We are very grateful to the reviewers for their positive feedback and constructive review of our paper, especially in the light of the interdisciplinary context and the extensive content. We first provide general comments:

(I) Connection of application and theory; extensive content. We agree and are thankful that the reviewers see the importance of the findings. As reviewer 2 suggests, we have focused this work on the application of the theory to repertoire classification. We pursue a separate paper to investigate the theory in more detail. We will elaborate more on the connection between theory and application: The repertoires with 100,000s of instances require a powerful look-up model. We need a model that can store many patterns on which look-up is performed. The theory justifies the usage of modern Hopfield networks with their storage capacity for this challenging task. As reviewer 4 points out, we exploit this storage capacity to solve the immune repertoire classification problem. **(II) SVM.** We thank the reviewers for their interest in the SVM method and will include a more detailed analysis of it. The SVM method is suitable for fewer samples but fails at higher motif complexity due to rigid k-mer representation. DeepRC performs better at higher motif complexity (end-to-end DL with CNN/LSTM embedding offers high flexibility), e.g. datasets "13", "16" in "Simulated". Although in simple settings DeepRC and SVM are on par, DeepRC clearly outperforms SVM in more difficult settings. The difference is significant at p-values of $3\mathrm{e}{-}66$ ("Simulated"), $1\mathrm{e}{-}222$ ("Real-World data with implanted signals"), and $4\mathrm{e}{-}121$ (all 4 dataset categories) with McNemar's test. Nevertheless, we will add more datasets with higher motif complexity to further highlight this trend. **(III) Other compared methods.** We compare to all proposed methods for immune repertoire classification (burden test, logistic MIL) and several methods that we could adapt to this task. We emphasize that a neural net without the Hopfield attention mechanism is already included in the comparison (logistic MIL). Transformers are computationally not feasible because the query matrix would be quadratic in the number of instances. Nevertheless, we will include two other DL models with mean and max pooling instead of attention pooling.

Reviewer 1 *"[…] the authors show an equivalence between […] Hopfield networks and […] transformer models. […] Immune repertoire datasets are then used for experimental evaluation."* We thank the reviewer for outlining this – we also consider these our main contributions. *"The connection between the theoretical developments and the specific NN structure seem tenuous."* We will present this better (see I) to show that the connection is indeed strong. A fixed query is possible for modern Hopfield networks but not used in transformers. *"In particular, it seems that any existing transformer model using attention (e.g., [Yan et al., ACML 2018]) would also enjoy the same guarantees"* The reviewer is right. However, transformers are computationally not feasible here due to large matrix $\mathbf{Q}$, see III. We will cite and add this.*"The paper is well-written; […] helpful to define […] "immune repertoire classification" earlier […]"* We will define this term earlier in the manuscript.

Reviewer 2 *SVM comparison.* See II. *Storage capacity of SVM and other methods.* How many patterns can be stored (shattered) in an NN or SVM is best given by the VC dimension. However, this is no working memory that can actively store instances. We are not aware of other models with this property, except transformers/Hopfield nets. *Why modern Hopfield?* Only Hopfield formalism allowed deriving the storage capacity. We are not aware of a rigorous proof for transformer models. *Comparison to existing methods; DL without Hopfield.* See III. *"[…] connection between the two topics […] both of these bodies of work may be of interest […]."* See I. We agree the transition is not entirely smooth and will improve on that. *Temperature.* We will introduce the temperature parameter in the main text and add more information on Hopfield nets. *Term "pattern".* We will define and elaborate. A data point (=repertoire) consists of many instances (=AA sequences), each mapped to a fixed-sized vector (=pattern). "Pattern" in terminology of MIL corresponds to an instance or representation thereof. The modern Hopfield net stores patterns. *Description of datasets.* We will add the detailed description of the datasets and differences to the main paper. *Verbose introduction/text; notation.* We will condense the introduction and improve notations in the main text. *Other comments.* Thanks, we will adapt as suggested.

Reviewer 3 *"The theory is rather briefly described in the main paper […]."* We will present the theory better (see I). *"[…] matching performance between the real data and simulated data […]"* Yes, a really good catch: The simulation design is based on recent findings and expertise in computational and experimental immunology. *"If these theoretical results [are] rather trivial[…]?"* No, the theoretical results are far from trivial and a new method (modern Hopfield net) is introduced (see I). *"Not sure if this is a theory paper […]"* It is an application paper presenting and using a new theory (see I). *"[…] SVM [has] identical […] performance on many cases […] this particular SVM [is] also a contribution."* Actually, DeepRC outperforms SVM on tasks with complex motifs (see II); will elaborate. *Other comments.* Thanks, handled as suggested.

Reviewer 4 *"[…] relationship between [$\boldsymbol{\xi}$] and $\mathbf{Q}$."* A $\mathbf{Q}$ with one row is $\boldsymbol{\xi}^T$. *"[…] $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are matrices but in Equation 4, $\boldsymbol{\xi}$ and $\mathbf{V}$ are 1-d vectors."* We will explain this better. In the transformer notation, single queries that are vectors become multiple queries that are matrices. *DeepTCR.* DeepTCR is not a method by itself but a collection of software building blocks. Nevertheless, we will include more compared methods (see III). *"[…] how long is each repertoire?"* Simulated data: 316K, LSTM-generated: 285K, real-world with implanted signals: 10K, real-world: 299K (see Tables A2/A3 for details). *Duplicates.* We will provide details on handling duplicates. *Why CV factor 5? Also for DeepRC?* A CV procedure with 5 folds (also for DeepRC) was the border of computational feasibility. *"SVM MinMax is close on the heels of DeepRC. Why […] DeepRC over SVMs?"* Actually, there is a large gap between the methods in difficult settings (see II). DeepRC as a DL method requires a larger amount of samples. However, DeepRC, as an end-to-end DL model, can learn to detect more complex/flexible motifs with its CNN/LSTM sequence embedding. *"Are the DeepRC-identified motifs biologically interpretable?"* Yes. They may be interpreted as reproducible (across individuals) traces of immune responses. I.e. they have been useful in fighting a given disease, e.g. recognize pathogens. *Other comments.* Edited as suggested.