Unsupervised generative models for in vitro selection experiments

R. Monasson

Department of Physics, Ecole Normale Supérieure & CNRS, Paris

in collaboration with

S. Cocco, A. Di Gioacchino (Ecole Normale Supérieure)

J. Procyk, P. Sulc (Arizona State University)

RNA meeting, Benasque, August 9th, 2022

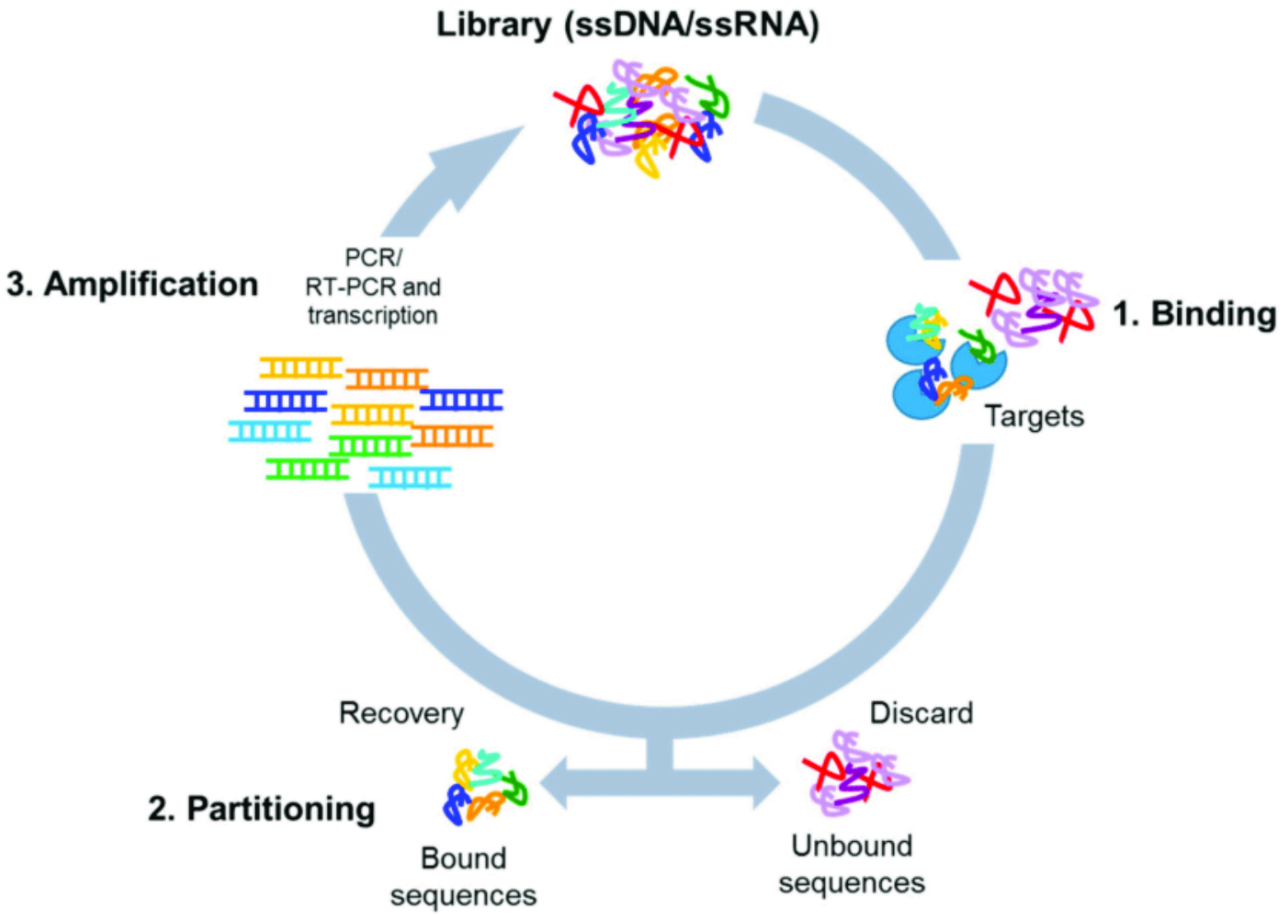
In vitro selection: a practical viewpoint and some issues

Selection is a fundamental biological process, at the core of Darwinian evolution.

In vitro realization:

Directed evolution experiments, e.g. SELEX (Systematic Evolution of Ligands by EXponential enrichment)

Very useful for bioengineering



Catuogno, Esposito, 2017

In vitro selection: a practical viewpoint and some issues

Sequences **v**	Counts (Round 1)	Counts (Round 2)	Counts (Round 3)	…
AUCCGU…	100	300	900	
UGGCAA…	1600	400	100	
GCUAAG…	1	2	4	
GCCUAU…	10	5	12	

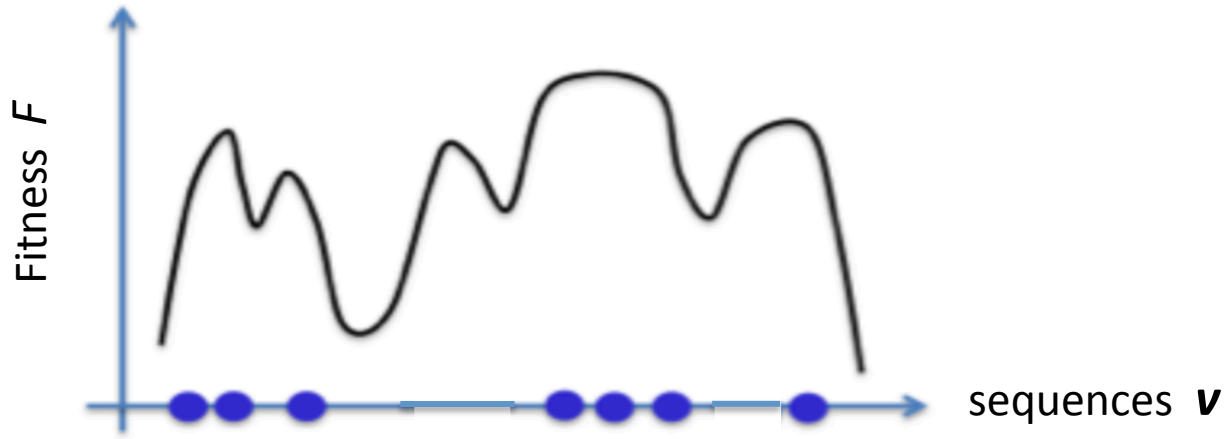
Relative Enrichment $$R.E.(v, r-1 \rightarrow r) = \frac{C_r(v)}{C_{r-1}(v)}$$ proxy for (exp) fitness

- Sampling in sequence space is generally very sparse, dependent on initial library
- Counts may be unreliable (biases in sub-sampling e.g. for small numbers, in amplification, sequencing errors …)
- Relationship between fitness and sequence?

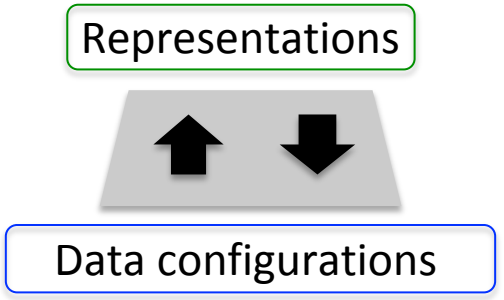
Selection: a practical viewpoint and some issues

Basic Assumptions

- Existence of sequence-to-fitness function
- R.E. is exponentially increasing with fitness (if biases are avoided)



Can be inferred by unsupervised methods, such as restricted Boltzmann machines (RBM), variational auto-encoders (VAE), …

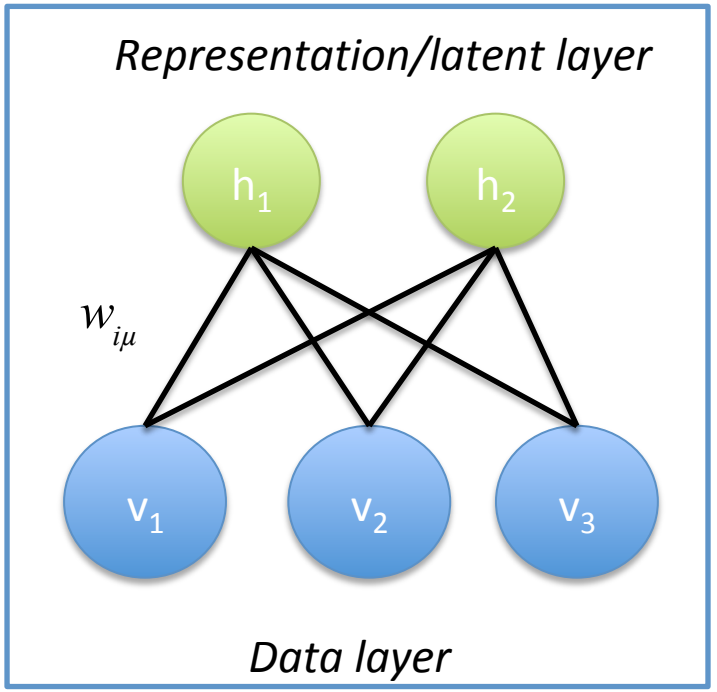


Objectives: Build scoring model,
Understand key features in data (interpretable representations …)
Generate new « data », possibly with desired features

Restricted Boltzmann Machines

- **Graphical model** constituted by two sets of random variables that are coupled together.

$$\log P(v,h) = \sum_i g_i\left(v_i\right) + \sum_{i,\mu} w_{i\mu}\left(v_i\right) h_\mu - \sum_\mu U_\mu(h_\mu)$$



- Marginal distribution: $P(v) = \int dh \;\; P(v,h)$

- Joint distribution of *v,h* define

$$\log P(v), \quad \log P(h|v), \quad \log P(v|h)$$

maximized over data set

Extract representation from data

Design data from representation

Why RBM?

- Simple(st) generative model implementing the data-representation duality

Why RBM?

- Simple(st) generative model implementing the data-representation duality
- Competitive with deeper architectures in some relevant situations
    - Not all applications are supported by huge data sets …

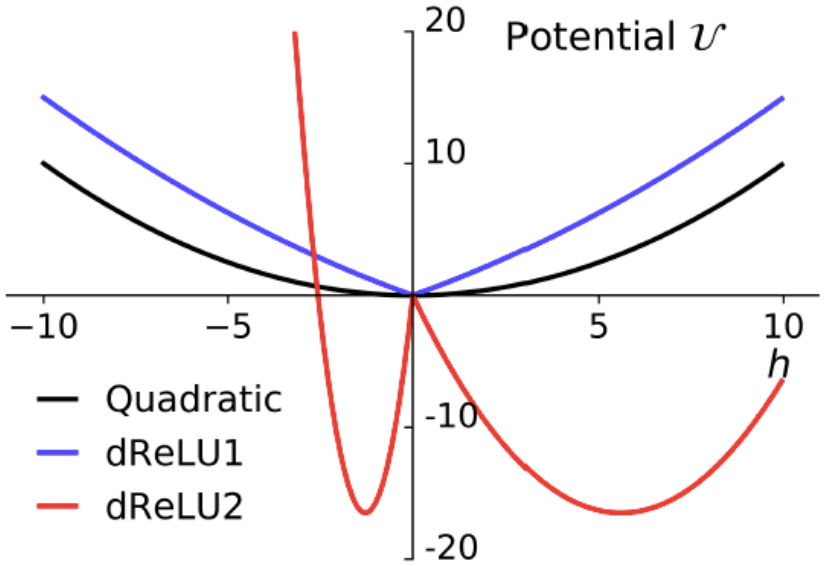
Analysis of T Cell Receptors (TCR) populations in patients suffering from pancreatic tumors

Luzka et al.,
Neoantigen quality predicts immunoediting and clonal evolution in pancreatic cancer survivors,
Nature 2022

Why RBM?

- Simple(st) generative model implementing the data-representation duality
- Competitive with deeper architectures in some relevant situations
  - Not all applications are supported by huge data sets …
  - No ad hoc assumption on the distribution of latent variables (potentials *U* are learned from data) ≠ VAE

$$\log P(v,h) = \sum_i g_i\left(v_i\right) + \sum_{i,\mu} w_{i\mu}\left(v_i\right) h_\mu - \sum_\mu U_\mu(h_\mu)$$



Why RBM?

- Simple(st) generative model implementing the data-representation duality
- Competitive with deeper architectures in some relevant situations
    - Not all applications are supported by huge data sets …
    - No ad hoc assumption on the distribution of latent variables (potentials *U* are learned from data) ≠ VAE
    - Interpretable (under some appropriate regularization conditions)

1. *Sparsity of representations*
2. *Sparsity of weights*
3. *Disentanglement of representations based on annotated data*

Performance vs. Interpretability Trade-off

Fernandez de Cossio Diaz, Cocco, RM, arxiv:2206.11600, 2022

Why RBM?

- Simple(st) generative model implementing the data-representation duality
- Competitive with deeper architectures in some relevant situations
    - Not all applications are supported by huge data sets …
    - No ad hoc assumption on the distribution of latent variables (potentials *U* are learned from data) ≠ VAE
    - Interpretable (under some appropriate regularization conditions)
- Appealing from a statistical mechanics point of view = deeply related to the Hopfield model (1982)

  Analytical understanding of « phases » in the hyperparameter space

Review by Decelle, Furtlehner 2020

Two applications

- to RNA riboswitches:

  Talk by **J. Fernandez de Cossio Diaz** in the « RNA Design » session later this week

- to SELEX experiments for the design of DNA aptamers binding to thrombin

Thrombin

Fibrinogen (soluble) → Fibrin (-> protein network)

ChemPubSoc Europe
DOI: 10.1002/cbic.201900265
CHEMBIOCHEM Full Papers

VIP *Very Important Paper*

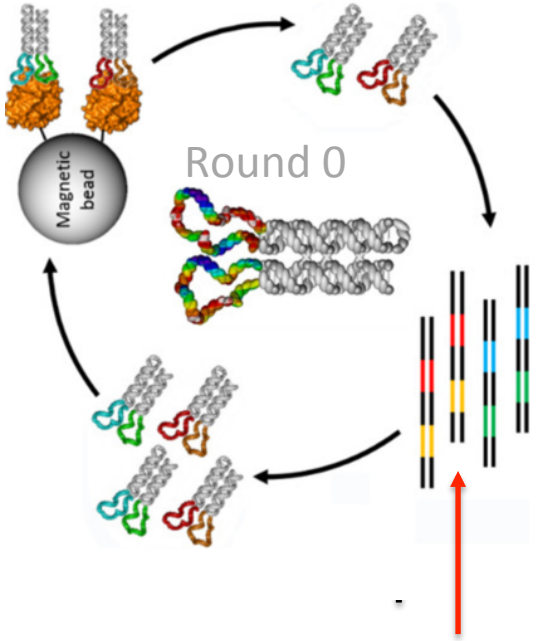
DNA-Nanoscaffold-Assisted Selection of Femtomolar Bivalent Human α-Thrombin Aptamers with Potent Anticoagulant Activity

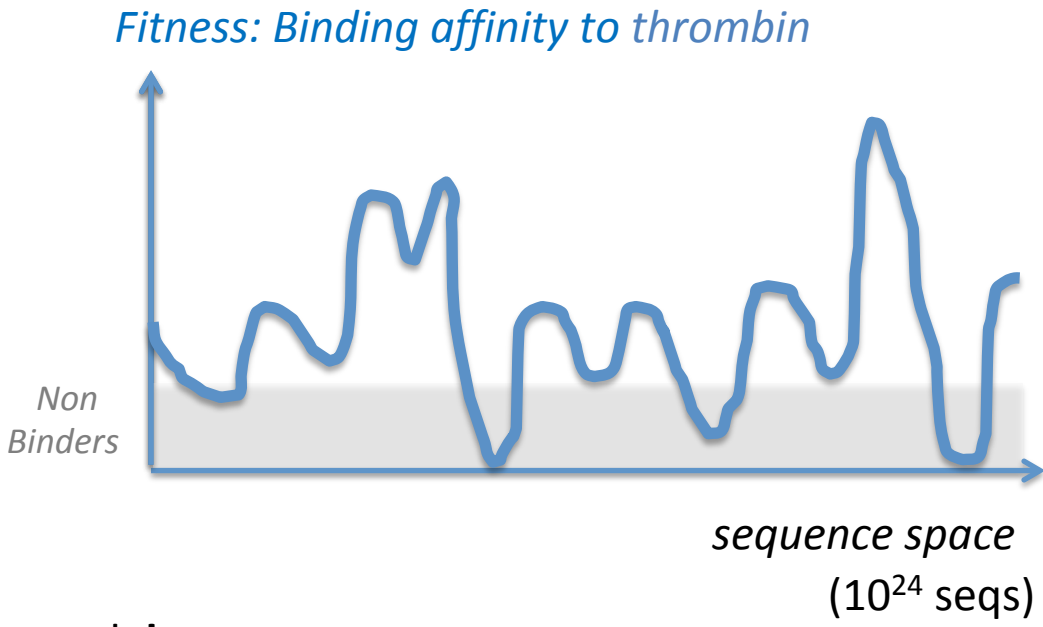
Yu Zhou, Xiaodong Qi, Yan Liu, Fei Zhang,* and Hao Yan*[a]

(2019)

Fitness and design of aptamers from SELEX data

Design of DNA aptamers that inhibit the coagulant activity of Thrombin



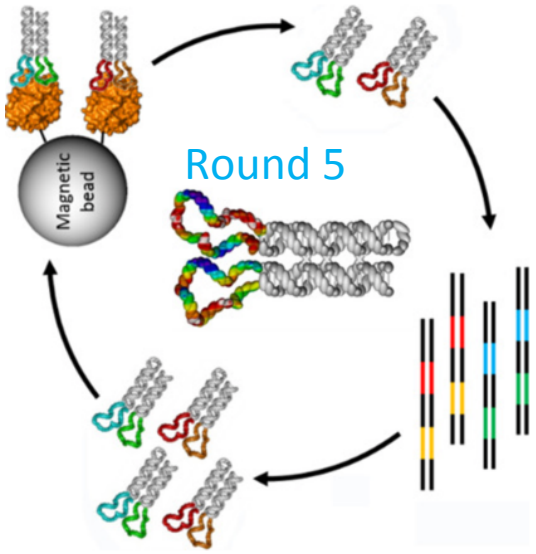


Initial set of 10^{15} random seqs of 20 nt for each **loop**

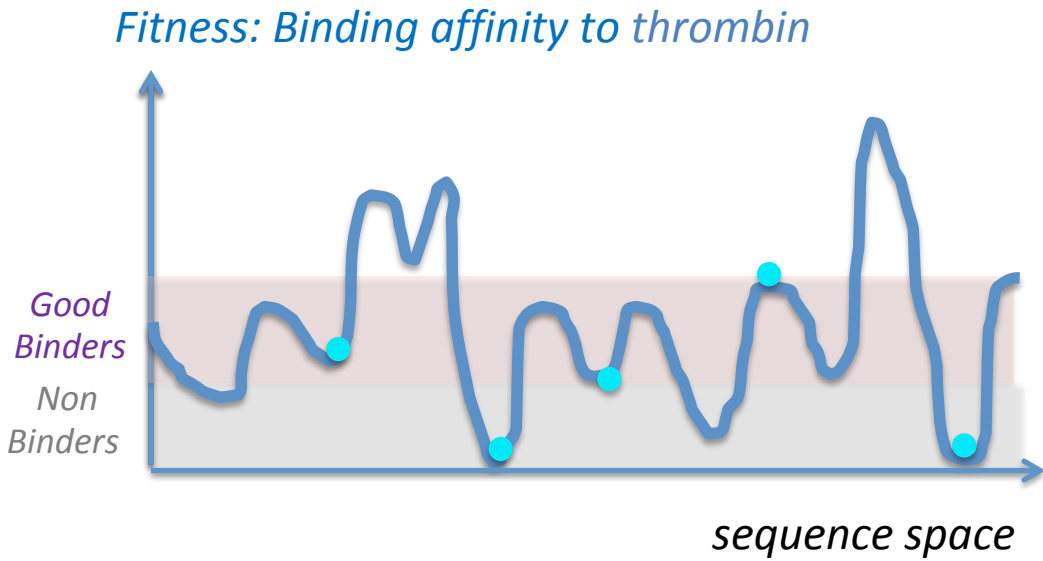
ATAGCTGATGAGCGCTACAC
ACGTTAGCTGTCGATAATGC
.
.

Zhou et al, ChemBioChem 2019

Fitness and design of aptamers from SELEX data

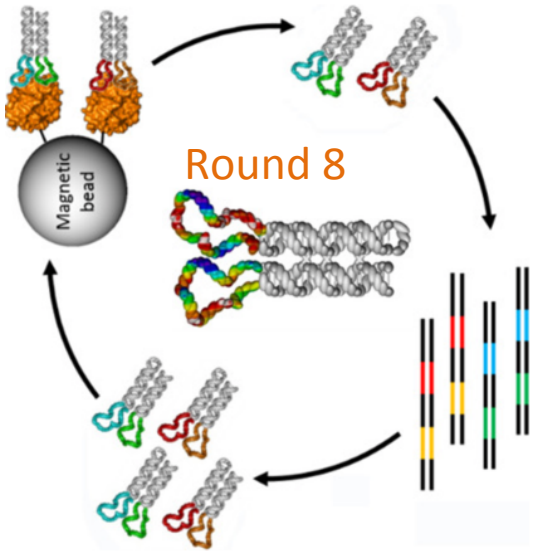


Sampling
10^6 Seqs

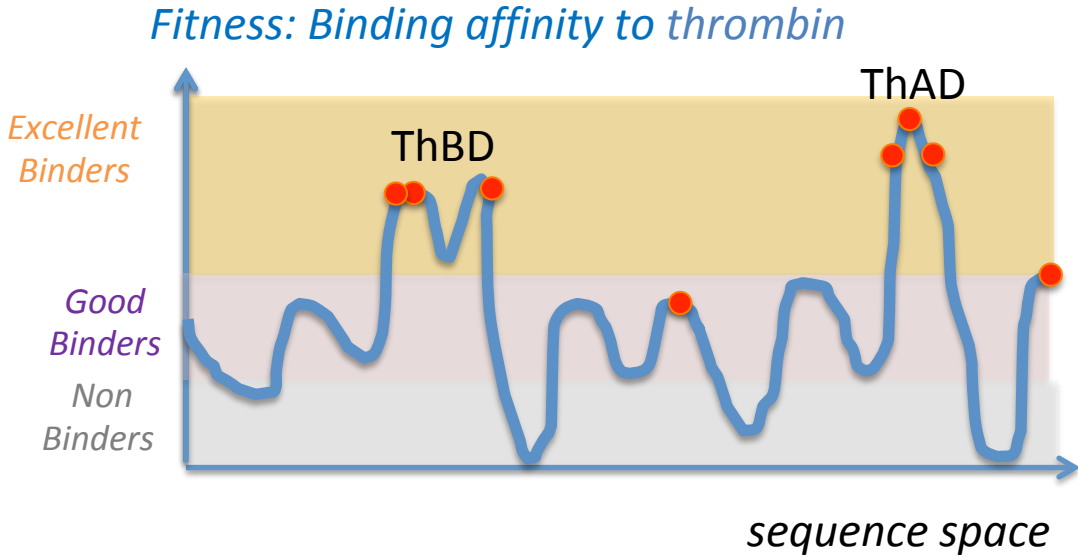


Zhou et al, ChemBioChem 2019

Fitness and design of aptamers from SELEX data



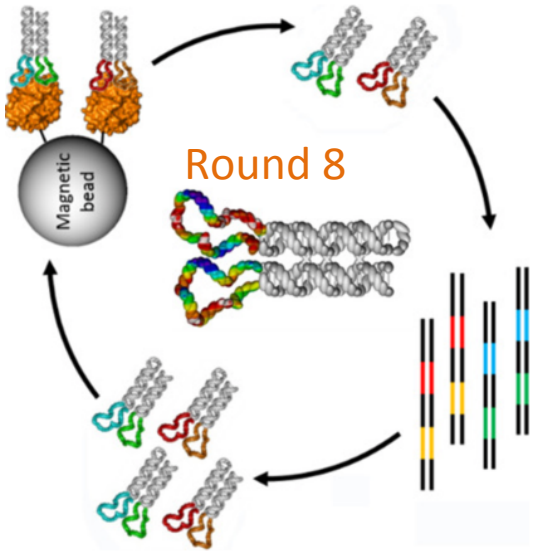
Sampling 10^6 Seqs



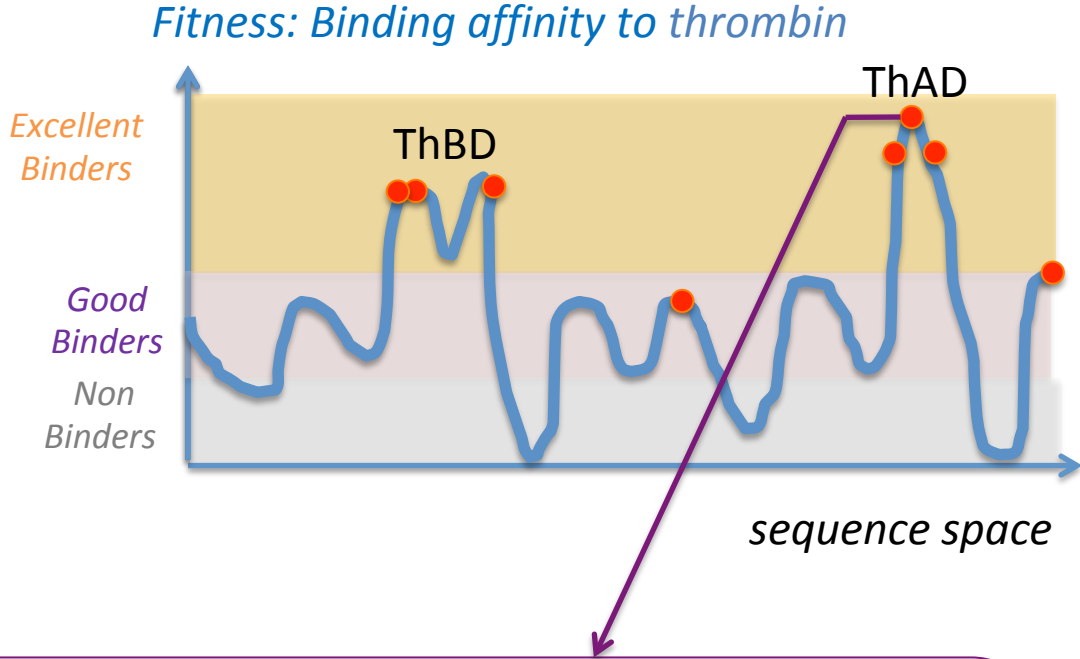
Excellent binders to thrombin are found

Zhou et al, ChemBioChem 2019

Fitness and design of aptamers from SELEX data



Sampling 10^6 Seqs

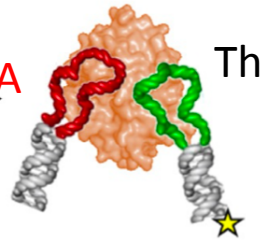


Excellent binders to thrombin are found

ThAD

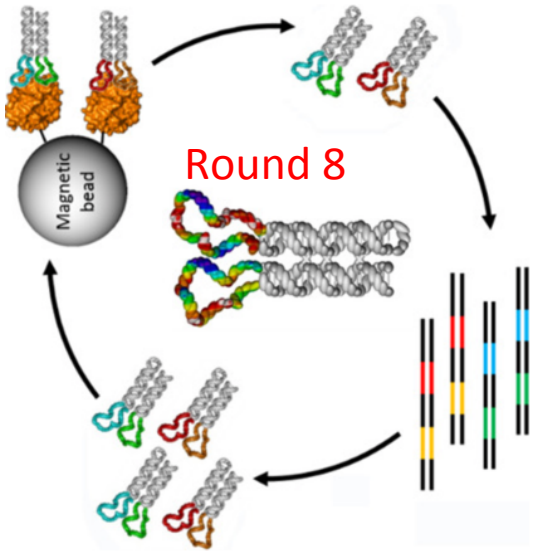
AGGGATGATGTGTGGTAGGC GTAGGATGGGTAGGGTGGTC

K_D=340 fM

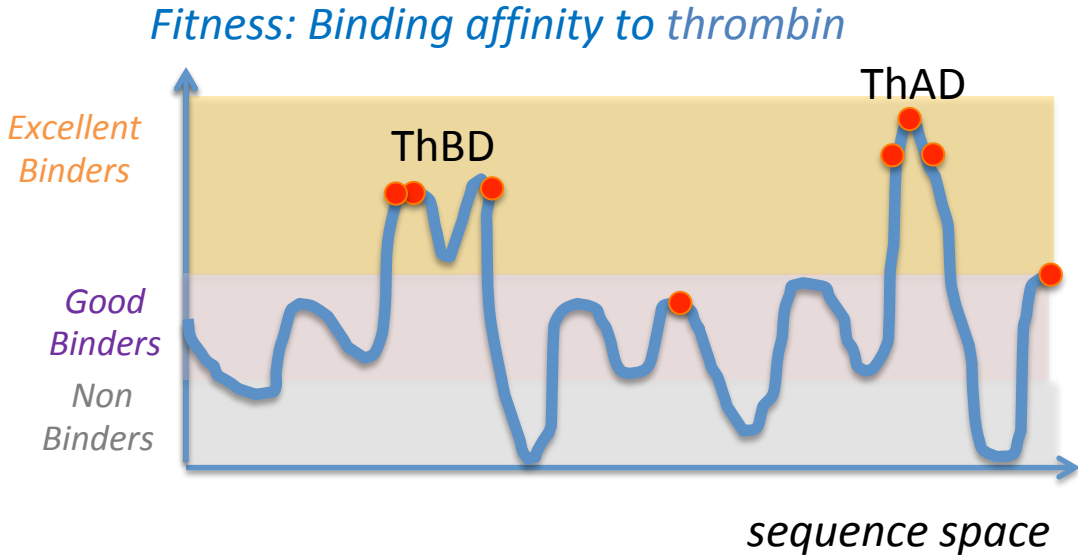


Zhou et al, ChemBioChem 2019

Fitness and design of aptamers from SELEX data



Sampling 10^6 Seqs



RBM models with two objectives

Predict selection at next rounds

Round 5 — Round 6 — Round 7 — Round 8

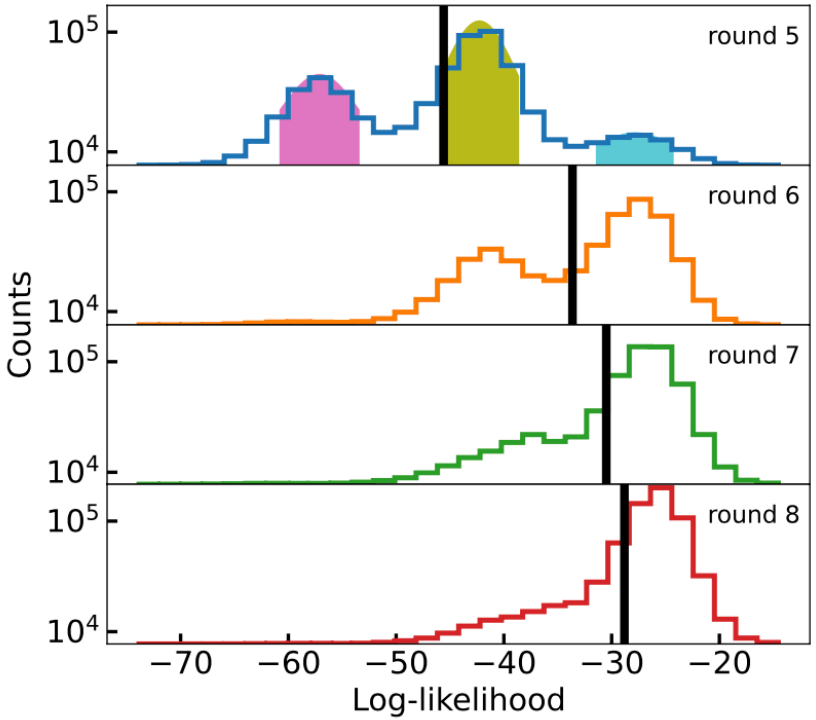
Design evolvable binders

RBM log-likelihoods predict enrichments at later rounds

- Histograms of RBM log-likelihoods

$$\log P_{r=6}(v)$$

(computable in time linear in L,M)

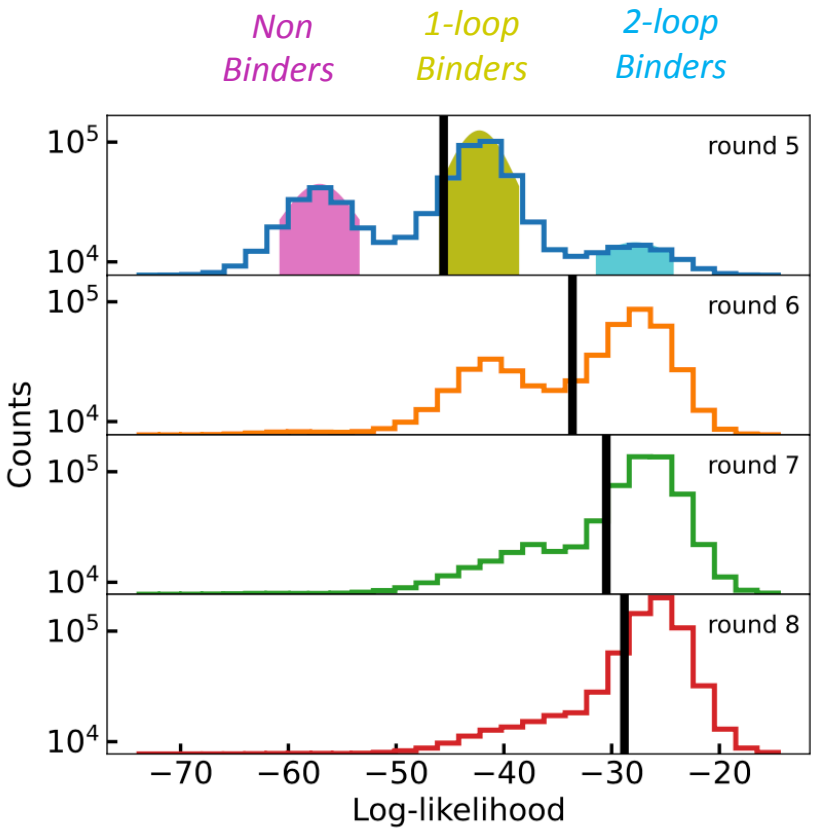


RBM log-likelihoods predict enrichments at later rounds

- Histograms of RBM log-likelihoods

$$\log P_{r=6}(v)$$

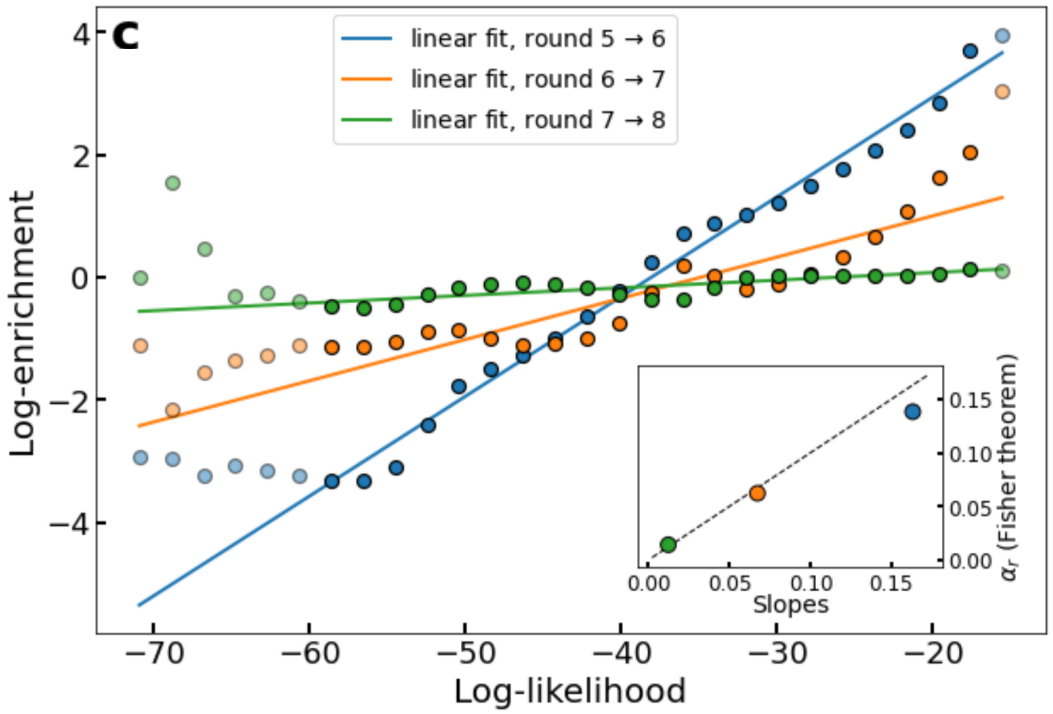
(computable in time linear in L,M)



- Precise connection with fitness ?

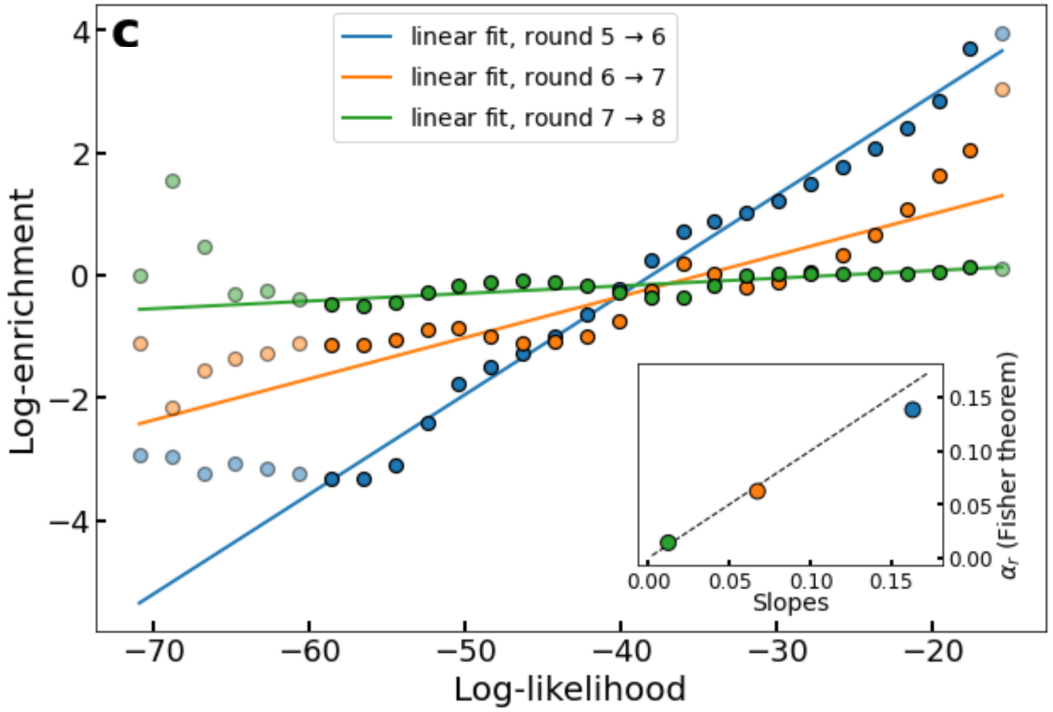
$$F_r(v) = \log RE(v, r-1 \rightarrow r) = \log \frac{C_r(v)}{C_{r-1}(v)}$$

RBM log-likelihoods predict enrichments at later rounds



Almost linear relation !

RBM log-likelihoods predict enrichments at later rounds



Almost linear relation !

Observation 1:

Slope decreases with rounds

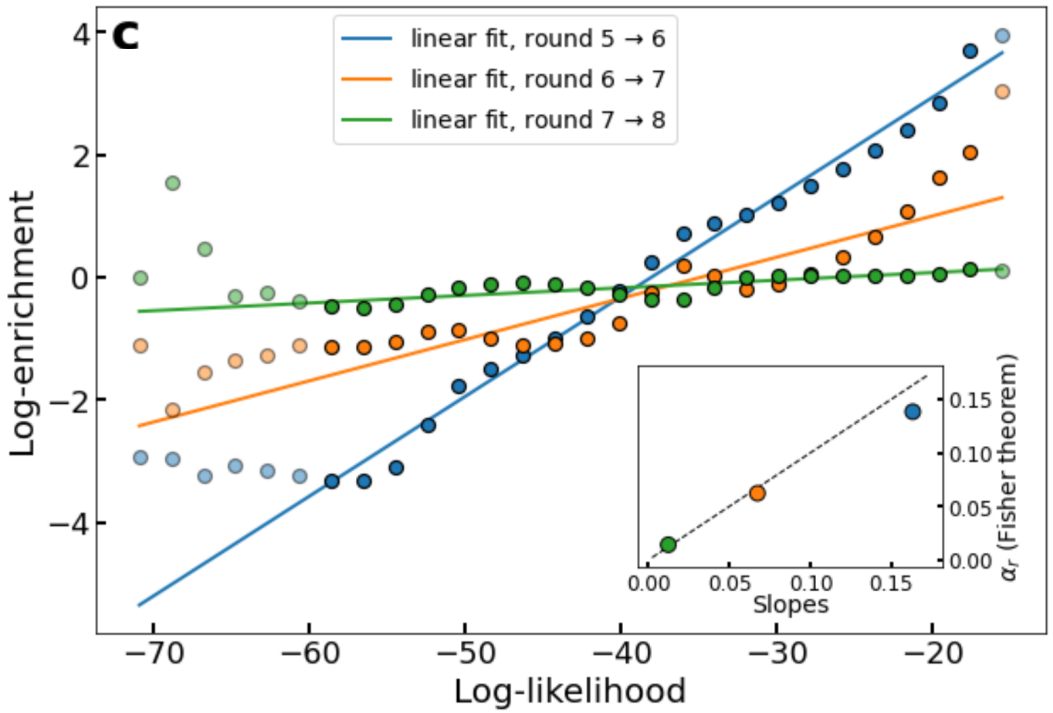
$$F_r(v) = \alpha_r \times F(v)$$

intensity of selection at round r

Thus, likelihood $p_r(v) \propto e^{\alpha_{r-1} F(v)} p_{r-1}(v) \propto \ldots \propto e^{\beta_r F(v)}$ with $\beta_r = \alpha_0 + \alpha_1 + \ldots + \alpha_{r-1}$

similar to inverse temperature in statistical physics

RBM log-likelihoods predict enrichments at later rounds



Almost linear relation !

Observation 1:

Slope decreases with rounds

$$F_r(v) = \alpha_r \times F(v)$$

intensity of selection at round r

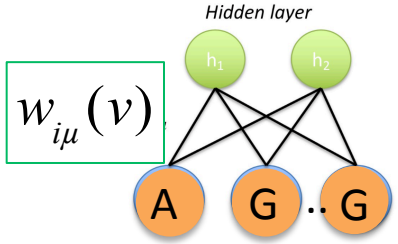
Thus, likelihood $p_r(v) \propto e^{\alpha_{r-1} F(v)} p_{r-1}(v) \propto \ldots \propto e^{\beta_r F(v)}$ with $\beta_r = \alpha_0 + \alpha_1 + \ldots + \alpha_{r-1}$

similar to inverse temperature in statistical physics

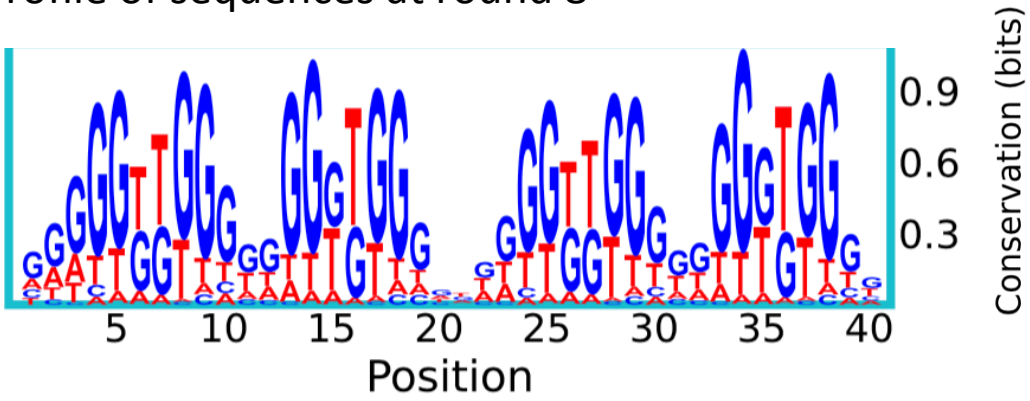
Observation 2: Slopes can be estimated from Fisher's fundamental theorem

$$\Delta \left\langle \log p \right\rangle_{r-1 \to r} = \alpha_{r-1} \times \mathrm{var}\left(\log p \right)_{r-1}$$

RBM weigths reveal nucleotidic motifs

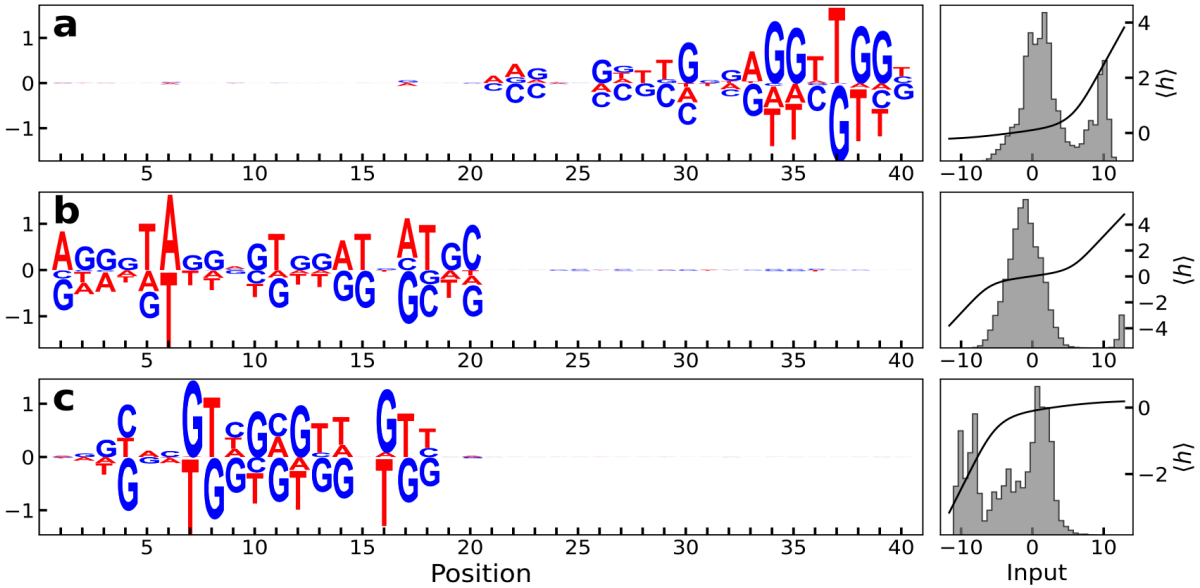


Profile of sequences at round 8



G Quadruplex motif

Three weights, showing that the two loops are independent:

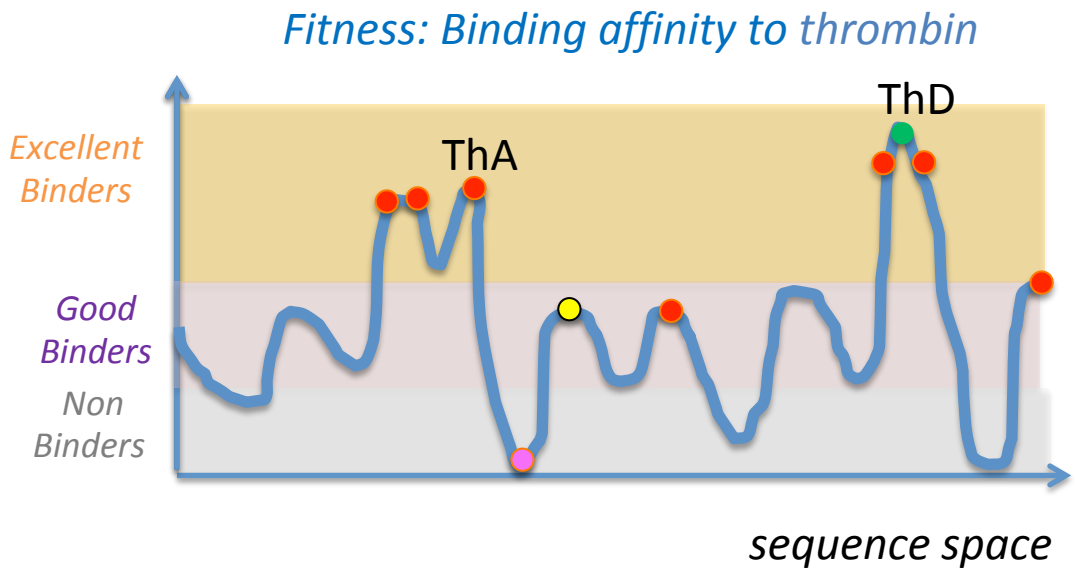


Variation of second half of G-quadruplex (pos 33-39)

Long-range AT-rich motif

Correlated shifts of boundaries of G-quadruplex

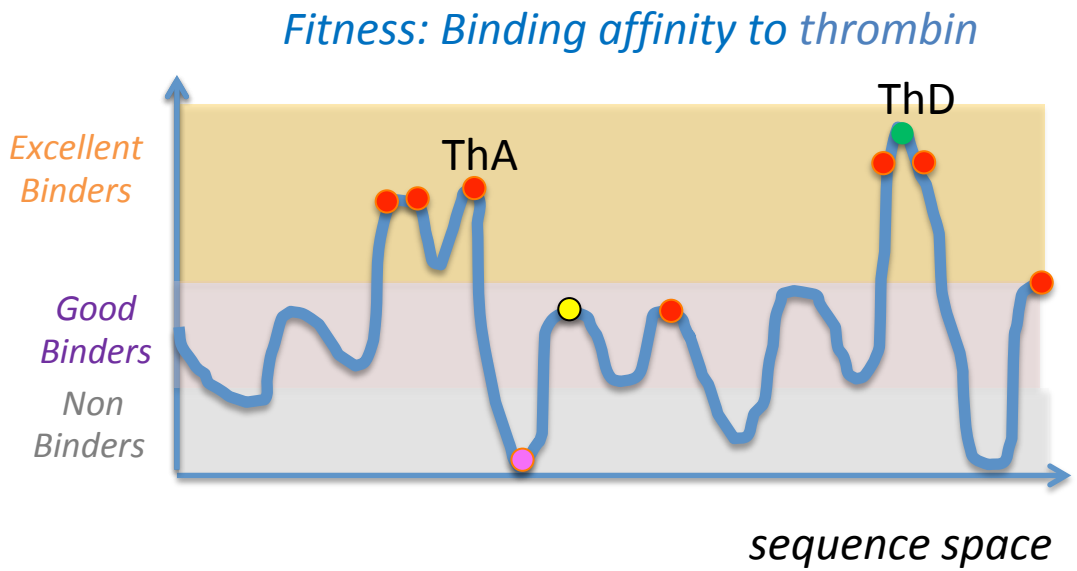
Sampling of RBM to design evolvable binders



Sequences obtained by sampling RBM **inferred from round 8 data (with counts)** are very close variants of best binders (ThA or ThD)

Sequences obtained by sampling RBM **inferred from round 8 UNIQUE data** are diverse

Sampling of RBM to design evolvable binders



Sequences obtained by sampling RBM **inferred from round 8 data (with counts)** are very close variants of best binders (ThA or ThD)

Sequences obtained by sampling RBM **inferred from round 8 UNIQUE data** are diverse

RBM used to

- Predict binding for sequences with ● low counts
- Design new binders ● by MC-sampling
- Identify deleterious mutations ● that strongly damage excellent binders

} **Overall 27 predictions** (6 for non-binding, 21 for binding)

Validation by gel shift assays:
25 trues: **93% Accuracy**

Di Gioacchino et al,
BiorXiv 2022.03.12.484094, under review