ is then processed by an output network o.

Overview of DeepRC method



To make the computational effort for this number of instances feasible, we compute the prediction and weight updates using only the 10% of instances with the highest attention weights per bag and apply strong random instance-dropout during training.

Experimental Evaluation

We analyzed the performance of DeepRC and other immune repertoire classification methods on simulated and real data in 4 categories. They include 31 datasets with 785–$5k$ repertoires per dataset and on average $300k$ instances per repertoire. DeepRC outperforms all compared methods, significantly outperforming the second best method, an SVM with a MinMax k-mer kernel.

	Real-world	Real-world data with implanted signals				LSTM-generated data				Simulated
	CMV	s.m. 1%	s.m. 0.1%	m.m. 1%	m.m. 0.1%	10%	1%	0.5%	0.1%	0.05%	avg.
DeepRC	**0.831** ± 0.002	**1.000** ± 0.000	**0.984**± 0.008	**0.999**± 0.001	**0.938**±0.009	**1.000**± 0.000	**1.000**± 0.000	**1.000**± 0.000	**1.000**± 0.000	**0.998**± 0.002	**0.865**± 0.211
SVM (MM)	**0.825** ± 0.022	**1.000** ± 0.000	0.578± 0.020	**1.000**± 0.000	0.531±0.019	**1.000**± 0.000	**1.000**± 0.000	**0.999**± 0.001	**0.999**± 0.002	0.985± 0.014	0.832± 0.203
SVM (J)	0.546 ± 0.021	0.988 ± 0.003	0.527± 0.016	**1.000**± 0.000	0.574±0.019	0.981± 0.041	**1.000**± 0.000	**1.000**± 0.000	0.904± 0.036	0.768± 0.068	0.543± 0.076
KNN (MM)	0.679 ± 0.076	0.744 ± 0.237	0.486± 0.031	0.674± 0.182	0.500±0.022	0.699± 0.272	0.717± 0.263	0.732± 0.263	0.536± 0.156	0.516± 0.153	0.629± 0.126
KNN (J)	0.534 ± 0.039	0.652 ± 0.155	0.484± 0.025	0.695± 0.200	0.508±0.025	0.698± 0.285	0.606± 0.237	0.523± 0.164	0.550± 0.186	0.539± 0.194	0.501± 0.007
Log. Regr.	0.613 ± 0.044	**1.000** ± 0.000	0.585± 0.045	**1.000**± 0.000	0.512±0.015	**1.000**± 0.000	**1.000**± 0.000	**1.000**± 0.000	0.697± 0.164	0.466± 0.103	0.832± 0.204
Log. MIL (KMER)	0.582 ± 0.065	0.541 ± 0.074	0.506± 0.034	0.994± 0.004	0.620±0.153	**0.997**± 0.004	0.718± 0.112	0.637± 0.144	0.571± 0.146	0.528± 0.129	0.662± 0.216
Log. MIL (TCRB)	0.515 ± 0.073	0.503 ± 0.032	0.501± 0.016	0.992± 0.003	0.782±0.030	0.541± 0.086	0.566± 0.162	0.468± 0.086	0.505± 0.067	0.500± 0.121	0.501± 0.015
Burden test	0.699 ± 0.041	**1.000** ± 0.000	0.640± 0.048	**1.000**± 0.000	0.891±0.016	**1.000**± 0.000	**1.000**± 0.000	**1.000**± 0.000	**0.999**± 0.003	0.792± 0.280	0.543± 0.070
Motif binary		1.000 ± 0.000	0.704± 0.028	0.994± 0.003	0.620±0.038	1.000± 0.000	1.000± 0.000	1.000± 0.000	0.999± 0.003	0.999± 0.003	0.899± 0.158
Motif nonbinary		0.920 ± 0.004	0.562± 0.028	0.647± 0.030	0.515±0.031	1.000± 0.000	1.000± 0.000	0.989± 0.011	0.722± 0.085	0.626± 0.094	0.727± 0.189

Interpretability and Retrieval of Important Subsequences

DeepRC allows for the retrieval of important sequences and subsequences via the attention weights of sequence instances or the Integrated Gradients method. Sequences with high attention weights in the real-world dataset correspond to those identified by [2]. For simulated datasets, we successfully retrieved the implanted indicative subsequences from trained DeepRC models.

retrieved motif	S F E N	S F E N	S ^ ^ N	Ss^N
implanted motif	SFEN	SF^{d}EN	SZZN	SZ^{d}ZN
witness rate	0.01%	0.01%	0.1%	0.1%

^... positional feature (sequence center)
Z ... wildcard character
d ... 50% deletion chance

https://github.com/ml-jku/DeepRC
https://github.com/ml-jku/hopfield-layers

References

[1] Mora, T. and Walczak, A. M. How many different clonotypes do immune [...]? *Current Opinion in Systems Biology*, 18:104-110, 2019.
[2] Emerson, R. O., DeWitt, W. S., et al. Immunosequencing identifies signatures of [...]. *Nature Genetics*, 49(5):659, 2017.
[3] Georgiou, G., Ippolito, G. C., et al. The promise and challenge of high-throughput [...]. *Nature Biotechnology*, 32(2):158, 2014.
[4] Greiff, V., Yaari, G., et al. Mining adaptive immune receptor repertoires for biological and [...]. *Current Opinion in Systems Biology*, 2020.
[5] Carbonneau, M.-A., Cheplygina, V., et al. Multiple instance learning: a survey of problem [...]. *Pattern Recognition*, 77:329-353, 2018.
[6] Krotov, D. and Hopfield, J. J. Dense associative memory for [...]. *Advances in Neural Information Processing Systems*, pp. 1172-1180, 2016.
[7] Demircigil, M., Heusel, J., et al. On a model of associative memory with huge [...]. *Journal of Statistical Physics*, 168(2):288-299, 2017.
[8] Hopfield, J. J. Neural networks and physical systems with [...]. *Proceedings of the National Academy of Sciences*, 79(8):2554-2558, 1982.
[9] Ramsauer, H., Schäfl, B., et al. Hopfield networks is all you need. *arXiv*, 2008.02217, 2020.
[10] Vaswani, A., Shazeer, N., et al. Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.

DeepRC code & datasets

modern Hopfield networks blog