PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection

Andrea Di Gioacchino[1‡], Jonah Procyk[2‡], Marco Molari[1,3,4], John S. Schreck[5], Yu Zhou[2], Yan Liu[2], Rémi Monasson[1]*, Simona Cocco[1]*, Petr Šulc[2]*

1 Laboratoire de Physique de l'Ecole Normale Supérieure, PSL & CNRS UMR8063, Sorbonne Université, Université de Paris, Paris, France, **2** School of Molecular Sciences and Center for Molecular Design and Biomimetics, The Biodesign Institute, Arizona State University, Tempe, Arizona, United States of America, **3** Biozentrum, University of Basel, Basel, Switzerland, **4** Swiss Institute of Bioinformatics, Basel, Switzerland, **5** National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, Colorado, United States of America

‡ These authors contributed equally to this work and are ordered alphabetically.
* remi.monasson@phys.ens.fr (RM); simona.cocco@phys.ens.fr (SC); psulc@asu.edu (PŠ)

Citation: Di Gioacchino A, Procyk J, Molari M, Schreck JS, Zhou Y, Liu Y, et al. (2022) Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. PLoS Comput Biol 18(9): e1010561. https://doi.org/10.1371/journal.pcbi.1010561

Editor: Jinyan Li, University of Technology Sydney, AUSTRALIA

Received: May 23, 2022

Accepted: September 12, 2022

Published: September 29, 2022

Data Availability Statement: The source codes used for training and sampling from RBM, as well as the trained supervised learning models, is available at github.com/adigioacchino/RBMsForAptamers. The data used for this work is publicly available through the Zenodo service, with the following Digital Object Identifier (DOI): 10.5281/zenodo.6341687.

Funding: This work was supported by National Science Foundation (No. 2155095 and TG-BIO210009 to P.S.), Agence Nationale de la

Abstract

Selection protocols such as SELEX, where molecules are selected over multiple rounds for their ability to bind to a target of interest, are popular methods for obtaining binders for diagnostic and therapeutic purposes. We show that Restricted Boltzmann Machines (RBMs), an unsupervised two-layer neural network architecture, can successfully be trained on sequence ensembles from single rounds of SELEX experiments for thrombin aptamers. RBMs assign scores to sequences that can be directly related to their fitnesses estimated through experimental enrichment ratios. Hence, RBMs trained from sequence data at a given round can be used to predict the effects of selection at later rounds. Moreover, the parameters of the trained RBMs are interpretable and identify functional features contributing most to sequence fitness. To exploit the generative capabilities of RBMs, we introduce two different training protocols: one taking into account sequence counts, capable of identifying the few best binders, and another based on unique sequences only, generating more diverse binders. We then use RBMs model to generate novel aptamers with putative disruptive mutations or good binding properties, and validate the generated sequences with gel shift assay experiments. Finally, we compare the RBM's performance with different supervised learning approaches that include random forests and several deep neural network architectures.

Author summary

We show that two-layer neural networks, Restricted Boltzmann Machines (RBM), can be successfully trained on sequence ensemble datasets from selection-amplification experiments. We train the RBM using datasets from aptamer selection experiments on thrombin protein, and show that the model can successfully generalize to the test set to predict

Recherche (RBMPro CE30-0021-01 and ANR-19 Decrypted CE30-0021-01 to S.C. and R.M.), and European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (No 101026293 to A.D.G.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

binders and non-binders. The log-likelihood assigned to a sequence by the RBM is correlated with the sequence fitness as quantified by the amplification between different rounds of selection. We further show that that the model is interpretable and by inspecting the weights of the model, we can identify structural motifs that are characteristic of the good binders. We explore the usage of the RBMs to identify which of the possible protein exosites the aptamers bind to. We show that the RBM can also be used for unsupervised clustering. Finally, we use RBMs to generate novel aptamers, and we experimentally verify predicted binding and non-binding sequences generated from the RBM.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Discovery and design of molecules that can specifically bind a given target molecule is a key problem in diagnostics, therapeutics and molecular biology in general. Multiple different experimental approaches exist to select specific molecular target binder such as antibodies, short peptides, proteins or small molecules. Single stranded oligonucleotides (DNA or RNA) have also been shown to be able to specifically bind with high affinity to a plethora of various targets, including small metabolites, proteins, nucleic acids, viruses, exosomes, and cells of specific tissue [1–10], showing promise for applications that range from diagnostics to targeted disease therapy [11]. These short oligonucleotides, called aptamers, are selected from an initial pool of sequences by a procedure known as Systematic Evolution of Ligands by Exponential Enrichment (SELEX) [12, 13]. This method consists of multiple rounds of selection, where aptamers that bind strongly enough to the protein target are selected and amplified for the next round, until few strong binders are obtained. The advantages of using DNA or RNA include low cost of synthesising these molecules and relative ease of their manipulation in the laboratory setting as opposed to other selection methods such as peptide or antibody selection [14, 15]. Oligonucleotides can be denatured and refolded many times, allowing for multiple selection rounds. On the other hand, as they are composed of four possible types of bases (A, C, G and T/U), they do not offer such chemical diversity as antibodies. Thus, the range of targets that aptamers can be selected to bind strongly to is limited to some extent. However, chemical modifications of the nucleic bases can increase the chemical space of the aptamers and provide diverse sequence libraries from which strong binders can be selected against a variety of targets [16].

With the advance of next generation sequencing and high-throughput biological and molecular dataset production, various machine learning methods have been used to process biological sequences datasets, with applications including classifications, binding prediction, and molecular design [17]. While a significant improvement has recently been achieved in using deep learning for protein or RNA structure predictions [18, 19], predictions of binding interactions and *de novo* design of molecular binders remain outstanding significant challenges. So far, it is primarily the prediction of interaction between a small molecule ligand and a target protein that has received attention from the machine learning community, as such approaches are at the basis of the drug screening pipeline [20]. Motif-finding and clustering-based methods, combined with secondary structure prediction tools, have been previously

developed for processing SELEX datasets [21–23, 23–25]. Currently, the SELEX dataset processing typically involves clustering and identifying a common motif in aligned sequences and then selecting representative aptamers from the last round of selection and verifying their binding affinity to the target.

A challenging task in the analysis of SELEX experiments is the quantification of the aptamer fitness, which determines the sequence landscape evolution at each selection round. Several approaches have been introduced in the past, based on *in silico* molecular dynamics simulations [26, 27], on clustering in sequence space together with enrichment measurements [28], and on additional, direct fitness estimation experiments [29]. These methods proved useful to estimate the fitness of a limited number of selected sequences or of large classes of similar sequences, but they seem unable to assign in a reliable way a fitness score to each molecule observed in final rounds of SELEX.

Over the last decade, deep neural networks (DNN) have become a popular machine learning tool in many areas, such as image recognition or natural language processing, and are now increasingly applied in chemical and biological data processing workflow [30–33]. However, training DNNs typically requires large datasets, which can be challenging and expensive to obtain from biological experiments. DNNs have many free parameters, which makes it difficult to identify and interpret particular features of the molecule that are attributed to its ability to bind a given target. The presence of errors in the sequence dataset, coming e.g. from experimental error in affinity measurements or sequencing errors, adds further difficulties to training as well as to interpretability. Machine-learning methods for sequence ensembles include inverse models from statistical physics, such as direct coupling analysis (DCA) methods [34], which have been previously successfully used to infer native contacts and guide folding of RNA and proteins based on homologous sequence alignment [35, 36], as well as to generate functional enzymes based on functional protein alignments [37] and protein recognizing RNA [38]. They infer parameters of maximum-entropy models, which are fixed by the requirement that the conservation of single residues and pairs of residues given by the model match the values observed in the sequence alignment. More recently, Restricted Boltzmann Machine (RBM) architectures, a neural network with a bipartite graph structure, have been successfully applied as a generative model for protein domain sequences [39], as well as a predictor of peptides that will be presented on Major Histocompatibility Complexes [40]. They present an intermediate level of complexity between the direct coupling models and DNNs, as they can be trained to recognize multi-residue coupling as opposed to pairwise interactions, but due to limited number of weights between the two neuron layers, the parameters can still be interpreted and rationalized.

Here, we apply RBM models to a set of DNA sequences obtained from the prior experimental work of some of us that used SELEX method to obtain thrombin aptamers [41] (Fig 1). We show that the sequence likelihood assigned by the RBM can be directly related to the fitness of that sequence in the experimental selection. Moreover an RBM model that is trained on an earlier round of the selection is able to predict fitness of sequences in the next rounds not seen during the training, showing remarkable generalization capabilities. We further show that we can identify the sequence motifs conferring large likelihood to an aptamer sequence and that RBM's hidden unit input can be used to cluster sequences. We show the capability of the RBM to predict binding affinity and generate new monovalent aptamers, which are good binders to one of the two thrombin binding sites, by gel shift assays. We investigate how taking into account the individual sequence counts from the experiment in the training data changes the properties of the inferred RBM model. Lastly, we also explore several supervised learning approaches that include random forest and various DNN architectures, but find them difficult to train and with poor generalization performance on our dataset.

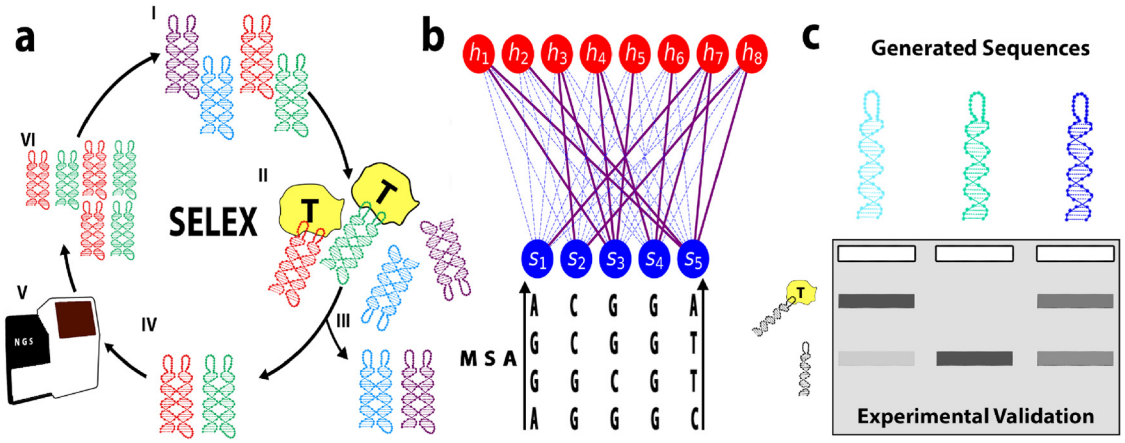

Fig 1. Schematic view of the SELEX experiment and the RBM-based analysis. a: The SELEX procedure used to obtain DNA aptamers that bind to thrombin consists of the following steps: I) We start with an initial library of DNA sequences. II) DNA aptamers compete with each other to bind to thrombin. III) Sequences that are unbound (or bound too weakly) are washed away. IV) Remaining bound sequences dissociate after the sample is heated up. V) Binding sequences are sequenced. VI) Using polymerase chain reaction (PCR), multiple copies are made of the remaining sequences, resulting into a new library of aptamers for the next round of selection. **b**: The sequenced aptamers from respective rounds of the SELEX protocol are used to train the parameters of the Restricted Boltzmann Machine model. In this unsupervised neural network architecture, a layer of visible units carry the aptamer sequence, while the layer of hidden units extract representations. The weighted connections between the two layers are learned through maximization of the log-likelihood of the sequences obtained through SELEX. **c**: Single loop sequences generated using the Restricted Boltzmann Machine model are experimentally validated using gel assays.

https://doi.org/10.1371/journal.pcbi.1010561.g001

Results

Dataset obtained from SELEX procedure

In a prior work [41], some of us used the SELEX method to obtain a bivalent DNA nanostructure that binds to a thrombin protein. In this DNA SELEX procedure, an initial library of about 10^{15} unique DNA sequences with all about the same length were exposed to the target tethered to a surface. The non-binding sequences were then washed away, while the binding sequences were collected (and optionally also sequenced). After amplification with PCR they served as the sequence library for the next cycle of SELEX. Cycles were repeated until binders of the desired binding affinity were found. The washing intensity was increased in later rounds to obtain stronger binders. In the particular experimental dataset used in Ref [41], the SELEX procedure was performed on a DNA nanotile (Fig 1), consisting of a joined-double helix region with two loops of 20 nucleotides each. While the double-helix nanotile structure was conserved across all DNA structures, the two respective loops were variable, starting from the initial random library. The SELEX procedure is schematically shown in Fig 1 and consisted of eight selection rounds. The binding molecules were sequenced in rounds 5 (891959 sequences out of which 891914 unique), 6 (736436 sequences out of which 735974 unique), 7 (750926 sequences out of which 744597 unique) and 8 (725431 sequences out of which 719413 unique), and form the datasets we use here for training our models.

For each round, our dataset includes the sequence of the two (left and right) respective variable loop regions of the DNA nanotile, as well as the number of counts of the two-loop sequence, corresponding to the number of times it was sequenced in the experiment. In typical SELEX protocols, the sequences with the largest number of counts in the last rounds are considered the best binders.

Restricted Boltzmann Machine model

We use a Restricted Boltzmann Machine (RBM) to learn the probability distribution over the set of aptamers based on the sequences collected through the SELEX procedure. An RBM is a probabilistic model, represented by a bipartite graph consisting of L "visible" and M "hidden" units (shown schematically in Fig 1B). It assigns a probability $p(\boldsymbol{s}, \boldsymbol{h})$ to a system state, given by two parts: the configuration of visible units, $\boldsymbol{s} = (s_1, \ldots, s_L)$, where s_i = A, C, G or T are the nucleotides on site i along the aptamer sequence, and the configuration of the hidden units, $\boldsymbol{h} = (h_1, \ldots, h_M)$, meant to extract latent factors of variation in the visible configurations. The likelihood of a sequence $\boldsymbol{s}$ is formally obtained by marginalizing over all possible latent configurations (not observed in the data), $p(\boldsymbol{s}) = \int d\boldsymbol{h}\, p(\boldsymbol{s}, \boldsymbol{h})$. The number L of visible units can be set to 40 to model full two-loop sequences or restricted to 20 to describe each loop independently. These two possibilities will be referred to as, respectively, D (Double loop) and S (Single loop) in the following.

Training a RBM consists in finding the parameters (in particular, the couplings between the layers) so that the log-likelihood of the observed data,

$$\mathcal{L} = \sum_{\boldsymbol{s} \in \text{round r}} \log p(\boldsymbol{s}) \ , \tag{1}$$

is maximized. Here the sum over $\boldsymbol{s}$ is over the sequences observed at a fixed selection round, say, r, of the SELEX experiment. Each sequence may therefore appear multiple times, depending on the number of its counts. We will denote this model with C (Count). An alternative is to include in the sum in Eq (1) unique sequences only. The resulting model, labelled with U (Unique), has different properties, which we will discussed below.

The maximization of $\mathcal{L}$ is a computationally difficult problem, but several effective techniques to obtain good parameter values have been developed, for instance contrastive divergence [42] and persistent contrastive divergence [43]. As described in Methods Sec Restricted Boltzmann Machine: Definition, training, sampling, we train, following Ref [39], the RBM using persistent contrastive divergence and using double Rectified Linear hidden units, with a L_1^2 regularization scheme. This regularization favors sparse weights, and enhances interpretability of the trained model.

RBM's log-likelihood is an accurate predictor of the aptamer's fitness

Fig 2A shows the distributions of log-likelihoods of sequences collected at SELEX rounds r = 5 to 8, estimated with an RBM trained on double-loop aptamer sequences with counts measured at round 6 (RBM-DC, see S1 Appendix). At round 5 three peaks are apparent. The logos of the sequences in each peak are shown in Fig 2B. The peak at low log-likelihoods is characterized by highly variable sequences, weakly enriched in C, G nucleotides. The peak at intermediate values correspond to sequences with a structured loop (the left one, for most sequences), including a G-quadruplex motif. In the high log-likelihood peak a similar G-quadruplex motif appears on both left and right loops (for more details, see also Sec The log-likelihoods of the aptamers can be explained by the additive contributions of their left and right loops). From round 6 to 8 the peaks at low and intermediate log-likelihood values are progressively depleted, and the peak at high log-likelihood gets more and more populated. This enrichment strongly suggests a positive correlation between the score assigned by the RBM and the fitness.

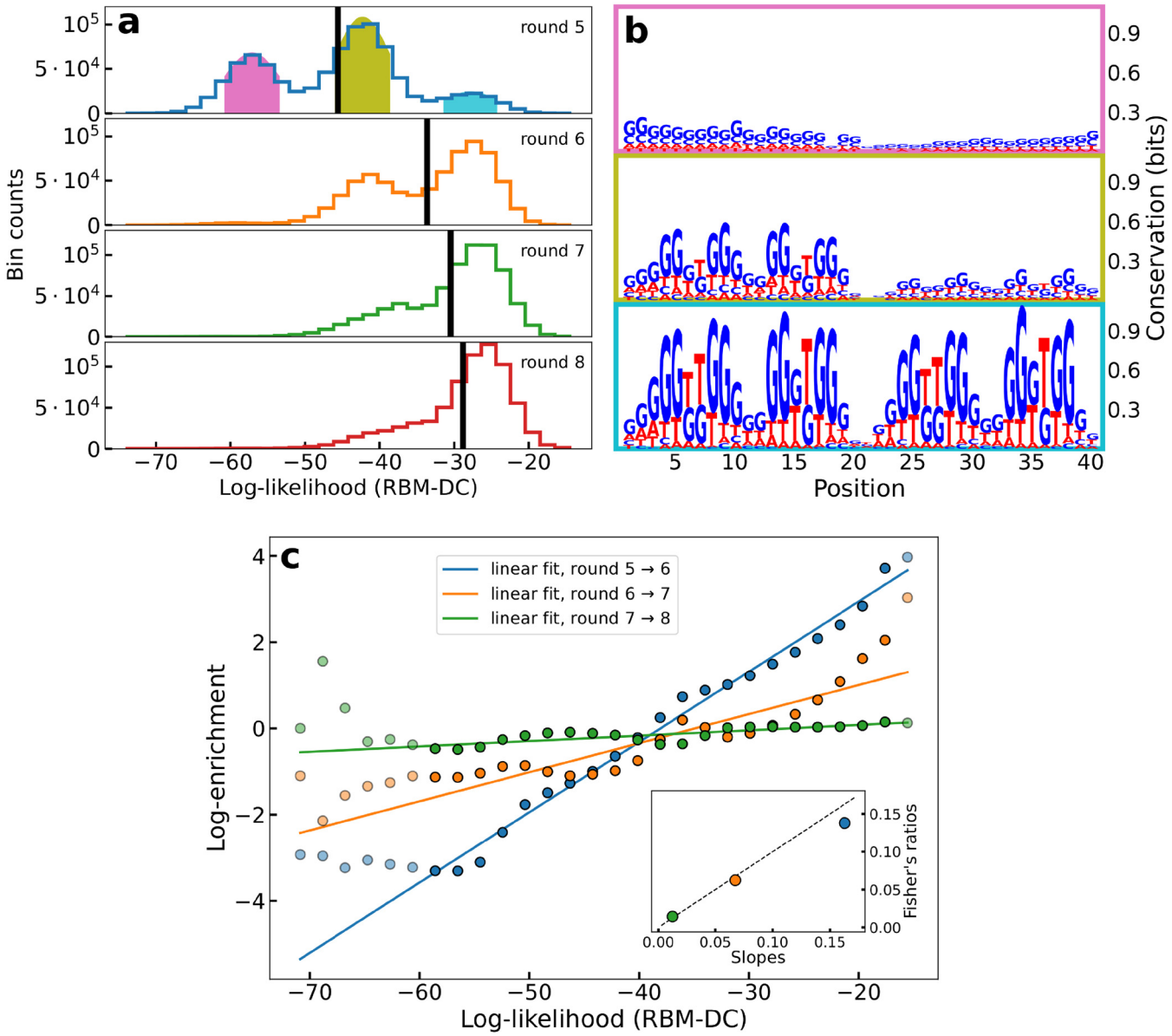

Fig 2. The RBM log-likelihood is strongly correlated to sequence fitness. a: Histograms of the log-likelihood of all sequences in the dataset at different rounds, obtained through RBM-DC trained on the sequences from round 6. The black line denotes the average log-likelihood. **b**: Logos of the sequences in each colored-shaded peak of the log-likelihood observed at round 5. **c**: Enrichment log-ratios log $\mathcal{E}$ vs. log-likelihoods log p averages over the sequences in each bin of the histograms of panel A. The three sets of points corresponds to rounds $r = 6, 7, 8$. Linear fit are estimated from points with log-likelihood in the interval $(-60, -17)$ (not shaded in the plot) only, to exclude under-sampled bins cumulatively representing 0.5%, 0.3%, and 0.3% of the sequences, respectively at rounds 6, 7 and 8. Inset: scatter plot of the slopes of the linear fits (x-axis) and of the log-likelihood Fisher ratios (y-axis); linear dashed line: $y = x$.

https://doi.org/10.1371/journal.pcbi.1010561.g002

In a population genetic framework, the fraction q of aptamers with sequence $\mathbf{s}$ changes from round $r-1$ to round r according to

$$q_r(\mathbf{s}) = \frac{e^{\alpha_{r-1}F(\mathbf{s})}}{\left\langle e^{\alpha_{r-1}F(\mathbf{s}')}\right\rangle_{\mathbf{s}'\in r-1}} q_{r-1}(\mathbf{s}), \tag{2}$$

where $\langle O(\mathbf{s}')\rangle_{\mathbf{s}'\in r-1} = \sum_{\mathbf{s}'} q_{r-1}(\mathbf{s}')O(\mathbf{s}')$ denotes the average of the observable O over the distribution of sequences at round $r-1$. The fitness $F(\mathbf{s})$ encompasses the capability of an aptamer $\boldsymbol{s}$ of

binding its target, as well as other chemical properties, such as its affinity to PCR amplification. Parameter α_{r-1} represents the selection strength from round $r-1$ to r, which can be tuned in practice e.g. by varying with the intensity of washing in SELEX selection.

According to Eq (2), formally valid for an infinite-size population only, the fitness $\alpha_{r-1}\, F(\mathbf{s})$ is, up to a sequence-independent additive constant, equal to the logarithm of the enrichment ratio $\mathcal{E}_r(\mathbf{s}) = C_r(\mathbf{s})/C_{r-1}(\mathbf{s})$, where $C_r(\mathbf{s})$ is the number of counts of sequence $\mathbf{s}$ at round r. However, the extreme subsampling of sequences at each round in our dataset prevents us from using empirical enrichment ratios $\mathcal{E}$ to estimate the fitnesses, and their correlation with log-likelihoods, see S1 Fig. For instance, only $f_{\text{shared}} = 0.5\%$ of the sequences observed in round 7 or round 8 are present in both rounds, and among these sequences, about $f_1 = 70\%$ have count $C = 1$ in both rounds. In earlier rounds, e.g. 5 and 6, the situations is even worse, with fractions $f_{\text{shared}} = 0.01\%$ and $f_1 = 93\%$.

To obtain more reliable enrichment ratios we gather all sequences $\mathbf{s}$ having similar log-likelihoods $\log p(\mathbf{s})$, and introduce their cumulative number of counts, $C(\ell, r)$. More precisely, $C(\ell, r)$ is defined as the number of counts in the ℓ^{th} bin of the histogram of log-likelihoods in Fig 2A. We then define the effective enrichment ratio of bin ℓ through $\mathcal{E}_r(\ell) = C_r(\ell)/C_{r-1}(\ell)$. Fig 2C shows the scatter plots of the enrichment log-ratios $\log \mathcal{E}_r(\ell)$ vs. the log-likelihoods ℓ, for rounds $r = 6, 7, 8$. Very strong correlations are observed, with coefficients of determination $R^2 = 0.99$, 0.83 and 0.66 and slopes 0.16, 0.07, 0.01 for, respectively, the pairs of rounds $5 \to 6$, $6 \to 7$, and $7 \to 8$. The smaller values of the slopes of the linear regressions at later rounds suggests that the effective selection strength α_{r-1} appearing in Eq (2) is weaker in the last SELEX rounds than in the previous ones. This interpretation is supported by the fact that the 10 different single-loop aptamers with largest count numbers at round 8 do not increase exponentially in the last rounds considered here, as shown in S2 Fig.

The linear relationship between the RBM log-likelihood $\log p(\mathbf{s})$ and the sequence fitness $F(\mathbf{s})$ suggests an alternative way to estimate the selection strengths α_r. Fisher's fundamental theorem (see for instance [44] for a review) postulates that the selection strength can be estimated through the ratio of the increase of the average fitness and of the its variance, $\alpha_{r-1} = (\langle F\rangle_r - \langle F\rangle_{r-1})/\text{var}(F)$. We compute these Fisher's ratios using $\log p$ as a proxy for F to estimate the selection strengths at the various rounds. Results are shown in the inset of Fig 2C, and agree with those obtained directly from the slopes of the linear regressions. The precise relation between the fitness and the log-likelihood is further examined in Discussion section.

The log-likelihoods of the aptamers can be explained by the additive contributions of their left and right loops

To examine the cooperative binding of the left and right loops of the aptamer nanostructure at a given round of SELEX, we have trained RBM models on the 20 nucleotide-long single loops only. In practice, RBM-SC trained on all left (L) loop subsequences, on all right (R) loop subsequences, or on both of them show very similar properties (S3 Fig), and we hereafter report results with the latter model. We show in Fig 3A the log-likelihoods of the L and R loops for all aptamers at round 5. We observe the presence of four peaks in the joint distribution, corresponding to all possible combinations of the two peaks at, respectively, low ($\simeq \mathcal{L}_-$) and high ($\simeq \mathcal{L}_+$) log-likelihoods present in the marginal distributions for L or R loops.

As shown in Fig 3B, aptamer sequences previously characterized as having low (in pink), intermediate (in olive) and high (in turquoise) log-likelihoods, see Fig 2A, occupy the four corners of the joint-distribution plot. Therefore, high-log-likelihood aptamers have both L and R loops with high log-likelihoods $\mathcal{L}_+$, while the L and R loops of the low-log-likelihood aptamers

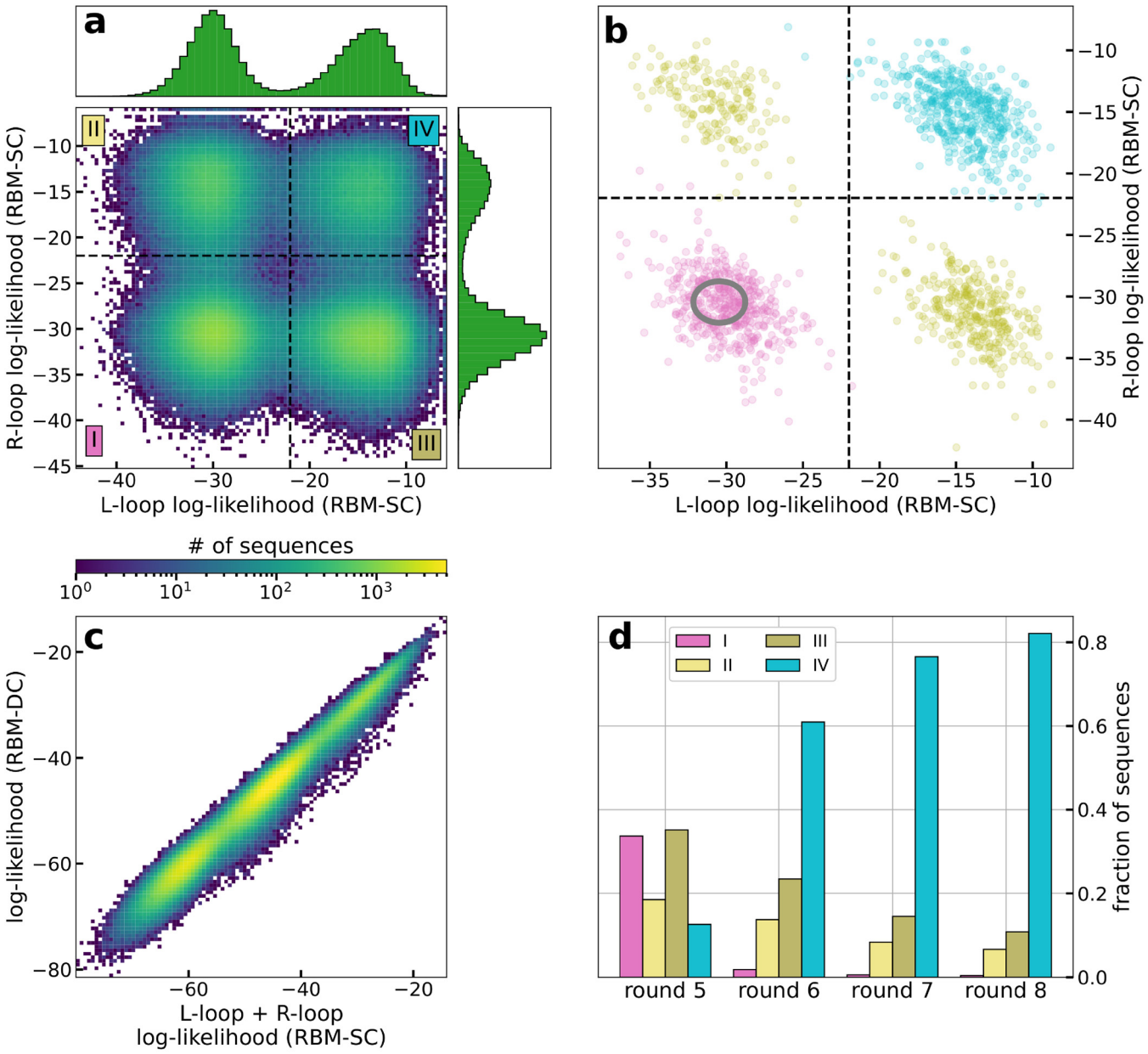

Fig 3. Contribution of left and right loops in the RBM log-likelihood. a: Joint distribution of the log-likelihoods of the L and R loop subsequences at round 5, estimated with RBM-SC trained on subsequences at round 8. The insets show the marginal distributions for both loops. **b**: Same as panel A for 2000 aptamer sequences attached to each of the three colored peaks in Fig 2A (same color code). The gray ellipse shows the distribution of the log-likelihoods of uniform random sequences (center: mean values, ellipse: 2 standard deviations from the mean). **c**: Log-likelihoods of the aptamers in round 5 (estimated with RBM-DC see S1 Appendix) vs. sums of the log-likelihoods of their L and R loops (estimated with RBM-SC). **d**: Fractions of sequences in regions I to IV of panel A at rounds 5, 6, 7 and 8, estimated with the same RBM-SC model as in panel A.

https://doi.org/10.1371/journal.pcbi.1010561.g003

have both low log-likelihoods $\mathcal{L}_-$. Intermediate aptamers have one loop, either L or R, with high loglikehood value $\mathcal{L}_+$ and the other with low log-likelihood $\mathcal{L}_-$.

Fig 3C shows the scatter plot of the log-likelihoods of the full aptamers (estimated with RBM-DC) vs. the sums of the log-likelihoods of their L and R loops (estimated with RBM-SC). We observe an excellent linear correlation ($R^2 = 0.99$), indicating that both loops contribute additively to the score of the full aptamer. This linearity also explains the three peak structure

of the aptamer log-likelihoods in Fig 2A, approximately located at $2\mathcal{L}_-$, $\mathcal{L}_- + \mathcal{L}_+$, and $2\mathcal{L}_+$. Moreover, thanks to this linearity, the selection of the aptamer population from one SELEX round to the next one (Fig 2) can be predicted also at the level of single-loop aptamers (see S4 Fig).

Fig 3D shows the fractions of sequences in the four regions labelled I to IV of the L and R log-likelihoods at successive rounds of selection, see Fig 3A. As observed in Fig 2A for the full aptamer sequences we see a progressive enrichment in sequences for which both L and R loops have high log-likelihoods. However, we also observe a substantial fraction of sequences ($> 15\%$) at round 8, in which one loop only has high log-likelihood. The cognate 20-nucleotide sequences, with low log-likelihood on the other loop, will be called parasite in the following, as they are likely to be selected only due to the ability of the other loop to bind thrombin. To check this hypothesis we generate random aptamer sequences, in which the 40 nucleotides are drawn uniformly at random. As shown in Fig 3B these random aptamer sequences are located in the $(\mathcal{L}_-, \mathcal{L}_-)$ corner, and do not differ much from the pink sequences in terms of log-likelihood, see gray ellipse in Fig 3B. Notice that removing the parasite sequences from the training set of RBM-SC does not significantly modify the estimation of log-likelihoods, see S5 Fig, which shows the robustness of the RBM model against the presence of random sequences in the data. The identification of parasite sequences has important consequences for the design of new aptamers based on the RBM model, as discussed in the next section.

RBM parameters reveal functional features of the aptamer sequences

We next extract the features that contribute the most to the likelihood of the sequences by studying weights between hidden units and visible layer (Fig 1B). To enhance the interpretability of the RBM weights connecting input and and hidden layers we enforce their sparsity through appropriate regularisation (see Methods Sec Restricted Boltzmann Machine: Definition, training, sampling and Ref [39]). Fig 4A–4C (left) show the sequence logo of the three weights of RBM-DC with largest Frobenius norms (S6 Fig); the height of nucleotide symbol s in position i for hidden unit μ represents the value of the weight $w_{\mu i}(s)$.

We first observe that the weights are strongly localized either on the left or the right loop. The lack of correlation between the left and right loop sequences holds for all weights (S7 Fig), and is compatible with the additivity of their contributions to the aptamer log-likelihood in Fig 3C.

A closer look at the sequence-dependence of the logos in Fig 4A–4C shows they are G-rich and match parts of G-quadruplex motifs. For instance, the hidden unit focusing on the right loop in Fig 4A, is strongly activated when the motif AGGTTGG is present on the L loop in positions 33–39. Other L subsequences lead to much weaker activities (in absolute value), see right subpanel in Fig 4A. A similar observation holds the left loop in Fig 4C, with the motif GNNTGGTGTGGNTGG in positions 4–18 which is compatible with a G-quadruplex structure. Other features are also detected by the RBM. As an example the weight logo in Fig 4B is identifying long-range correlations across positions 1–20 associated consisting in a AT-rich motif and is present in some of the training sequences (see histogram in right subpanel).

Another relevant set of parameters learned from the data are the local fields acting on the visible variables. These parameters follow quite closely the nucleotidic profile of with the dataset, so they reflect a general enrichment in G-content, particularly in the positions most used to form G-quadruplexes (see S8 Fig for the local fields of RBM-DC trained at round 6 and for the conservation logo of the sequences used to train the model).

We then explore the capability of RBM to provide low-dimensional representation of sequences. Prior experimental work [41] identified four different families of thrombin-binding

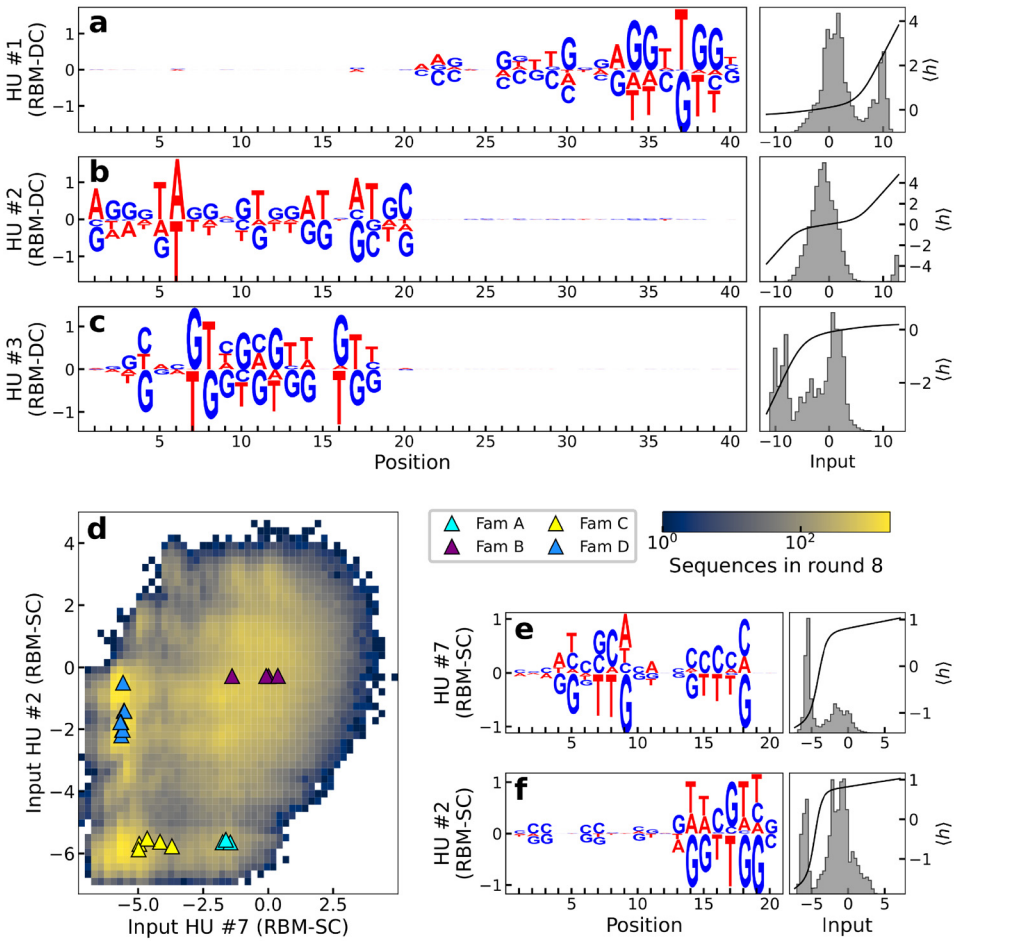

Fig 4. Weights learned by the RBMs have biological interpretations. a-c: Left: logos of three weights $\mu = 1, 2, 3$ with largest Frobenius norms for RBM-DC trained on round 8 aptamer data; Right: histograms of the inputs $I_\mu = \sum_i w_{\mu i}(s_i)$ (where $w_{\mu i}$ is the weight of the connection between hidden unit μ and visible unit i for nucleotide s_i) to the corresponding hidden units for the sequences **s** in the dataset (gray) and average activity (black). **d**: The four families identified in [41] are separated in different clusters in the two-dimensional subspace spanned by the inputs to hidden units 2 and 7 of RBM-SC (trained on loop subsequences at round 8). **e, f**: Logo, distribution of inputs and average activity of the same hidden units as in panel D.

https://doi.org/10.1371/journal.pcbi.1010561.g004

aptamers (named A, B, C and D), based on sequence alignment and manual curation. We show in Fig 4D the value of inputs I_μ of two hidden units of single-loop RBM-SC, ranked 2 and 7 in terms of weight Frobenius norms able to cluster these four families. Each hidden unit's activity (see Methods) has a bimodal distribution (Fig 4E and 4F), and the combinations of these modes identify the four families.

RBM trained from unique sequences generate diverse aptamers capable of binding thrombin

After having established that the RBM log-likelihoods and the fitnesses of the aptamers in our dataset are strongly interrelated, we now use the RBM model to generate new sequences *in silico* (see Methods Sec Restricted Boltzmann Machine: Definition, training, sampling). Note that the number of available sequences at any round, $<10^6$, is much smaller than the number of possible sequences over 20 nucleotides, $4^{20} \simeq 10^{12}$. Hence, it is a non trivial problem to

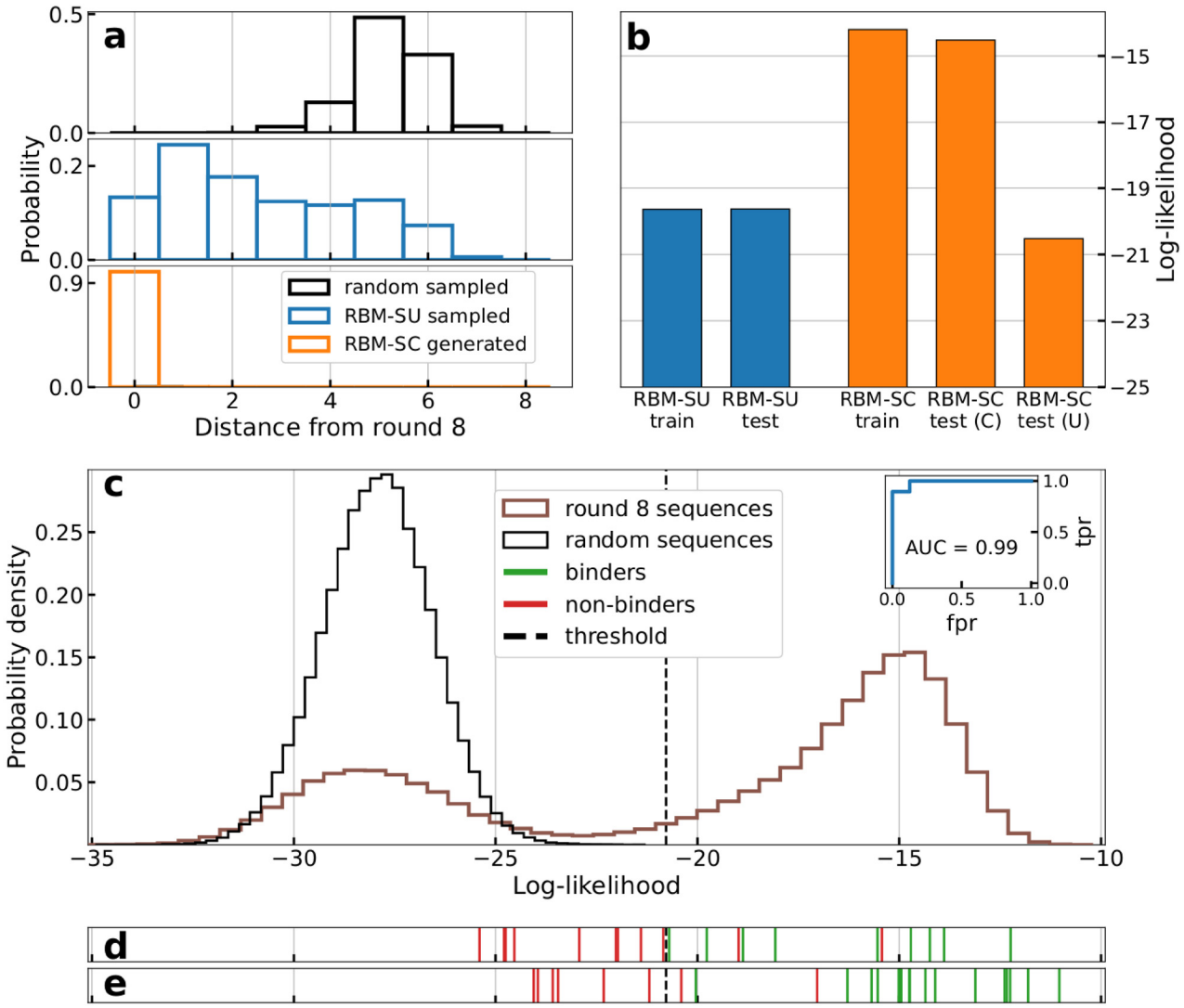

Fig 5. RBM can be used to design new aptamers binding thrombin. a: Histograms of distances to dataset of the best (top 5% in terms of log-likelihoods) sequences generated by RBM-SC (orange) and RBM-SU (blue) trained on round 8 data. The black histograms show the distribution obtained with sequences generated uniformly at random. **b**: Average log-likelihoods of training (90% of the unique sequences observed at round 8, chosen at random) and test (remaining 10% of unique sequences) sets for RBM-SU and RBM-SC (after re-introduction of counts). For RBM-SC, the average log-likelihood of the test set is shown either when weighing each sequence with (C) or without (U) its counts. For the RBM-SU model we show the unweighted average-log likelihood on the training and testing sets respectively. **c**: Histogram of the log-likelihoods of all unique aptamers observed in the last round (blue line) and of uniformly random sequences (orange line), computed with RBM-SU trained on the same data (blue line). Inset: AUC computed on the sequences generated by the RBM-SC model (panel E). **d**: Vertical lines locate the log-likelihoods of sequences experimentally validated to be binders (green) or non binders (red). Sequences taken from a preliminary set described in S1 Table. Results allow us to determine the binding/non binding threshold, located with the black dashed line. **e**: Same as panel D, for sequences designed with the RBM-SU model trained on round-8 sequences, as described in Sec RBM trained from unique sequences generate diverse aptamers capable of binding thrombin, see Table 1.

https://doi.org/10.1371/journal.pcbi.1010561.g005

reconstruct the full likelihood landscape from such undersampled data, and use it to generate new binders.

Sampling RBM-SC trained on round-8 data reveals a lack of diversity in the sampled sequences: all the generated sequences with high log-likelihoods are already present in the dataset (Fig 5A). RBM-SC rightly assigns high scores to the strong binders present at the end of SELEX procedure, but is unable to generate diverse sequences with high scores (Fig 5A).

We then train another model, called RBM-SU, by maximizing the sum of the log-likelihoods of unique sequences in round 8 dataset (composed of 382094 unique single-loop sequences), see Eq (1). Details of the training procedure are given in S1 Appendix. The rationale for this approach is two fold. First, 8^{th} round data are expected to include better binders and much less parasite sequences than earlier rounds. Second, discarding the sequence counts prevents the model from being dominated by few very good binders to thrombin.

The effective diversity of training data is reflected in the generated sequences from RBM-SU model. A large fraction of sequences generated by RBM-SU with top log-likelihoods are not present in the dataset, contrary to what found with RBM-SC, see Fig 5A. In addition, about 30% of generated sequences are 4 or more nucleotides away from the dataset, as is the case for the majority of randomly generated sequences of length 20 nucleotides. Furthermore, we show in Fig 5B that RBM-SU exhibits excellent generalization properties. The log-likelihood of test data (unique sequences present at round 8 but not used for training) is very close to the one of the training data. On the contrary, RBM-SC essentially assigns high scores to high-count sequences in the training data, and shows poor generalization.

We have next experimentally tested the binding to thrombin of some aptamer sequences to validate the ability of the RBM-SU to predict binding and to generate *de novo* binders. The 20-nucleotide DNA sequences are first inserted into the loop of a hairpin with fixed 18 base-pair-long stem. To estimate the binding affinity to thrombin we use native gel shift assay, where we incubate the thrombin protein with the hairpin aptamer, see Methods Sec Thrombin binding assay and S2 Appendix.

A set of 16 sequences listed in S1 Table (excluding the control sequences listed in the Table), together with 4 binders, experimentally validated in [41] and named ThA, ThB, ThC and ThD, is first used to estimate the log-likelihood threshold above which a sequence is predicted to bind thrombin, see Fig 5D, where the log-likelihoods of tested sequences are represented as vertical red and green lines, for verified non-binders and binders respectively.

We then propose a set of 27 sequences to test (r1-r27 in Table 1): 2/3 of them are *de novo* designed sequences generated from the RBM-SU model, and the remaining 1/3 are present in round 8. *De novo* sequences are chosen to test the power of the RBM model to produce good thrombin binders, or to predict critical mutations transforming binders into non binders. Sequences already present in the round-8 data are chosen to test non trivial RBM predictions, *e.g.* sequences with low or high counts having, respectively, high or low log-likelihoods. The detailed description of these sequences and of the design criteria is found in Method Sec Design of single-loop aptamers with RBM.

Over the 27 sequences to test, 21 sequences were above threshold, and therefore predicted as binders and 6 sequences below threshold, predicted as non binders. The experimental gel assays are shown in Fig 6. Overall, 93% of the RBM predictions (binder or non binder) are confirmed by experiments. The log-likelihoods of the tested sequences, along with the RBM predictions and the experimental findings are reported in Table 1 and represented with the experimental results in Fig 5E.

These results show that the log-likelihood provided by RBM-SU is an accurate predictor of the capability to bind thrombin. We show in the inset of Fig 5C the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC = 0.99) for RBM-SU-generated sequences. Let us stress that RBM-SC, shows poor performance in discriminating good from bad binders among these sequences, see S9 Fig. This failure is expected from the poor generalization abilities of RBM-SC for sequences with low counts (Fig 5B).

Table 1. Sequences generated from RBM-SU, log-likelihoods, binding predictions (based on the comparison of the RBM-SU log-likelihood and the threshold in Fig 5D), and results from gel shift assay (B for binders, NB for non binders) and exosite binding assays. For comparison, data for ThA, ThB, ThC and ThD sequences from Ref [41] are shown. ThB and ThC have not been tested for binding with our experimental setup (so NA is used in the corresponding column), although they are expected to bind thrombin given the results obtained in Ref [41].

Label	Sequence	Log-likelihood		Binding Pred.	Binding Result	Exosite	Distance round 8
		RBM-SC	RBM-SU				
r1	AGTGATGATGTGTGGTAGGC	-11.5	-23.4	NB	NB*	NA	0
r2	AGTGTAGGTGTGGATGATGC	-11.4	-24.0	NB	NB	NA	0
r3	TAGGTTTTGGGTAGCGTGGT	-13.0	-22.3	NB	NB	NA	1
r4	AGGGATGATGTGTGGCAGGA	-17.3	-23.6	NB	NB	NA	1
r5	CTAGGACGGGTAGGGCGGTG	-15.9	-21.2	NB	NB	NA	1
r6	AGGGATGTGTGTGGTAGGCT	-14.1	-23.9	NB	NB*	NA	0
r7	AGGGATGCTGCGTGGTAGGC	-10.2	-20.0	B	B	II	0
r8	GAGGGTTGGTGTGGTTGGCA	-10.6	-11.0	B	B	I	0
r9	AGGGTTGGTGTGTGGTTGGC	- 9.8	-11.8	B	B	I	0
r10	ATGGTTGGTTTATGGTTGGC	-15.2	-14.7	B	B	I	1
r11	GAAGGGTGGTCAGGGTGGGA	-16.5	-15.7	B	B	I	2
r12	GGAGGGTGGGTCGGGTGGGA	-15.2	-15.0	B	B	NA	1
r13	GGGGTTGGTACAGGGTTGGC	-16.3	-14.9	B	B	I	2
r14	AGATGGGCAGGTTGGTGCGG	-16.3	-16.3	B	B	I	2
r15	AGATGGGTGGGTAGGGTGGG	-13.9	-14.3	B	B	NA	2
r16	ATAGGGTGGGTGGGTGGGTA	-13.1	-15.0	B	B	NA	1
r17	TGGTGGTTGGGTTGGGTTGG	-12.8	-12.3	B	B	I	1
r18	TGGGATGGGATTGGTAGGCG	-12.2	-20.4	B	NB	NA	0
r19	AGGGTTGGTTATGTGGTTGG	-19.3	-20.0	B	B	I	0
r20	ATTGGTTGGGTAGGGTGGTT	-10.4	-12.2	B	B	I	0
r21	AAACGGTTGGTGAGGTTGGT	-11.2	-12.4	B	B	I	0
r22	CGGGGTGGTGTGGGTGGGAG	-15.1	-14.7	B	B	NA	2
r23	TATTGGTTGGATAGGTTGGT	-13.8	-13.1	B	B	I	1
r24	AGGGTTGGGTGGTTGGATGA	-14.9	-14.1	B	B	I	1
r25	CGGGTTGGGGGGTTGGATTC	-17.0	-15.0	B	B	I	1
r26	CGGTTGGGGGGGTTGGATAC	-18.8	-15.5	B	B	I	1
r27	TGTGGGTTGGTGAGGTAGGT	-18.0	-17.0	B	NB	NA	1
ThA	AGGGATGATGTGTGGTAGGC	-6.0	-19.8	B	B	II	0
ThB	AGGGTAGGTGTGGATGATGC	-5.7	-20.7	NB	NA	II	0
ThC	TAGGTTTTGGGTAGGGTGGT	-6.8	-18.1	B	NA	I	0
ThD	GTAGGATGGGTAGGGTGGTC	-5.7	-13.9	B	B	I	0
p1	AGGGATGATGTGTGGTTGGC	-10.3	-17.1	B	B	I	0
p2	AGGGATGGTGTGTGGTAGGC	- 9.2	-16.2	B	B	II	0
p3	AGGGTTGATGTGTGGTAGGC	- 7.2	-19.1	B	B	II	0
p4	AGGGATGGTGTGTGGTTGGC	- 9.3	-13.1	B	B	I	0
p5	AGGGTTGATGTGTGGTTGGC	-11.1	-16.2	B	B	I	0
p6	AGGGTTGGTGTGTGGTAGGC	- 9.7	-15.2	B	B	II	0

*Aptamers and r1 and r6 did not show thrombin binding gel band, but their pattern indicates a possible weak interaction with thrombin.

https://doi.org/10.1371/journal.pcbi.1010561.t001

Competition assay for exosite binding site and binding strength measurements

Thrombin has two exosites, referred to as I and II, which can be bound by aptamers, *e.g.* ThA is known to bind exosite II, while ThD binds exosite I [41]. We first identify the target exosite

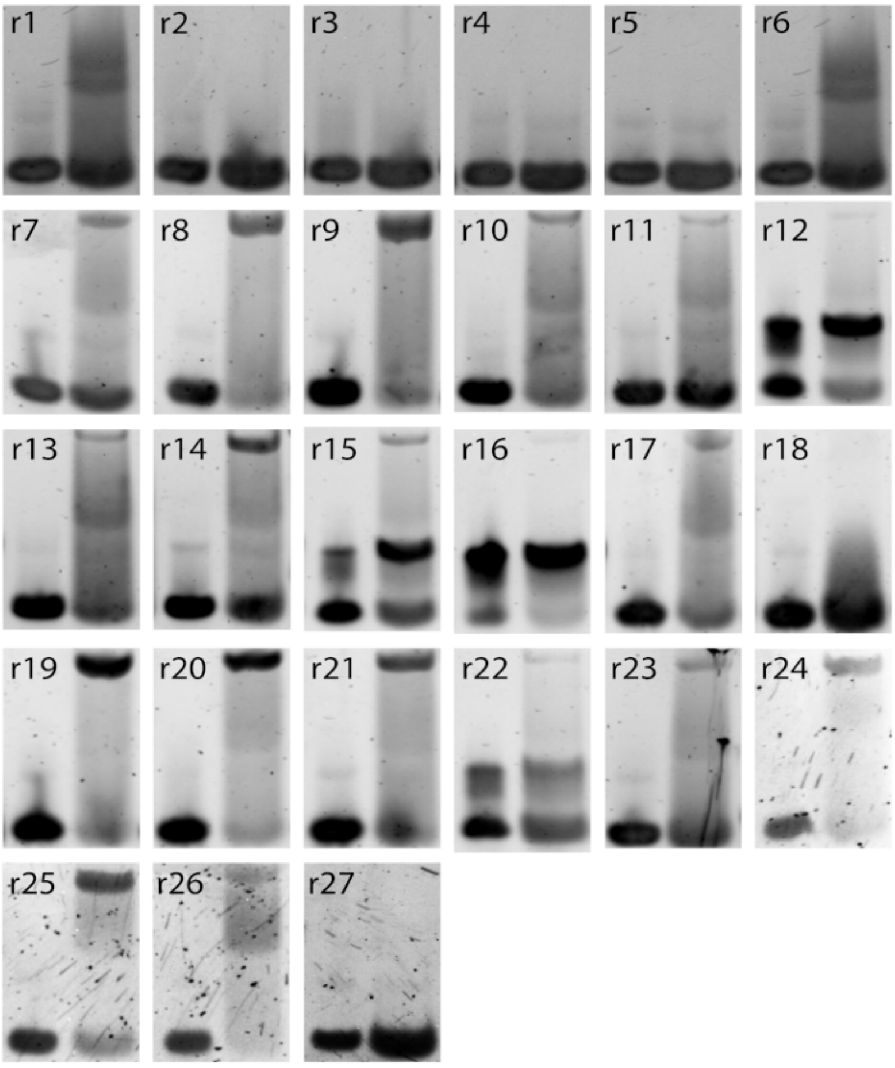

Fig 6. Experimental measurements of binding of respective designed sequences (r1 to r27) to thrombin. 5% native gel assay at 15˚C of stem loops (1–27) alone in the presence of Mg^{2+}/K^{+} (lane 1) and allowed to mix with α-thrombin for 30 minutes at 25˚C on the bench (lane 2). r12, r15, r16, and r22 aptamers were forming dimers with themselves but upon using samples without K^{+}, they were found to bind thrombin. Their entries above display the successful attempt (see Methods for further details). Aptamers r1 and r6 did not show a clear upper band that is indicative of thrombin-aptamer dimer, but the observed smear might indicate weak interactions with the thrombin.

https://doi.org/10.1371/journal.pcbi.1010561.g006

for all the binding aptamers among the r1-r27 by testing each of them (aside from those which were found to form dimer states, see Table 1) against ThA and ThD, see Methods Sec Exosite binding assay. In such a competition assay, the designed aptamers are preincubated with thrombin and are put in competition with a small amount of fluorescently labelled ThD or ThA [41]. If the pre-incubated and fluorescent strand bind the same exosite a thrombin/fluorescent strand complex is observed in the same position as in the thrombin binding assay. However, if the pre-incubated and fluorescent strands target different exosites thrombin is bound twice, causing a downward shift in the observed band (S10 Fig). As shown in Fig 7 and Table 1 we find that all thrombin-binding aptamers among sequences r1-r27 bind exosite I, except one.

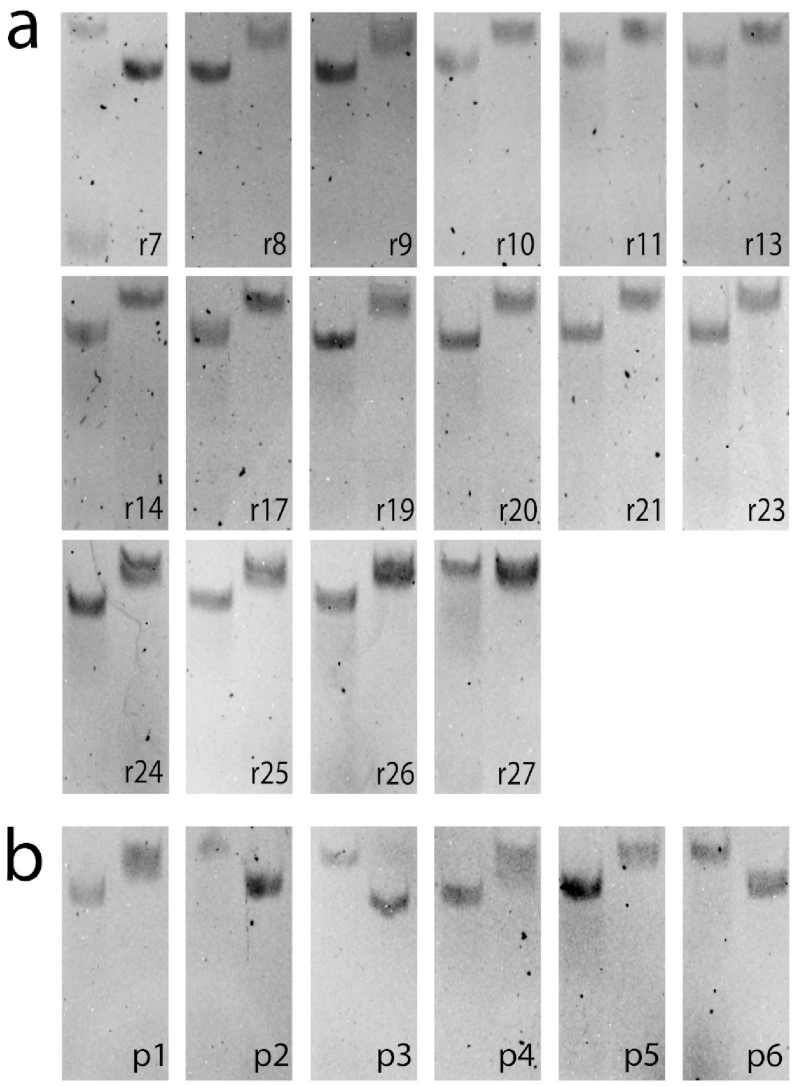

Fig 7. **a**: Binding site assay of all binding sequences and r27 (a nonbinder control) in the RBM generated dataset. **b**: Binding site assay of the 6 sequences that make up the sequence space between ThA and test sequence r9. For all gels, Lane 1 shows addition of ThA and lane 2 shows addition of ThD to the thrombin pre-incubated strand (labeled in black). Results are reported in Table 1.

https://doi.org/10.1371/journal.pcbi.1010561.g007

As we noticed that sequence r9, which is an exosite-I binder, is only 3 mutations away from ThA, which binds exosite II, we decided to test all six intermediate sequences, labelled as p1-p6 in Table 1. One mutation (Adenine vs. Thymine on site 17) seems to control the exosite binding preference along the mutational path, see Table 1 and S11 Fig. Analysis of the RBM-SU weights confirms that position 17 is particularly relevant on the aptamer sequence: many weights have non-zero values on this site (S12 Fig). To understand if the presence of A on site 17 (rarely encountered in round 8 sequences) is sufficient to guarantee binding to exosite II we specifically design four sequences (r24 to r27) with this feature and log-likelihoods above threshold, see Methods Sec Design of single-loop aptamers with RBM and Table 1. As reported above none of these sequence turns out to bind exosite II (while 3 out of 4 bind exosite I), showing that binding specificity is generally controlled by multiple-nucleotide motifs along the sequence.

Next we test if any of the *de novo* generated aptamer sequences with high RBM-SU log-likelihoods are stronger binders than previously identified ThD and ThA aptamers, the binders with the largest number of counts at the end of SELEX [41]. To determine the strongest binder using competition assays, thrombin is mixed with a mixture of the control and the test aptamers at equal ratios, with the control strand being fluorophore labelled (details in S2 Appendix). The stronger binder is considered to be the control or the test aptamer when fluorescence is observed, respectively, in the thrombin-aptamer gel band or in the stem loop band (the unbound aptamer), see S13 and S14 Figs. We observe that none of the designed aptamers binds thrombin more strongly than ThA to exosite II binders, or than ThD to exosite I binders. This result is expected: given the size of the original library ($\sim 10^{15}$) virtually all possible sequences of 20-nucleotide aptamer are initially present, so it is unlikely that SELEX misses stronger binders than ThA and ThD.

We then ask whether the outcomes of competition assays for the best binders could be predicted from the comparisons of their log-likelihoods. RBM-SC-based predictions have 100% success with respect to the above competition assays, always assigning larger scores to ThA and ThD than to the competing aptamers. Conversely, RBM-SU underestimates ThA and ThD binding strength, assigning, in particular, low log-likelihood to ThA and having a global performance of 38% on performance of RBM-generated sequences in the competition assays with ThA and ThD. However, for competitive assays between sequences r1-r27, RBM-SU scores are slightly more predictive than their RBM-SC counterparts, with fractions of successful predictions equal to, respectively, 67% and 59%. Interestingly RBM-SU and RBM-SC also depart from one another in their estimates of the log-likelihoods of exosite I and II binders. We observe in Fig 8 that aptamers binding exosite I have higher scores than their exosite II counterparts, explaining the overwhelming presence of exosite I binders among RBM-SU generated sequences. On the opposite, RBM-SC generally assigns higher log-likelihoods to exosite-II binders. The differences in the behaviours of these models are further examined in Discussion.

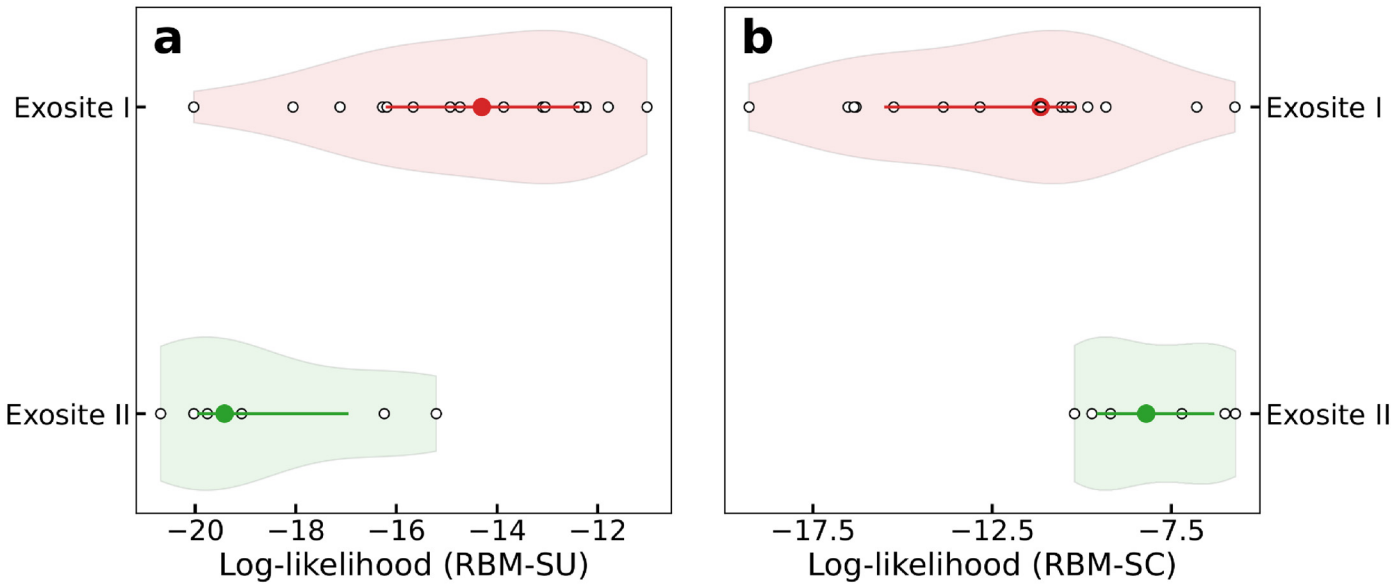

Fig 8. Aptamers binding to exosite I have larger log-likelihoods with RBM-SU, lower log-likelihood with RBM-SC. Violin plots showing the log-likelihoods of exosite I (light orange violin) and exosite II (light green violin) binders. Circles in darker colors denote the average log-likelihood over the class, lines denotes 25- and 75-percentiles, and white points corresponds to the log-likelihoods of the generated sequences. In panel A RBM-SU is used, while in panel B RBM-SC is used.

https://doi.org/10.1371/journal.pcbi.1010561.g008

Supervised learning approach

We also explored supervised learning approaches to train from the aptamer datasets. We considered several DNN architectures (ResNet, Siamese Network and variational autoencoder) as well as traditional methods (random forest and gradient boosted tree) that we trained to classify sequences as binders or non-binders (see S3 Appendix). Training was complicated by the fact that the aptamer dataset only contained positive examples (binders from different selection rounds with their respective counts obtained from the sequencing step). Hence, we either classified sequences with low counts as non-binders, or we generated random sequences not present in the dataset and treated them as non-binders. The first approach achieved between 70% to 84% accuracy on the validation dataset. The second approach had at least 99% accuracy on the validation dataset for all models. However, when evaluating models against the test set (sequences from Table 1), we observed 30% to 74% accuracy for the first approach, and 70% to 89% for the second approach, as the test set is heavily biased to binding sequences, and methods with high accuracy classified most non-binders as false positives. These results indicate that SELEX datasets are challenging for the commonly used supervised learning methods.

Discussion

In this work we proposed data-driven models of aptamer sequences obtained at different stages of directed evolution for thrombin binding. Our models are based on Restricted Boltzmann Machines (RBM), the simplest neural network architecture embedding the notion of representation (or latent factors) of sequence data.

One of our main findings is that the score (log-likelihood) assigned by the model to a sequence $\boldsymbol{s}$ was linearly related to its fitness $F(\boldsymbol{s})$ in the SELEX experiment. More precisely, repeated applications of Eq (1) at previous rounds of selection imply that the likelihood of a sequence $\boldsymbol{s}$ at round r is related to its fitness through

$$p_r(\mathbf{s}) \propto e^{\beta_r F(\mathbf{s})} \quad , \quad \beta_r = \alpha_0 + \alpha_1 + \ldots + \alpha_{r-1} \quad , \tag{3}$$

where α_{k-1} is the intensity of selection from round $k-1$ to k, see Eq (1), and the initial library is assumed to be roughly uniform over the sequence space ($\beta_0 = 0$). This equation can be conveniently rephrased in the language of statistical physics. The rounds of SELEX selection shape a Boltzmann-like distribution over the aptamer sequences, corresponding to an effective energy equal to minus the fitness, $-F$. The effective inverse temperature β_r at round r is the sum of the intensities of selection at the previous rounds, and measures the cumulative effect of these previous selections. As more rounds are carried out, the effective temperature $1/\beta_r$ diminishes, and the distribution of sequences concentrates around the fittest aptamers, *i.e.* the sequences $\boldsymbol{s}$ maximizing $F(\boldsymbol{s})$, see Fig 9. As more and more rounds r of SELEX are applied to the aptamer population the cumulative selection strength β_r seem to saturate, a phenomenon compatible with previous theoretical works [45] and observed in other SELEX experiments [26].

The values of the selection strengths α_r and of the cumulative selection strengths β_r can be extracted from our analysis; for definiteness we arbitrarily choose $\beta_6 = 1$ to fix the scale of the fitness F, as Eqs (2) and (3) are obviously unchanged under the rescaling $\alpha_r, \beta_r \to \lambda\, \alpha_r, \beta_r, F \to F/\lambda$. First, we report in S15 Fig the scatter plots of the log-likelihoods of the sequence data with models trained at different rounds, say, r and r'; the slopes of these scatter plots give access to the ratios $\beta_r/\beta_{r'}$ according to Eq (3). Second the linear fits of the log-likelihood (estimated with the RBM trained on round-6 data) vs. log. enrichment ratios, as well as the Fisher ratios shown in Fig 2C provide estimates of the ratios α_r/β_6. The outcome for α_r, β_r can be found in

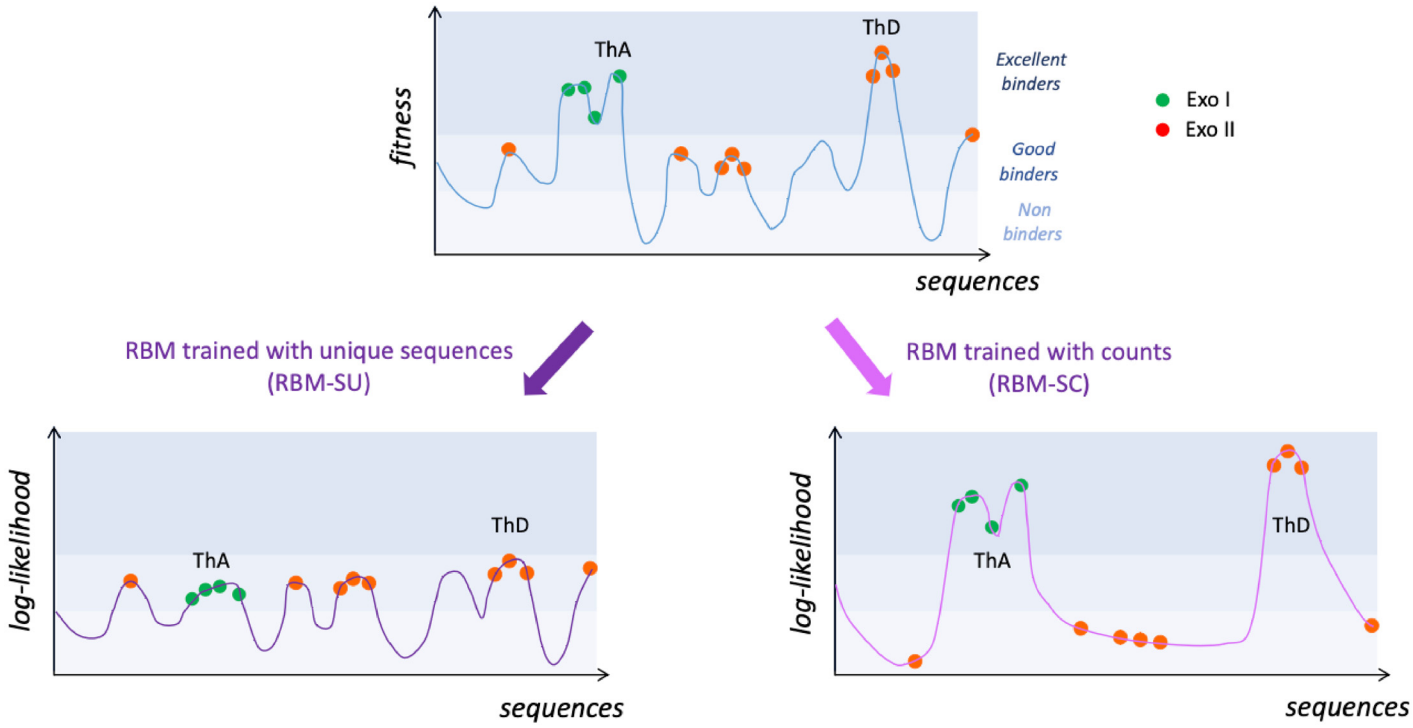

Fig 9. Sketches of the fitness and inferred landscapes. Top: fitness of the aptamer sequences as estimated by the SELEX experiment. After some rounds of selection, most sequences are good binders to thrombin and have low counts (very often, $C = 1$), while some are excellent binders and have large counts. Two excellent binders, ThA and ThD, are schematically shown. Bottom: log-likelihood landscapes defined by the RBM models, trained from unique sequences (RBM-SU, left) or taking into counts (RBM-SC, right). RBM-SU is able to capture the statistical features of the many good binders, but does not reproduce well the few high-fitness peaks. It can be used to generate new sequences (empty peak in the landscape). Conversely, RBM-SC accurately models the high peaks in the fitness landscape, but is unable to reproduce the detailed structure of the landscape at lower levels. It cannot be used to generate new binders.

https://doi.org/10.1371/journal.pcbi.1010561.g009

S2 Table, and allows us to accurately quantify the amount of selection on the aptamers throughout the SELEX process.

The double-loop nature of the aptamer sequences studied here is at the origin of two interesting phenomena. First, we find that $\log p_r(\boldsymbol{s})$ and, consequently, the fitness $F(\boldsymbol{s})$ are, to a very good accuracy, equal to the sum of two contributions coming from the left and from the right loops. This additivity property suggests a mechanistic picture of the binding of aptamers to thrombin. The enrichment factor of the set of molecules carrying the sequence $\boldsymbol{s}$ is proportional to the probability p_{bind} that they bind thrombin and to their amplification factor through PCR. Hence, $\log p_{\text{bind}}$ is proportional to the fitness and additivity of the latter implies that p_{bind} is the product of the binding probabilities of the left and right loops. The two loops of aptamers are thus progressively required, through successive SELEX rounds, to bind the thrombin target. While double-loop aptamers with one binding loop and one parasite subsequence exist in early rounds, they progressively disappear (Fig 2A). The bivalence of aptamers in the final rounds likely reflects the strong selection pressure imposed by SELEX.

The RBM model also allows for identification of the nucleotide motifs in the aptamer sequence that contribute most to the sequence likelihood, or, equivalently, to its fitness. Such motifs are indicative of a G-quadruplex group, a known functional motif in the DNA aptamers that bind thrombin [46]. Other RBM motifs could also allow one to help identify clusters of sequences (subfamilies), investigated in prior works through sequence alignments and manual curation.

A second major finding is that the RBM model is capable of generating new sequences, not present in the dataset, with good binding properties. We have generated 27 aptamer sequences from the RBM that were either predicted to bind or not bind to thrombin. Out of 21 sequences that were thought to be binders, 19 were confirmed to bind thrombin, and all 6 sequences generated as non-binders were rightly predicted so. These non-binder sequences were generated under the non-trivial constraint to differ as little as possible (in terms of mutated nucleotides) from known good binders.

We stress that the capability of RBM models to generate diverse aptamers crucially depends on how they are trained. Standard training, where the counts of sequences are taken into account result in models giving very high scores to the very best binders in the dataset, but unable to generalize beyond these few sequences (Fig 5B). On the contrary, discarding the count information and maximizing the log-likelihood of the set of unique sequences produces models with very good generalization properties, and able to design new and diverse binders, as confirmed in the experiments reported above. The choice of considering unique sequence is partially reminiscent of the reweighting procedure used in sequence-based modeling of proteins [34, 39], and allows the inferred log-likelihoods to reflect more accurately the probabilities for sequences with low number of counts, see Fig 9. Notice that, while unique-sequence-based training could *a priori* be sensitive to sequencing errors we estimate that the probability ϵ of misreading a nucleotide is $< 10^{-3}$ (see Methods Sec Estimation of sequencing error probability and S4 Appendix), in agreement with error rates with next generation sequencing methods [47]. As a result spurious sequences are $< 0.5\%$ of all unique sequences in the dataset, and have only marginal impact on the trained model. However, ensembles in other SELEX experiments using modified bases might experience higher sequencing error rates, which our approach would allow to identify and correct for (Methods Sec Estimation of sequencing error probability).

The properties of the two models are graphically summarized in Fig 9. RBM-SC, which takes into account counts, accurately models the high peaks of the fitness landscape, but discards the smaller peaks. It rightly assigns very high log-likelihoods to the excellent binders, such as ThA or ThD. However, at this level of fitness, the diversity of the sequences that can be generated is very poor. Conversely, RBM-SU, captures the statistical features of sequences at a much lower level of fitness. Many varied sequences can then be generated, the majority of which are good binders. RBM-SU is therefore able to generate more diverse and less strong binders, which makes it particularly appropriate for the design of evolvable aptamers [48]. In principle, RBM-SC inferred from sequences collected in an early round would have had similar properties to RBM-SU inferred from round-8 data. However, in the specific problem of double-loop aptamers we consider here, the presence of a large number of parasite single-loop sequences at the beginning of SELEX evolution could also affect the generative power of models trained at early rounds.

We next used a competitive binding assay both to first classify the binding site of the generated sequences and, in a second step, to assess the strength of binding to a given exosite. We find that the majority of sequences generated with RBM-SU preferentially bind to exosite I. In addition, sequences binding exosite I have on average higher log-likelihoods than the few exosite-II binding sequences. In particular, ThA, an exosite-II binder with a large number of counts in the SELEX experiment is not among the sequences with highest RBM-SU log-likelihoods. Furthermore direct competition experiments between the highest log-likelihood sequences and ThA or ThD (binding exosite I and having a large number of counts) showed that the latter aptamers outperform the former in terms of binding affinity. These apparently paradoxical results can be explained in two ways. First, the log-likelihoods were estimated with the model used for generating sequences, that is, RBM-SU. This model is very good at generate

diverse binders, but is not trained to reproduce counts. The absence of correlation between RBM-SU log-likelihoods and counts or binding affinities is therefore not surprising, whereas RBM-SC high scores show a good correlation with large counts as expected (S16 Fig). Second, these results are compatible with a selection mechanism involving binding to the two sites of thrombin. Binding to exosite II has been shown to facilitate binding to exosite I, presumably through allosteric structural change [41]. Due to this allostery mechanism, when exosite II is loaded (even with a different molecule), hairpin with a low-affinity loop to exosite I could be selected. This mechanism could produce a rather subtle parasitism, where only the best exosite II binders in a quasi-monoclonal population (few sequences with largest counts) are under strong selection, and allow for the presence of a more diverse exosite-I binder population. Further experimental investigations combined with theoretical analysis, *e.g.* using concepts developed in ecosystems dynamics in presence of parasite populations, could help to further investigate the selection dynamics.

We note that our RBM represents a higher level of complexity than the direct contact analysis methods (DCA) that have also been recently applied to protein ensemble selection experiments [49]. While the DCA method trained using the pseudo-likelihood method was not able to correctly predict binders and non-binders for our dataset, when we used contrastive divergence training for DCA, the assigned scores from the trained DCA model showed correlation with our trained RBM (see S5 Appendix). As opposed to DCA, which infers pairwise interactions, the RBM model's hidden units can be used for clustering of sequences or identification of multi-nucleotide motifs, such as G-quadruplexes, making them more readily interpretable. We have explored using supervised learning models, including DNNs, on our datasets predicting binders and non-binders, but as further detailed in S3 Appendix, we did not obtain good prediction accuracy for the outcomes of our experiments with designed sequences.

Conclusion

In this work, we presented an unsupervised learning approach for modeling sequence ensembles obtained from selection experiments based on Restricted Boltzmann Machines (RBM). The approach was applied to previously obtained data from SELEX experiment to find thrombin bivalent aptamers nanostructures that bind two different exosites. More precisely, our approach consisted of the following steps: 1) developing a method that estimated sequencing error rates, which could be used for curation of the sequence data, 2) showing that the log-likelihood of the trained RBM accurately predicted aptamer fitness in terms of its propensity to be enriched in later rounds of the experimental selection protocol, 3) using RBM to identify contributions of the two aptamer loops to exosite binding, 4) showing that inspection of the parameters of the trained RBM identified functional features (such as G-quadruplex) of the selected sequences, 5) using the trained model to generate novel sequences, whose ability to bind thrombin was verified experimentally, and 6) comparing RBMs with different supervised learning models trained on the same dataset, with the result that RBM generalized better.

We emphasize that the calculation of log-likelihood and hence of the fitness of any designed sequence by RBMs is very efficient, making them faster than other approaches based on e.g. docking or free-energy estimation from molecular simulation. Furthermore, the structure of the model allows us to capture and identify complex features that could include covarying residues or motifs. We showed that RBM training can be flexibly adapted depending on the scope, e.g. taking into account sequence counts or not allows one to design stronger or more diverse binders. We anticipate that RBMs will be also useful for the modeling of other aptamer datasets with more complex selection protocol, such as competition assays where aptamers are selected to bind to a desired target, e.g. cancerous tissues, and at the same time not to bind to the

control, e.g. healthy tissue. We believe our approach has the potential to generate alternative or better binders for these complex targets, as well as to unveil the sequence motifs that are enriched or avoided in these high-quality aptamers. The same approach can be also useful to model RNA and DNA regulatory sequences and their interaction with proteins in the key processes such as transcription regulation [31, 32, 38, 50]. Lastly, our modeling and design methods are also readily applicable to other selection-amplification protocols, such as phage display for antibody discovery [51, 52] or directed protein evolution studies [49, 53], which have much larger space of possible sequences (20^L for length L) compared to aptamers (4^L).

Methods

Estimation of sequencing error probability

Sequencing errors are potentially harmful, as they could lead to more unique sequences in the dataset and possible biases in the RBM models. We introduce an inference approach to estimate the sequencing error rate, based on the presence of spurious single-site mutations of sequences with high number of counts. In practice the method consists in selecting a subset of sequences with high number of counts, referred to as "peak" sequences, and in comparing the expected number of sequences one mutation away from these peaks due to sequencing errors to the actual number in the data. Our analysis, detailed in S4 Appendix, indicates that the error rate (per nucleotide) is smaller than $\epsilon^* \sim 10^{-3}$.

We use this bound to estimate the expected number of spurious sequences present in the dataset. We obtain $N_{\text{spurious}} \sim 1000$ unique sequences (see S4 Appendix), corresponding to $\sim 0.5\%$ of the total number of unique sequences present in the data.

Restricted Boltzmann Machine: Definition, training, sampling

The probability of a visible and hidden units state in an RBM model is defined by

$$p(\boldsymbol{s}, \boldsymbol{h}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{L} g_i(s_i) - \sum_{\mu=1}^{M}\mathcal{U}_\mu(h_\mu) + \sum_{\mu,i} h_\mu w_{\mu i}(s_i)\right), \tag{4}$$

where Z is the normalization, g_i, and $w_{\mu i}$ are parameters to be inferred from the data during training, and

$$\mathcal{U}_\mu(h) = \frac{1}{2}\gamma_{\mu+}(h_+)^2 + \frac{1}{2}\gamma_{\mu-}(h_-)^2 + \theta_{\mu+}h_+ + \theta_{\mu-}h_-, \tag{5}$$

where $h_+ = \max(h, 0)$, $h_- = \min(h, 0)$ and $\gamma_{\mu+}$, $\gamma_{\mu-}$, $\theta_{\mu+}$, $\theta_{\mu-}$ are again model parameters to be inferred from the data during training. This specific form of the function $\mathcal{U}_\mu$, which is called "double Rectified Linear Unit" combines the usage of a relatively low number of parameters with the possibility of learning high-order correlations in the data [39]. An advantage of choosing Double ReLU potentials is that the likelihood $\log p(\boldsymbol{s})$ of a sequence $\boldsymbol{s}$, obtained by marginalizing $p(\boldsymbol{s}, \boldsymbol{h})$ over $\boldsymbol{h}$, has an explicit analytical expression in terms of error functions.

It has been suggested that, for RBMs, sparsity of the weight parameters, together with a high number of hidden units, can improve the generative properties of the machine and its interpretability [39, 54]. To prevent the model from overfitting, we hence enforce sparsity of weights and we empirically set M to value above which the model's log-likelihood on validation dataset does not further increase. We resort to a L_1^2 regularization scheme, which consists

in adding to the log-likelihood of the data, $\mathcal{L}$ in Eq (1), a term of the form [39, 40]

$$-\lambda \sum_{\mu=1}^{M} \left(\sum_{i=1}^{L} \sum_{s_i} |w_{\mu i}(s_i)| \right)^2 , \tag{6}$$

hence enhancing sparsity homogeneously across hidden units. The value of the hyperparameter λ must be, in general, chosen carefully to balance model interpretability (obtained for sparse weights, *i.e*. large λ) and expressivity (to learn data features). We observed little effects of changes in hyperparameters (see also S17 Fig), provided that they are not too different from the one given in S1 Appendix. This is also the case for the number M of hidden units chosen: we used $M \simeq 70$ for RBMs with $L = 20$ visible units, and $M \simeq 90$ for $L = 40$. Precise values of M are given, for each RBM used, in S1 Appendix, but we noticed that using different numbers have little effects on the results discussed in this work (see also S17 Fig).

Once the parameters in Eq (4) are obtained, we can sample from the marginal distribution $p(\boldsymbol{s})$ to generate new sequences. Sampling can be done in several ways [55]. Here we use alternate Gibbs sampling (AGS), which consists in sampling the RBM's visible units while keeping the hidden units fixed and vice-versa, in an alternate manner, until the Monte Carlo Markov Chain equilibrates. To increase the probability of sampling high log-likelihood sequences we can sample from $p(\boldsymbol{s})^2$ instead of $p(\boldsymbol{s})$ using the so-called duplication trick [39]. We write

$$p(\boldsymbol{s})^2 = \left(\int d\boldsymbol{h}\, p(\boldsymbol{s}, \boldsymbol{h}) \right)^2 = \int d\boldsymbol{h}_1 \int d\boldsymbol{h}_2\, p(\boldsymbol{s}, \boldsymbol{h}_1)\, p(\boldsymbol{s}, \boldsymbol{h}_2) \ . \tag{7}$$

This squared likelihood distribution can therefore be sampled with standard AGS after duplication of the hidden layer of the trained RBM model.

The average hidden unit μ's activity for a given sequence $\mathbf{s}$ is defined as $\langle h_\mu \rangle = \int d\mathbf{h}\, h_\mu\, p(\mathbf{h}|\mathbf{s})$. Note that $\langle h_\mu \rangle$ only depends on the sequence $\mathbf{s}$ through the input $I_\mu = \sum_i w_{\mu i}(\mathbf{s}_i)$. When the average activity is close to 0, the corresponding hidden unit has vanishing contribution to the sequence log-likelihood, while for both large negative or positive values of average activity the contribution of the hidden unit to the log-likelihood is positive. Therefore the sign of the weights $w_{i\mu}$ assigned to a particular sequence motif is not indicative itself of the presence or absence of a given pattern, as the contribution in $p(\mathbf{h}|\mathbf{s})$ depends on the product $h_\mu I_\mu$ and can only be null or positive.

Design of single-loop aptamers with RBM

The RBM-SU distribution $p(\boldsymbol{s})$ can be sampled to generate sequences $\boldsymbol{s}$ of interest, and test the validity of the model. We describe below how we generated sequences in Table 1.

Determination of threshold. We fix the threshold, which allows us to distinguish good from bad binders based on their log-likelihoods to minimize the number of misclassified sequences among the preliminary set of sequences given in S1 Table. As a range of possible values are possible, we actually take the median of this interval.

Sequences with high likelihoods. We first sample through AGS (see Sec Restricted Boltzmann Machine: Definition, training, sampling) 4000 sequences from $p(\boldsymbol{s})$ and from $p(\boldsymbol{s})^2$. We then choose 10 among these sequences (named r9 to r17 and r22, r23 in Table 1), which have both high log-likelihood and large distances (numbers of different nucleotides) to round 8 data. In practice these sequences are at Hamming distance 1 or 2 from the closest sequences in the original dataset, since further away sequences have substantially lower log-likelihoods. All 10 generated sequences are experimentally confirmed to be good binders (Table 1), and are indicated as green lines in Fig 5E.

Sequences with critical mutations for binding/non-binding status. We next use our RBM to predict critical mutations capable of changing the binder/non-binder status of aptamers. First we exhaustively look for the smallest possible number of mutations leading to a substantial decrease of the log-likelihood of known good binders. In particular, sequence r1 has 1 mutation with respect to a control sequence that we tested for binding (named d10 in S1 Table), r2 and r3 are both 1 mutation away from, respectively ThB and ThC, both identified as good binders in Ref [41]. All these generated sequences are confirmed to be unable to stably bind thrombin after this single-point mutation (Table 1 and Fig 6) and they correspond to red vertical lines to the left of the threshold in Fig 5E. All these mutations removed a G from the sequence, and G nucleotides are necessary to form G-quadruplex motifs, known to be important for thrombin aptamers. To show that our RBM can also identify other positions in the aptamer that are key to thrombin binding, we also design two more sequences, r4 and r5, which have 2 mutations with respect to aptamers found in the SELEX dataset and validated as good binders (respectively, d10 and d18, see S1 Table). The mutations are again chosen so that the log-likelihood is decreased as much as possible, but without removing G nucleotides from the original sequences. We find the sequences lost their ability to bind thrombin after the 2 mutations, as predicted by the RBM (Fig 6), so they correspond to two vertical red lines to the left of the threshold in Fig 5E.

Sequences in dataset with mismatches between counts and log-likelihoods. We further test the performance of the RBM model by searching for sequences with (1) relatively low log-likelihoods but with large numbers of counts (139 or more, see S3 Table) in the SELEX experimental data from Ref [41], of for sequences with high log-likelihood but with few counts (11 or less, see S3 Table). The sequences chosen in case (1) are r6, r7, r18, r19 (see Table 1); one of them (r6) is below, and the other 3 are slightly above the identified log-likelihood threshold. Sequences chosen in case (2) are r8, r9, r20, r21 (Table 1). The RBM predictions are confirmed in all cases but one (r18), which corresponds to a red vertical line at the right of the threshold in Fig 5E.

Sequences sharing a rare mutation with ThA, a strong exosite-II binder. Last of all we design *de novo* sequences (r24 to r27 in Table 1) under the following two-fold criterion. First these sequences are required to have Adenine in position 17, which is uncommon in the training dataset (A is the second least common nucleotide in that position, being present in about 13% of the sequences in round 8; it is found in ThA, which strongly binds exosite II). Second, the sequences are required to have large log-likelihoods, exceeding the threshold value. Remarkably, the only non-binder among r24-r27 is the one with lowest log-likelihood, r27. However, while mutating away from A in ThA change the binding specificity from exosite II to I (Fig 8) sequences r24 to r27 are all exosite-I binders, showing that the presence of A17 is not sufficient for exosite-II specificity.

Thrombin binding assay

All RBM designed sequences were first assessed for their ability to bind either of the cationic exosites of human alpha-thrombin. Each sequence was placed as the loop of a 18 bp stem loop with the full sequences reported in the S4 Table. As done previously [41], we used a 5% native gel shift assay to qualitatively assess the binding of each stem loop to thrombin. Each sequence was tested with two gel lanes, the first lane always corresponding to the stem loop without thrombin and the second lane consisting of equimolar amounts (500 nM) of thrombin and the stem loop. The presence of an upper band, consisting of a stem loop bound to thrombin complex, in the second lane indicates a binding sequence. Sequences without the upper band (non-binding sequences) either very weakly interact with thrombin, characterized by a smear but no

band in the second lane, or do not interact with thrombin at all matching their negative control lane. Sequences ThA and ThD were selected from the previous study as positive controls for their high affinity for thrombin and known binding sites [41]. Results for all RBM generated sequences are shown in Fig 6 and summarized in Table 1. Results for all DCA generated sequences are shown in S18 Fig and summarized in S1 Table. To quantify the interaction of the stem loop and thrombin, we tested both control sequences independently and together in varying concentrations of thrombin (S10 Fig). The results clearly indicate the stem loop/thrombin band occurs from a 1:1 interaction of thrombin and each stem loop, and the simultaneous binding of two stem loops on opposite exosites of thrombin downshifts the stem loops/thrombin band from the singular case.

A secondary band prominently appeared among four of the sequences during the binding assay, (r12, r15, r16, and r22). These sequences showed no binding to Thrombin at first. Upon further investigation, the secondary band was found to most likely be a dimer state of the DNA loop from interaction of the G-quartet motifs. The four sequences have a higher G-content than all other RBM-generated sequences. Additionally, a G-quadruplex dimer would require K^+ cations to form, indicating a testable transition from the single loop to dimer state. The sequences' Thrombin binding ability was re-assessed by the same experiment, with two small changes. The first was remaking the DNA samples without K^+ in their buffer, so their transition from single stem loop to a dimer state could be observed [56]. The second change was the heating the DNA samples to 90˚C for 5 minutes before immediately chilling them in ice. Samples (r12, r15, r16, r22) in Fig 6 show the results of this final experiment, with all dimer-susceptible sequences showing an ability to bind Thrombin. Accordingly, we classify these sequences as binders and suggest their absence from the original dataset is due to G-quadruplex dimer formation during the original SELEX procedure. A clear shift from the monomer state in 1x TAE Mg^{2+} (no K^+) buffer (lane 1) to the dimer state upon addition of buffer with K+ (lane 2) is also observed for all dimer-susceptible sequences. Note this transition still contains some fraction of the dimer state in lane 1 where the sample contains no K+. This is due to presence of K+ in the gel matrix itself as well as the running buffer.

Exosite binding assay

RBM-generated sequences that were able to bind to thrombin were tested to determine which exosite (I or II) of thrombin they bind to. Each aptamer was pre-incubated with thrombin for 30 minutes at 25˚Cat an equimolar ratio in two separate samples. Small amounts (1/10th the pre-incubated strand) of fluorescent labeled exosite II binder ThA [41] was added to the first sample and fluorescent labeled exosite I binder ThD to the second. Using the same strategy as our thrombin binding assay, our samples were run in a 5% native gel with 5 mM K^+ for proper DNA/thrombin binding. If the pre-incubated strand bound the same exosite as the fluorescent strand, the thrombin/fluorescent strand complex band would be observed in the same position as seen in our thrombin binding assay. However, if the pre-incubated strand bound the opposite exosite as the fluorescent strand, both strands bind thrombin causing the same downward shift as observed for our exosite verified control strands mixing (S10 Fig). Accordingly, sequences with no binding affinity to thrombin matched control samples with no test strand. By comparing the outcome of both lanes for a sample we are able to firmly assign the binding site of our test sequences. The gel results are shown in Fig 7 and summarized in Table 1.

Supporting information

S1 Fig. Panel A: probability density function of the counts observed for the double aptamers in each round. Notice the log scale on the y axis. Panel B: for each pair of consecutive rounds,

we plot here the logarithm of the ratio of counts of the sequences present in both rounds (left) and the corresponding histogram (right), against the log-likelihood of the sequence computed with the RBM-DC model.
(PDF)

S2 Fig. Evolution of counts of the 10 left (panel A) and right (panel B) aptamers with largest number of counts at round 8. Counts have been re-scaled by a factor so that the total number of counts in each round is constant.
(PDF)

S3 Fig. Log-likelihood computed with the RBM-SC model and with the RBM-LC model (trained on left single-loop sequences at round 8, see S1 Appendix) in panel A or RBM-RC model (trained on right single-loop sequences at round 8, see S1 Appendix) in panel B for the single-loop sequences observed at round 8. The slope and the R^2 values of the linear fit are respectively 0.96 and 0.98 for panel A, and 1.05 and 0.97 for panel B. Panel C: log-likelihood computed with the RBM-DC model for the double-loop sequences observed at round 5, compared with the sum of the log-likelihood obtained by using RBM-LC to score the left loop and RBM-RC to score the right loop. The slope and the R^2 value of the linear fit are, respectively, 0.99 and 0.99.
(PDF)

S4 Fig. Relationship between log-enrichment and log-likelihoods of single-loop aptamers. Panels A, C show the histograms of log-likelihoods at each round, as computed by RBM-SU6 (panel A) and RBM-SC6 (panel C). Panels B, D show the scatter plot of log-enrichment of each bin in the left panels, and the corresponding log-likelihood. In the inset, the slope of each linear fit appearing in the main plot is compared with the same quantity estimated as a Fisher's ratio (see Sec RBM's log-likelihood is an accurate predictor of the aptamer's fitness). The dashed black line is the $x = y$ line.
(PDF)

S5 Fig. Left side: histograms of log-likelihoods of left (blue) and right (orange) loops computed with RBM-SU (panel A) or RBM-SC (panel C) for sequences observed in round 8 (unique in panel A, with their counts in panel B), together with that of $5 \cdot 10^5$ random uniform sequences (light green); the black line is the 99-quantile of the light green histogram, and parasite sequences are defined as those which have lower log-likelihood than the black line, while at the same time the other loop of the 40-nt aptamer has log-likelihood larger than the threshold. Right side: log-likelihood of the RBM trained after excluding parasite sequences at round 8 (RBM-NPU for panel B, RBM-NPC for panel D) versus that of the RBM-SU (panel B) or RBM-SC (panel D) model. A linear fit for the points at the right-hand side of the black line (which is the same of panels a for panel B, and of panel C for panel D) gives a slope of 1.0 and a R^2 of 0.92 for panel B, and a slope of 1.0 and a R^2 of 0.96 for panel D. For points at the left-hand side of the black line the slope is 2.6 with an R^2 of 0.79 for panel B, and the slope is 2.0 with an R^2 of 0.33 for panel D.
(PNG)

S6 Fig. Panel A: Frobenius norms of the weights for RBM-DC. The logos corresponding to the 3 weights with largest Frobenius norm are given in Fig 4A–4C. Panel B: Frobenius norms of the weights for RBM-SC. The logos corresponding to the weight with the 2nd largest Frobenius norm and the one with the 7th largest Frobenius norm are given in Fig 4E and 4F.
(PDF)

S7 Fig. Panel A: Frobenius norms obtained for each weight of RBM-DC computed using only the first 20 visible units (L-norm in the x axis) or the last 20 visible units (R-norm in the y axis). Panel B: RBM-DCL and RBM-DCR are two RBMs with 20 visible units used to score left and right loops. RBM-DCL (RBM-DCR) is obtained from RBM-DC by using only its first (last) 20 visible units and their fields, and the hidden units with L-norm > R-norm (R-norm > L-norm) with their potentials, ignoring their interactions with the last (first) 20 visible units. In this panel, we compare, for each unique double-loop sequence observed at round 5, the log-likelihood of the RBM-DC model with the sum of the log-likelihoods obtained by using RBM-DCL to score the left loop and RBM-DCR to score the right loop. The slope of the linear fit is 0.99 and the R^2 score is >0.99.
(PDF)

S8 Fig. Local field learned by the RBM-DC6 used in Fig 2 (panel A), compared with the conservation logo of the full dataset at round 6 (panel B).
(PDF)

S9 Fig. Panel A: Histogram of the log-likelihoods of all unique aptamers observed in the last round (blue line) and of uniformly random sequences (orange line), computed with RBM-SC trained on single-loop sequences from round 8, keeping information about the counts. Inset: AUC computed on the sequences generated by the RBM-SU model (panel C). Panel B: Vertical lines locate the log-likelihoods of sequences experimentally validated to be binders (green) or non binders (red). Sequences taken from a preliminary set described in S1 Table. Results allows us to determine the binding/non binding threshold, shown with the black dashed line. Panel C: same as panel B for sequences designed with the RBM-SU model, as described in Sec RBM trained from unique sequences generate diverse aptamers capable of binding thrombin (see Table 1).
(PDF)

S10 Fig. Panel A: lane 1 contains 5' 6FAM labeled control sequence ThA, lane 2 contains 5' 6FAM labeled control sequence ThD. Panel B: ThA mixed in varying ratios with Thrombin (1:0.32, 1:0.64, 1:1.08). Panel C: ThD mixed in varying ratios with Thrombin (1:0.32, 1:0.64, 1:1.08). Panel D: ThA + ThD mixed in varying ratios with Thrombin (1:1:0.32, 1:1:0.64, 1:1:1.08).
(PNG)

S11 Fig. RBM-SU (panel A) or RBM-SC (panel B) log-likelihood versus distance from ThA for sequences p1 to p6 in Table 1. Different mutations are represented with different line styles: dotted lines for mutations involving position 5 (mutating A into T when going from ThA to r9), dashed lines for mutations involving position 8 (mutating A into G when going from ThA to r9), and solid lines for mutations involving position 17 (mutating A into T when going from ThA to r9).
(PDF)

S12 Fig. Panel A: fields of RBM-SU. The largest field (in norm) corresponds to position 17 (gray box), which is the one that in S11 Fig determines the binding exosite. Panel B: sum of the norms of each weight of RBM-SU, at fixed sequence position. The largest sum corresponds again to position 17 (gray box).
(PDF)

S13 Fig. One sided competition assays of all exosite-I binders vs. a different fluorophore labeled strand in each well, r8, r14, and r19 respectively (panel A). Numbers to the left of each trial indicate the identity of the non-labeled strand. Additionally exosite-II binders were tested

against fluorophore labeled ThA with negative control labeled ThA (panel B) and select exosite-I binders were tested against ThD with negative control labeled ThD (panel C).
(PNG)

S14 Fig. Competition assay of r8F vs r14 and r14F vs r8 (panel A). r8F vs r19 and r19F vs r8 (panel B), and r19F vs r14 and r14F vs r19 (panel C). The F suffix indicates the strand is fluorophore labeled with a 5' 6FAM modification.
(PNG)

S15 Fig. Comparison of the log-likelihoods computed with RBM-SC trained at different rounds (named RBM-SC5, RBM-SC6, RBM-SC7 and RBM-SC8 if trained respectively on sequences observed in round 5, 6, 7, 8). Plots on the diagonal are the distribution of the log-likelihoods of each RBM. The sequences used to prepare each histogram are the full set of sequences observed in round 5, 6, 7, or 8 (discarding counts). In each-non diagonal plot, the slope m and the coefficient of determination r^2 for the linear fit are given.
(PDF)

S16 Fig. Panel A: Log-likelihoods (computed with RBM-SU) versus log number of counts for the unique single-loop sequences observed at round 8. ThA (counts: 10132, log-likelihood: -19.8) and ThD (counts: 8853, log-likelihood: -13.9) are highlighted with circles. Panels B, C: Log-likelihoods computed with RBM-SC (for panel B) or RBM-SU (for panel C) versus log number of counts for the 1000 unique single-loop sequences observed at round 8 with highest number of counts.
(PDF)

S17 Fig. Average log-likelihoods computed with 16 RBMs trained with different choices of hidden unit numbers and weight regularization on the single aptamers obtained from the 8-th round. The scale on the y-axis is kept constant across the different sub-plots to highlight how the difference in average log-likelihoods are much smaller than the difference between the log-likelihood of training (and test) data and that of random sequences. The green circle at 0.001 regularization strength correspond to the RBM used in the paper (RBM-SU).
(PDF)

S18 Fig. Thrombin binding assay of DCA generated sequences. Lane 1 has the stem loop alone, whereas lane 2 has the same stem loop exposed to thrombin. Binding sequences are indicated by a high visible band in lane 2.
(PNG)

S1 Table. Result of thrombin binding assays with all DCA-generated sequences and sequences of exosite I control d18 and exosite II control d10. B indicates a binder while NB indicates a nonbinder.
(PDF)

S2 Table. Ratios of the values β_r (with $r = 5, 6, 7, 8$), see Eq (3) can be estimated from the slopes given in S15 Fig. Here we fixed the fitness scale so that $\beta_6 = 1$, or, equivalently, each value in this table is given in units of β_6. Coefficients α_r (with $r = 5, 6, 7$) can be obtained: (i) as the slopes in Fig 2C (first column); (ii) using the Fisher's ratio, see inset of Fig 2C (second column); (iii) from the slopes in S15 Fig, since $\beta_{r+1} - \beta_r = \alpha_r$, see Eq (3). Each of these methods has different noise sources, but the values obtained are in quite good agreement.
(PDF)

S3 Table. For each sequence generated from RBM-SU trained on unique loop sequences observed in the last round, we provide here the distance from the closest single-loop

aptamer observed at round 8 (column Dist1, 382094 sequences) and the number of counts of each sequence at round 8. Since a good binder is expected to be found close to a sequence with many counts, we also provide in the other columns (Dist3, Dist10, Dist100) the distance to the closest single-loop aptamer with at least, respectively, 3, 10 or 100 counts in round 8 (respectively 74785, 22332, and 1177 sequences).
(PDF)

S4 Table. The full sequences from all experiments carried out in this work, with their loop region underlined for easy identification. r1–27 correspond to sequences generated from sampling our RBM. All sequences with p labels (p1-p6) are along the mutation pathway from sequence ThA to r9. Sequences d1-d9 and d11-d17 are were generated from sampling from the DCA parameters. Sequences d10, d18, ThA, and ThD were used as controls throughout.
(PDF)

S1 Appendix. Details of RBMs' training.
(PDF)

S2 Appendix. Detailed experimental methods.
(PDF)

S3 Appendix. Comparison with DNN and Traditional Machine Learning approaches.
(PDF)

S4 Appendix. Inference of sequencing error probability.
(PDF)

S5 Appendix. Comparison with Direct Coupling Analysis.
(PDF)

Acknowledgments

We thank G. Mayer and M. Famulok for helpful discussions and comments. A.D.G. would like to thank G. Isacchini and J. Fernandez-de-Cossio-Diaz for helpful discussions and comments. We also acknowledge Research Computing at Arizona State University for providing HPC resources that have contributed to the research results reported within this paper.

Author Contributions

Conceptualization: Rémi Monasson, Simona Cocco, Petr Šulc.

Data curation: Andrea Di Gioacchino, Jonah Procyk, Marco Molari.

Formal analysis: Andrea Di Gioacchino, Jonah Procyk, Marco Molari, John S. Schreck, Rémi Monasson, Simona Cocco, Petr Šulc.

Funding acquisition: Rémi Monasson, Simona Cocco, Petr Šulc.

Investigation: Andrea Di Gioacchino, Jonah Procyk, Rémi Monasson, Simona Cocco, Petr Šulc.

Methodology: Andrea Di Gioacchino, Jonah Procyk, Marco Molari, John S. Schreck, Yu Zhou, Yan Liu, Rémi Monasson, Simona Cocco, Petr Šulc.

Project administration: Rémi Monasson, Simona Cocco, Petr Šulc.

Resources: Rémi Monasson, Simona Cocco, Petr Šulc.

Software: Andrea Di Gioacchino, Jonah Procyk, Marco Molari.

Supervision: Rémi Monasson, Simona Cocco, Petr Šulc.

Validation: Andrea Di Gioacchino, Jonah Procyk.

Visualization: Andrea Di Gioacchino, Jonah Procyk, Marco Molari.

Writing – original draft: Andrea Di Gioacchino, Jonah Procyk, Marco Molari, Rémi Monasson, Simona Cocco, Petr Šulc.

Writing – review & editing: Andrea Di Gioacchino, Jonah Procyk, Rémi Monasson, Simona Cocco, Petr Šulc.

References

1. Mayer G, Lohberger A, Butzen S, Pofahl M, Blind M, Heckel A. From selection to caged aptamers: identification of light-dependent ssDNA aptamers targeting cytohesin. Bioorganic & medicinal chemistry letters. 2009; 19(23):6561–6564. https://doi.org/10.1016/j.bmcl.2009.10.032 PMID: 19854646
2. Lennarz S, Alich TC, Kelly T, Blind M, Beck H, Mayer G. Selective Aptamer-Based Control of Intraneuronal Signaling. Angewandte Chemie. 2015; 127(18):5459–5463. https://doi.org/10.1002/anie.201409597 PMID: 25754968
3. Schüller A, Matzner D, Lünse CE, Wittmann V, Schumacher C, Unsleber S, et al. Activation of the glmS ribozyme confers bacterial growth inhibition. Chembiochem. 2017; 18(5):435–440. https://doi.org/10.1002/cbic.201600491 PMID: 28012261
4. Schmitz A, Weber A, Bayin M, Breuers S, Fieberg V, Famulok M, et al. A SARS-CoV-2 Spike Binding DNA Aptamer that Inhibits Pseudovirus Infection by an RBD-Independent Mechanism. Angewandte Chemie International Edition. 2021; 60(18):10279–10285. https://doi.org/10.1002/anie.202100316 PMID: 33683787
5. Rosenthal M, Pfeiffer F, Mayer G. A Receptor-Guided Design Strategy for Ligand Identification. Angewandte Chemie International Edition. 2019; 58(31):10752–10755. https://doi.org/10.1002/anie.201903479 PMID: 31050104
6. Ortega AD, Takhaveev V, Vedelaar SR, Long Y, Mestre-Farràs N, Incarnato D, et al. A synthetic RNA-based biosensor for fructose-1, 6-bisphosphate that reports glycolytic flux. Cell Chemical Biology. 2021;. https://doi.org/10.1016/j.chembiol.2021.04.006 PMID: 33915105
7. Renzl C, Kakoti A, Mayer G. Aptamer-Mediated Reversible Transactivation of Gene Expression by Light. Angewandte Chemie. 2020; 132(50):22600–22604. https://doi.org/10.1002/ange.202009240
8. Domenyuk V, Gatalica Z, Santhanam R, Wei X, Stark A, Kennedy P, et al. Poly-ligand profiling differentiates trastuzumab-treated breast cancer patients according to their outcomes. Nature communications. 2018; 9(1):1–9. https://doi.org/10.1038/s41467-018-03631-z PMID: 29572535
9. Civit L, Taghdisi SM, Jonczyk A, Haßel SK, Gröber C, Blank M, et al. Systematic evaluation of cell-SELEX enriched aptamers binding to breast cancer cells. Biochimie. 2018; 145:53–62. https://doi.org/10.1016/j.biochi.2017.10.007 PMID: 29054799
10. Hornung T, O'Neill HA, Logie SC, Fowler KM, Duncan JE, Rosenow M, et al. ADAPT identifies an ESCRT complex composition that discriminates VCaP from LNCaP prostate cancer cell exosomes. Nucleic acids research. 2020; 48(8):4013–4027. https://doi.org/10.1093/nar/gkaa034 PMID: 31989173
11. Zhou J, Rossi JJ. Cell-type-specific, aptamer-functionalized agents for targeted disease therapy. Molecular Therapy-Nucleic Acids. 2014; 3:e169. https://doi.org/10.1038/mtna.2014.21 PMID: 24936916
12. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990; 249(4968):505–510. https://doi.org/10.1126/science.2200121 PMID: 2200121
13. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. Nature. 1990; 346(6287):818–822. https://doi.org/10.1038/346818a0 PMID: 1697402
14. Sola M, Menon AP, Moreno B, Meraviglia-Crivelli D, Soldevilla MM, Cartón-García F, et al. Aptamers against live targets: is in vivo SELEX finally coming to the edge? Molecular Therapy-Nucleic Acids. 2020; 21:192–204. https://doi.org/10.1016/j.omtn.2020.05.025 PMID: 32585627
15. Proske D, Blank M, Buhmann R, Resch A. Aptamers—basic research, drug development, and clinical applications. Applied microbiology and biotechnology. 2005; 69(4):367–374. https://doi.org/10.1007/s00253-005-0193-5 PMID: 16283295

16. Elskens JP, Elskens JM, Madder A. Chemical modification of aptamers for increased binding affinity in diagnostic applications: Current status and future prospects. International Journal of Molecular Sciences. 2020; 21(12):4522. https://doi.org/10.3390/ijms21124522 PMID: 32630547

17. D'Souza S, Prema K, Balaji S. Machine learning models for drug–target interactions: current knowledge and future directions. Drug Discovery Today. 2020; 25(4):748–756. https://doi.org/10.1016/j.drudis.2020.03.003 PMID: 32171918

18. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

19. Townshend RJ, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, et al. Geometric deep learning of RNA structure. Science. 2021; 373(6558):1047–1051. https://doi.org/10.1126/science.abe5650 PMID: 34446608

20. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Advanced Drug Delivery Reviews. 2021;. https://doi.org/10.1016/j.addr.2021.05.016 PMID: 34019959

21. Hoinka J, Zotenko E, Friedman A, Sauna ZE, Przytycka TM. Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers. Bioinformatics. 2012; 28(12):i215–i223. https://doi.org/10.1093/bioinformatics/bts210 PMID: 22689764

22. Song J, Zheng Y, Huang M, Wu L, Wang W, Zhu Z, et al. A sequential multidimensional analysis algorithm for aptamer identification based on structure analysis and machine learning. Analytical chemistry. 2019; 92(4):3307–3314. https://doi.org/10.1021/acs.analchem.9b05203

23. Alam KK, Chang JL, Burke DH. FASTAptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. Molecular Therapy-Nucleic Acids. 2015; 4:e230. https://doi.org/10.1038/mtna.2015.4 PMID: 25734917

24. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic acids research. 2015; 43(W1):W39–W49. https://doi.org/10.1093/nar/gkv416 PMID: 25953851

25. Jiang P, Meyer S, Hou Z, Propson NE, Soh HT, Thomson JA, et al. MPBind: a Meta-motif-based statistical framework and pipeline to Predict Binding potential of SELEX-derived aptamers. Bioinformatics. 2014; 30(18):2665–2667. https://doi.org/10.1093/bioinformatics/btu348 PMID: 24872422

26. Zhou Q, Xia X, Luo Z, Liang H, Shakhnovich E. Searching the Sequence Space for Potent Aptamers Using SELEX in Silico. Journal of Chemical Theory and Computation. 2015; 11(12):5939–5946. https://doi.org/10.1021/acs.jctc.5b00707 PMID: 26642994

27. Zhou Q, Sun X, Xia X, Fan Z, Luo Z, Zhao S, et al. Exploring the Mutational Robustness of Nucleic Acids by Searching Genotype Neighborhoods in Sequence Space. The Journal of Physical Chemistry Letters. 2017; 8(2):407–414. https://doi.org/10.1021/acs.jpclett.6b02769 PMID: 28045264

28. Pressman A, Moretti JE, Campbell GW, Müller UF, Chen IA. Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. Nucleic Acids Research. 2017; 45(14):8167–8179. https://doi.org/10.1093/nar/gkx540 PMID: 28645146

29. Pressman AD, Liu Z, Janzen E, Blanco C, Müller UF, Joyce GF, et al. Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA. Journal of the American Chemical Society. 2019; 141(15):6213–6223. https://doi.org/10.1021/jacs.8b13298 PMID: 30912655

30. Koo PK, Anand P, Paul SB, Eddy SR. Inferring sequence-structure preferences of RNA-binding proteins with convolutional residual networks. BioRxiv. 2018; p. 418459.

31. Zrimec J, Buric F, Kokina M, Garcia V, Zelezniak A. Learning the regulatory code of gene expression. Frontiers in Molecular Biosciences. 2021; 8. https://doi.org/10.3389/fmolb.2021.673363 PMID: 34179082

32. Koo PK, Ploenzke M. Deep learning for inferring transcription factor binding sites. Current opinion in systems biology. 2020; 19:16–23. https://doi.org/10.1016/j.coisb.2020.04.001 PMID: 32905524

33. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, et al. Deep diversification of an AAV capsid protein by machine learning. Nature Biotechnology. 2021; 39(6):691–696. https://doi.org/10.1038/s41587-020-00793-4 PMID: 33574611

34. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Reports on Progress in Physics. 2018; 81(3):032601. https://doi.org/10.1088/1361-6633/aa9965 PMID: 29120346

35. De Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A, et al. Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic acids research. 2015; 43(21):10444–10455. https://doi.org/10.1093/nar/gkv932 PMID: 26420827

36. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

37. Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, et al. An evolution-based model for designing chorismate mutase enzymes. Science. 2020; 369(6502):440–445. https://doi.org/10.1126/science.aba3304 PMID: 32703877

38. Zhou Q, Kunder N, De la Paz JA, Lasley AE, Bhat VD, Morcos F, et al. Global pairwise RNA interaction landscapes reveal core features of protein recognition. Nature communications. 2018; 9(1):1–10. https://doi.org/10.1038/s41467-018-04729-0 PMID: 29955037

39. Tubiana J, Cocco S, Monasson R. Learning protein constitutive motifs from sequence data. eLife. 2019; 8:e39397. https://doi.org/10.7554/eLife.39397 PMID: 30857591

40. Bravi B, Tubiana J, Cocco S, Monasson R, Mora T, Walczak AM. RBM-MHC: A Semi-Supervised Machine-Learning Method for Sample-Specific Prediction of Antigen Presentation by HLA-I Alleles. Cell systems. 2021; 12(2):195–202. https://doi.org/10.1016/j.cels.2020.11.005 PMID: 33338400

41. Zhou Y, Qi X, Liu Y, Zhang F, Yan H. DNA-Nanoscaffold-Assisted Selection of Femtomolar Bivalent Human alpha-Thrombin Aptamers with Potent Anticoagulant Activity. ChemBioChem. 2019; 20 (19):2494–2503. https://doi.org/10.1002/cbic.201900578 PMID: 31083763

42. Hinton GE. Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation. 2002; 14(8):1771–1800. https://doi.org/10.1162/089976602760128018 PMID: 12180402

43. Tieleman T. Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient. In: Proceedings of the 25th International Conference on Machine Learning. ICML'08. New York, NY, USA: Association for Computing Machinery; 2008. p. 1064–1071. Available from: https://doi.org/10.1145/1390156.1390290.

44. Neher RA, Shraiman BI. Statistical genetics and evolution of quantitative traits. Reviews of Modern Physics. 2011; 83(4):1283–1300. https://doi.org/10.1103/RevModPhys.83.1283

45. Hartl DL, Dykhuizen DE, Dean AM. Limits of Adaptation: The Evolution of Selective Neutrality. Genetics. 1985; 111(3):655–674. https://doi.org/10.1093/genetics/111.3.655 PMID: 3932127

46. Padmanabhan K, Padmanabhan K, Ferrara J, Sadler JE, Tulinsky A. The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. Journal of Biological Chemistry. 1993; 268 (24):17651–17654. https://doi.org/10.1016/S0021-9258(17)46749-4 PMID: 8102368

47. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific reports. 2018; 8(1):1–14. https://doi.org/10.1038/s41598-018-29325-6 PMID: 30026539

48. Wagner A. Robustness and evolvability in living systems. Princeton university press; 2013.

49. Bisardi M, Rodriguez-Rivas J, Zamponi F, Weigt M. Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. Molecular biology and evolution. 2022; 39(1):msab321. https://doi.org/10.1093/molbev/msab321 PMID: 34751386

50. Lou TF, Weidmann CA, Killingsworth J, Hall TMT, Goldstrohm AC, Campbell ZT. Integrated analysis of RNA-binding protein complexes using in vitro selection and high-throughput sequencing and sequence specificity landscapes (SEQRS). Methods. 2017; 118:171–181. https://doi.org/10.1016/j.ymeth.2016.10.001 PMID: 27729296

51. Hammers CM, Stanley JR. Antibody phage display: technique and applications. The Journal of investigative dermatology. 2014; 134(2):e17. https://doi.org/10.1038/jid.2013.521 PMID: 24424458

52. Kretzschmar T, Von Rüden T. Antibody discovery: phage display. Current opinion in biotechnology. 2002; 13(6):598–602. https://doi.org/10.1016/S0958-1669(02)00380-4 PMID: 12482520

53. Sesta L, Uguzzoni G, Fernandez-de Cossio-Diaz J, Pagnani A. AMaLa: Analysis of Directed Evolution Experiments via Annealed Mutational Approximated Landscape. International journal of molecular sciences. 2021; 22(20):10908. https://doi.org/10.3390/ijms222010908 PMID: 34681569

54. Tubiana J, Monasson R. Emergence of Compositional Representations in Restricted Boltzmann Machines. Phys Rev Lett. 2017; 118:138301. https://doi.org/10.1103/PhysRevLett.118.138301 PMID: 28409983

55. Roussel C, Cocco S, Monasson R. Barriers and dynamical paths in alternating Gibbs sampling of restricted Boltzmann machines. Physical Review E. 2021; 104(3):034109. https://doi.org/10.1103/PhysRevE.104.034109 PMID: 34654094

56. Kogut M, Kleist C, Czub J. Why do G-quadruplexes dimerize through the 5'-ends? Driving forces for G4 DNA dimerization examined in atomic detail. PLoS computational biology. 2019; 15(9):e1007383. https://doi.org/10.1371/journal.pcbi.1007383 PMID: 31539370