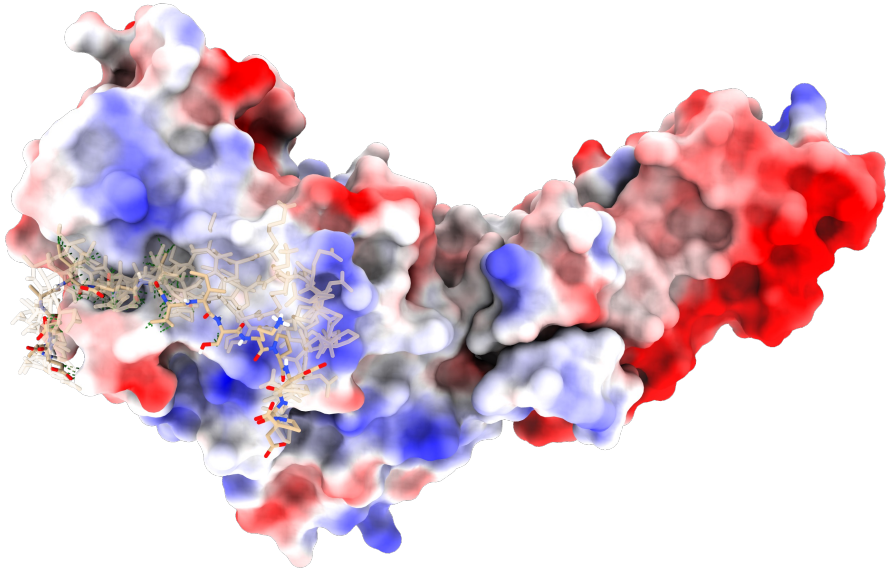
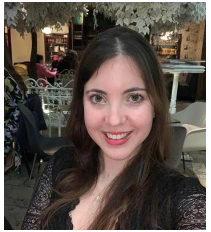
Harnessing Sequence Generative Models for Inhibitory Peptide Design: a Case Study





Dr. Jérôme Tubiana

BeVAS 18/04/2023

Joint work with:
Haim J. Wolfson lab
Maayan Gal lab
Lucia Adriana-Lifshits

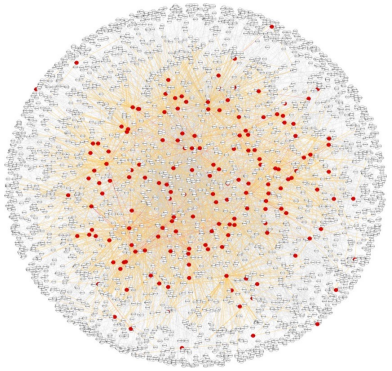
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

HFSP

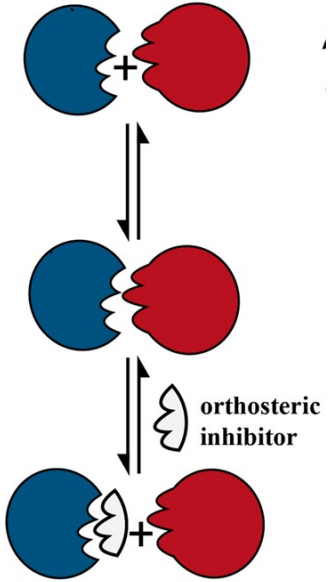
Edmond J. Safra Center for Bioinformatics
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Protein-protein Interaction (PPIs) Inhibitors

- Aberrant PPIs are associated with various diseases.
- PPIs are attractive targets for basic research, therapeutic & pesticide purpose.
- Interfering with PPIs with small molecules is challenging, due to their physio-chemical properties.
  - PPI interfaces are larger, flatter, hydrophobic
  - 40% of PPIs involve a disordered partner
- mAb widely successful, but only for extracellular targets

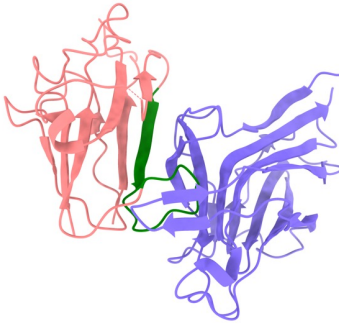


The Human P-P interactome
(~20K nodes, 650K edges)
Bonetta et al Nature 2010.



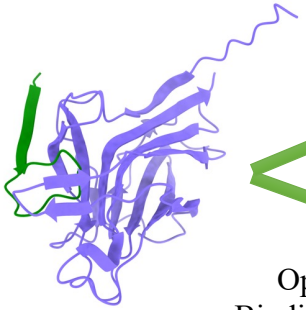
Credit: *Lu et al. Nature Signal Transduction and Targeted Therapy 2020*

Inhibiting PPIs with peptides: principle, benefits and challenges



EphB4-EphB2 complex

Extract interacting fragment

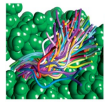


Optimize Binding affinity

Display experiments: Mutagenesis & selection

flexPepDock

Raveh, London, Schueler-Furman
PLOS CB 2010



PepCrawler

Donsky and Wolfson
Bioinformatics 2011

In-silico docking & binding energy optimization

✅ Peptides are suitable for binding protein-protein interfaces

✅ Native protein-peptides interactions are highly specific

✅ Initial peptide can be derived from native substrate in >50% cases (London et al. *Proteins 2011*)

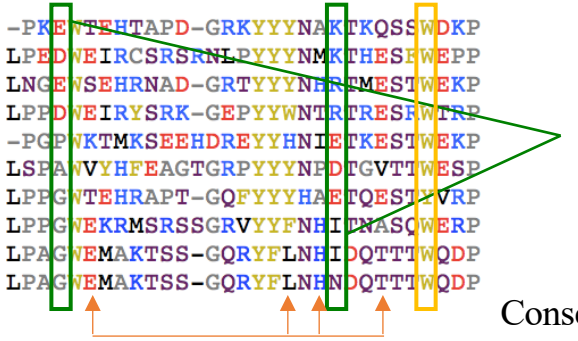
❌ Flexibility makes computational modeling of protein-peptide interactions challenging (sampling & scoring)

❌ Display experiments can be difficult to setup

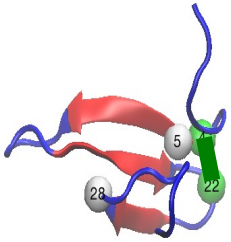
❌ Unclear how to efficiently explore the vast search space (20^L peptides of length L)

❌ Unclear how to select for specificity and other desirable properties (bioavailability, immunogenicity)

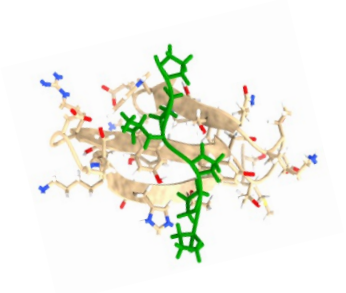
Evolutionary-based sequence generative models for protein design



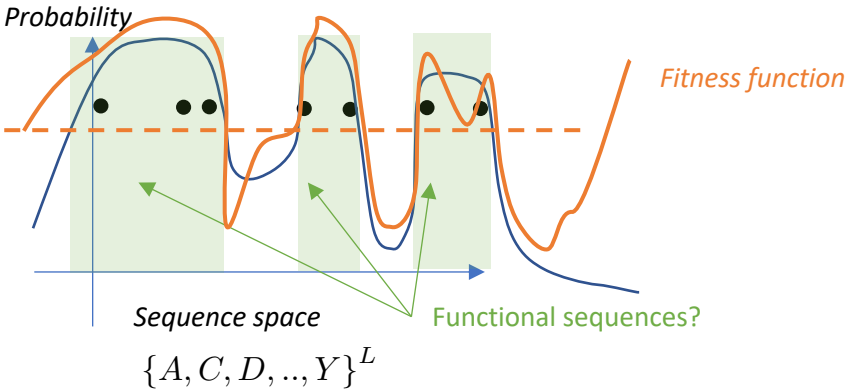
Conservation



Coevolution



High-order coevolution



$$\{A, C, D, .., Y\}^L$$

Non-exhaustive list of successful SGM-based design experiments:

1. Chorismate Mutase (Russ et al. Science 2020)
2. Luciferases (Hawkins-Hooker et al. PLOS CB 2021)
3. Malate Dehydrogenase (Repecka et al. Nat. Mach. Int. 2021)
4. Nanobody libraries (Shin et al. Com. 2021)
5. GFP (Biswas et al. Nat. Methods 2021)
6. SH3 domains (Lian et al. BiorXiv 2022)
7. Copper Superoxide Dismutase (Johnson et al. BiorXiv 2023)
8. Cas9 PAM-interacting domain (Malbranke et al. BiorXiv 2023)

Evolutionary-based generative models for peptide design

Table 1 Peptide generation studies using deep generative models. Abbreviations: NML, neural language model; VAE, variational autoencoder; GAN, generative adversarial network; AMP, antimicrobial peptide; ACP, anticancer peptide; CPP, cell-penetrating peptide; PMO, phosphorodiamidate morpholino oligomer

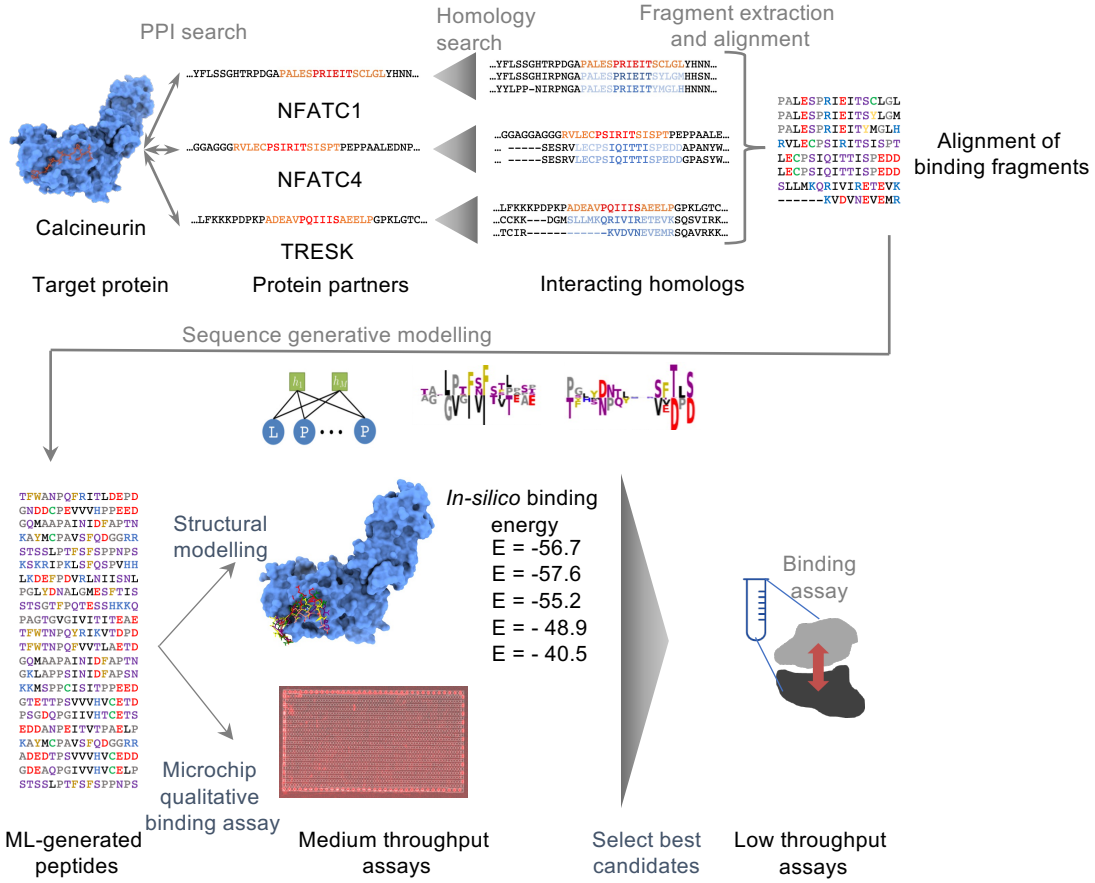
Method	Feature Representation	Application	Citation	Year
NML	One-hot	AMP generation	Müller *et al.*[56]	2018
NLM	Character sequence	AMP generation	Nagarajan *et al.*[52]	2018
NLM	Character sequence	ACP generation	Grisoni *et al.*[46]	2018
NLM	Learned representation using one-hot	Signal peptide generation	Wu *et al.*[55]	2020
NLM	Learned representation using structural and evolutionary data	AMP generation	Caceres-Delpiano *et al.*[54]	2020
NLM	One-hot	AMP generation	Wang *et al.*[41]	2021
NLM	Character sequence	CPP generation	Tran *et al.*[53]	2021
NLM	One-hot	AMP generation	Capecchi *et al.*[42]	2021
NLM	Fingerprint, one-hot	PMO delivery peptide generation	Schissel *et al.*[57]	2021
VAE	Learned representation using character sequence	AMP generation	Das *et al.*[38]	2018
VAE	Learned representation using one-hot	AMP generation	Dean *et al.*[44]	2020
VAE	Learned representation using character sequence	AMP generation	Das *et al.*[45]	2021
GAN	Character sequence	AMP generation	Tucs *et al.*[50]	2020
GAN	Character sequence/PDB structure	ACP generation	Rossetto *et al.*[48]	2020
GAN	Learned representation using character sequence	AMP generation	Ferrell *et al.*[39]	2020
GAN	Character sequence	AMP generation	Oort *et al.*[40]	2021
GAN	Sequence of amino acid property vectors	Immunogenic peptide generation	Li *et al.*[51]	2021
GAN	Character sequence	AMP generation	Surana *et al.*[49]	2021

Table reproduced from *Deep generative models for peptide design*
F. Wan, D. Kontogiorgos-Heintz and C. de la Fuente-Nunez, Digital Discovery 2022

Adapting to PPI inhibitor design?

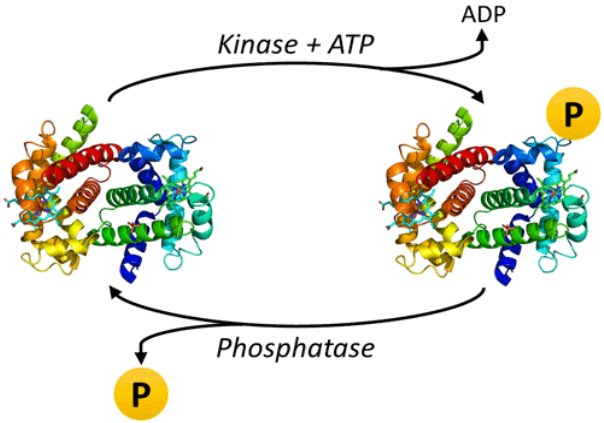
1/ How to gather sufficiently diverse MSA?
2/ How to go beyond binding affinity of natural peptides?

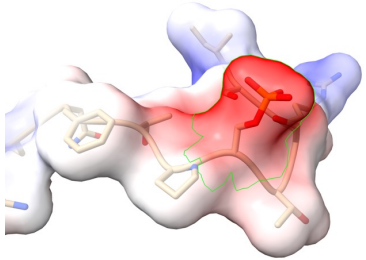
An integrative PPI inhibitor peptide design protocol



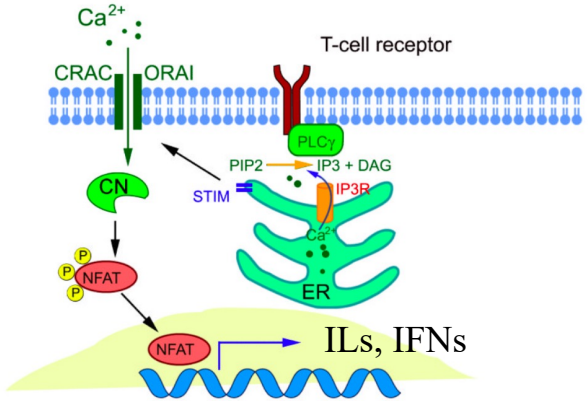
The Calcineurin (Cn) signaling pathway

Calcineurin (Cn) is a **calcium-dependent phosphatase** involved in multiple health & disease pathways.

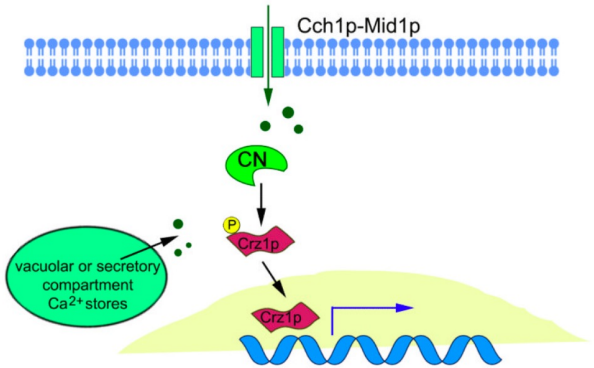




Phosphoserine (PDB: 1t29)



NFAT-mediated T-cell activation (Conserved in vertebrates)

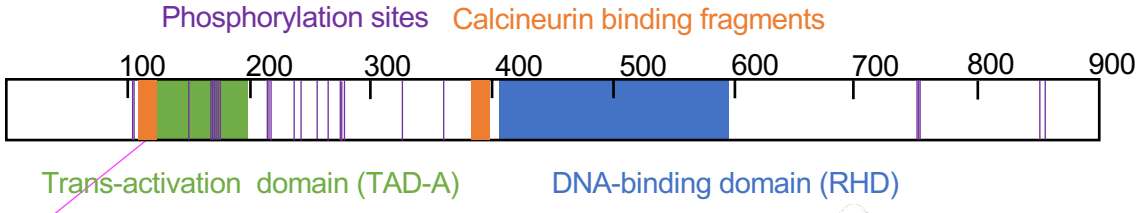


Crz1 environmental stress response pathway (Lower eukaryotes)

(Li et al. Trends Cell. Biology 2011)

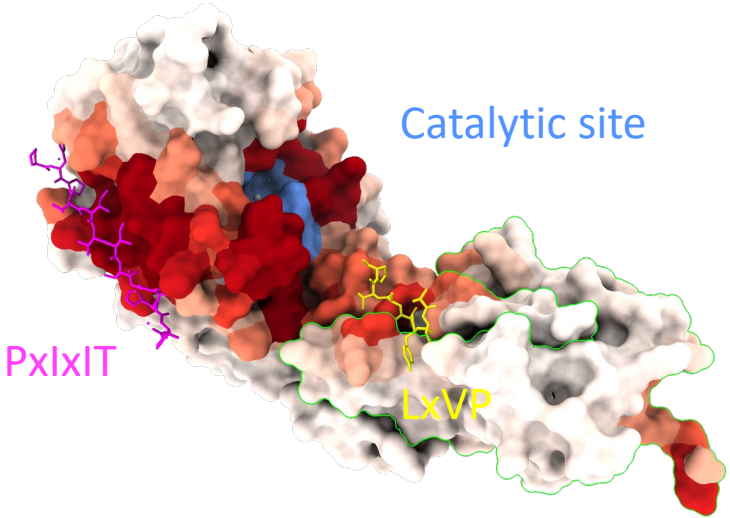
Structural basis of Calcineurin function

Most Cn substrates are disordered. They bind via two conserved short linear motifs (SLiMS)



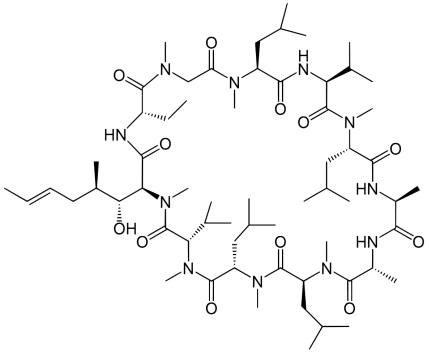
	PxIxIT
NFATc1	PALESPRIEITSCLGL
NFATc2	ASGLSPRIEITPSHEL
NFATc3	KPFECPSIQITSISPN
NFATc4	RVLECPSIRITSISPT
TRESK	ADEAVPQIIISAEELP
AKAP79	KRMEPIAIIITDTEIS
RCN1	GTTNTPSVIVHVCEDD
RIPOR2	SNSTNPEITITPAEFN
CAPN11	TFWTNPQFKISLPEGD
SFB3	RAYANPKFQFTYDSSV
CRZ1	AAPVTPIISIQEFNEG

Alignment of PxIxIT-containing fragments from substrates

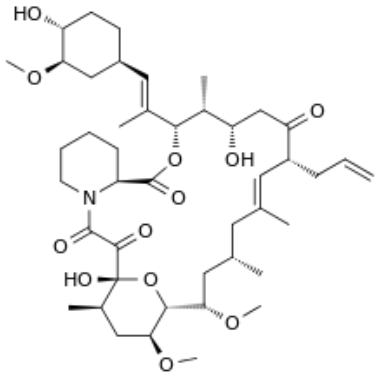


Calcineurin in **active conformation** bound to representative substrate fragments

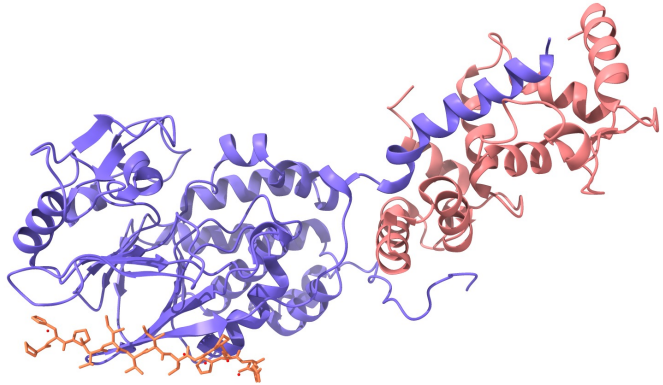
Calcineurin inhibitors



Cyclosporine A (CsA)



Tacrolimus (FK506)



The **PVIVIT** peptide
Aramburu et al. 1998 Science
(Combinatorial library + Display experiment)

Catalytic site inhibitors
Prescribed for transplantations since 80's

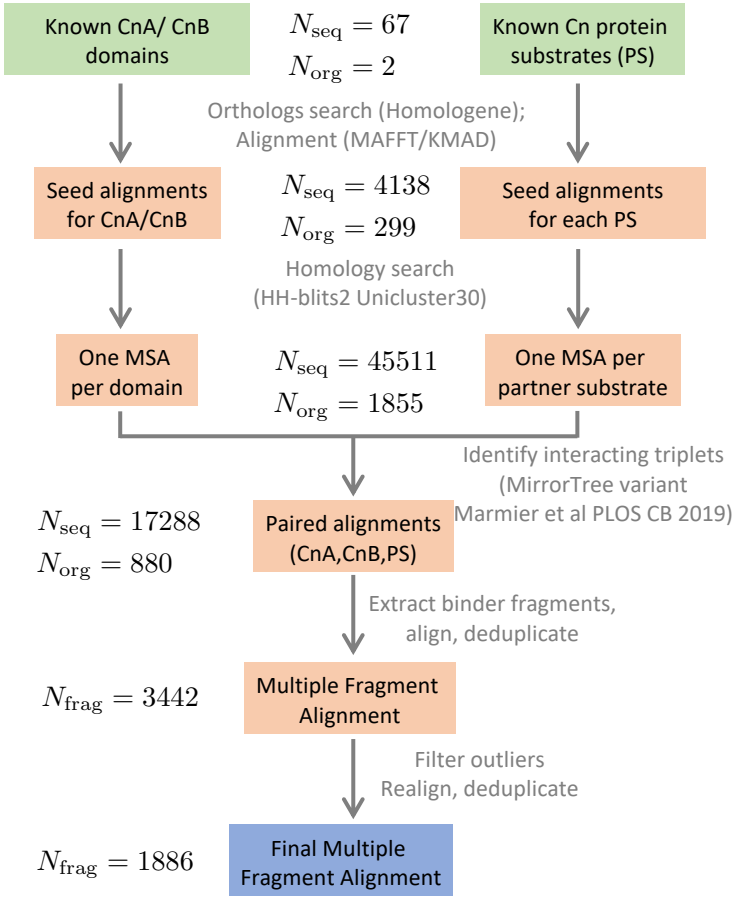
Protein Interaction inhibitor
Mice preclinical studies
(Noguchi et al. Nature Medicine 2004)

Step 1: Construction of an alignment of putative binding fragments

Input: a list of 67 Calcineurin substrates.
Sources: integrative high-throughput experiments

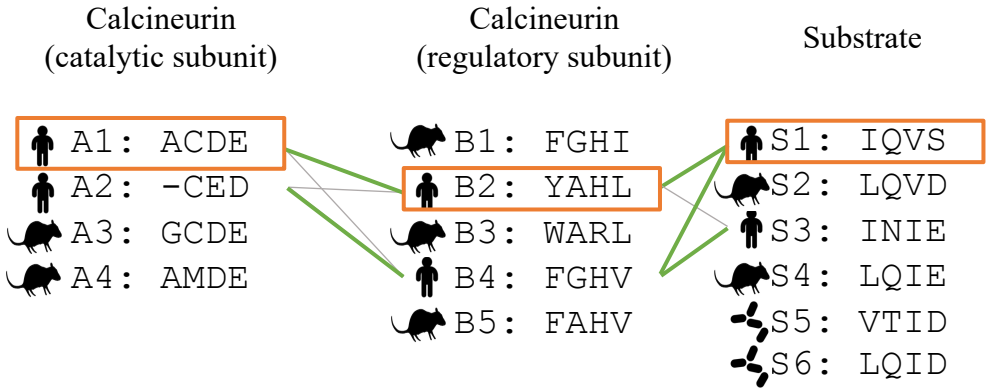
- Goldman A, Roy J, Bodenmiller B, Wanka S, Landry CR, Aebersold R, et al. The calcineurin signaling network evolves via conserved kinase-phosphatase modules that transcend substrate identity. *Mol Cell. 2014*
- Wigington CP, Roy J, Damle NP, Yadav VK, Blikstad C, Resch E, et al. Systematic Discovery of Short Linear Motifs Decodes Calcineurin Phosphatase Signaling. *Mol Cell. 2020 (Cyert lab)*

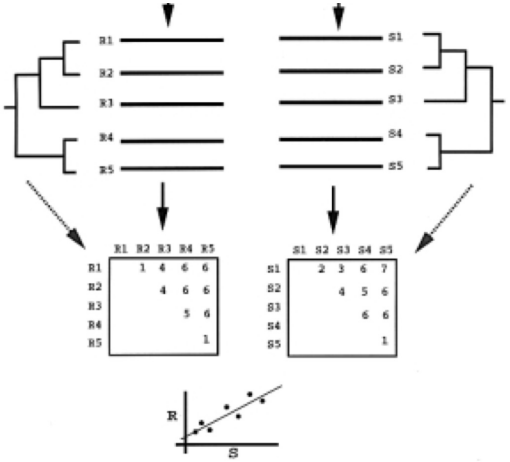
Gene	Organism	SLIM
NFATC1	Homo Sapiens	PRIEIT
NFATC2	Homo Sapiens	PRIEIT
NFATC3	Homo Sapiens	PSIQIT
NFATC4	Homo Sapiens	PSIRIT
TRESK	Homo Sapiens	PQIIIS
CRZ1	Saccharomyces Cerevisiae	PIISIQ
RCN1	Saccharomyces Cerevisiae	GAITID
SFB3	Saccharomyces Cerevisiae	PKFQFT
RGA2	Saccharomyces Cerevisiae	PQVLVS
ROD1	Saccharomyces Cerevisiae	PQIKIE
STE12	Saccharomyces Cerevisiae	PALSFS
RTS1	Saccharomyces Cerevisiae	PVLTVT
SLM1	Saccharomyces Cerevisiae	PNIYIQ
SLM2	Saccharomyces Cerevisiae	PEFYIE
RPL4A	Saccharomyces Cerevisiae	PQVTVH
RCN2	Saccharomyces Cerevisiae	PSITVN
DIG2	Saccharomyces Cerevisiae	PALNFS



The MSA pairing problem

Cn-substrate interactions are not systematically conserved across homologs





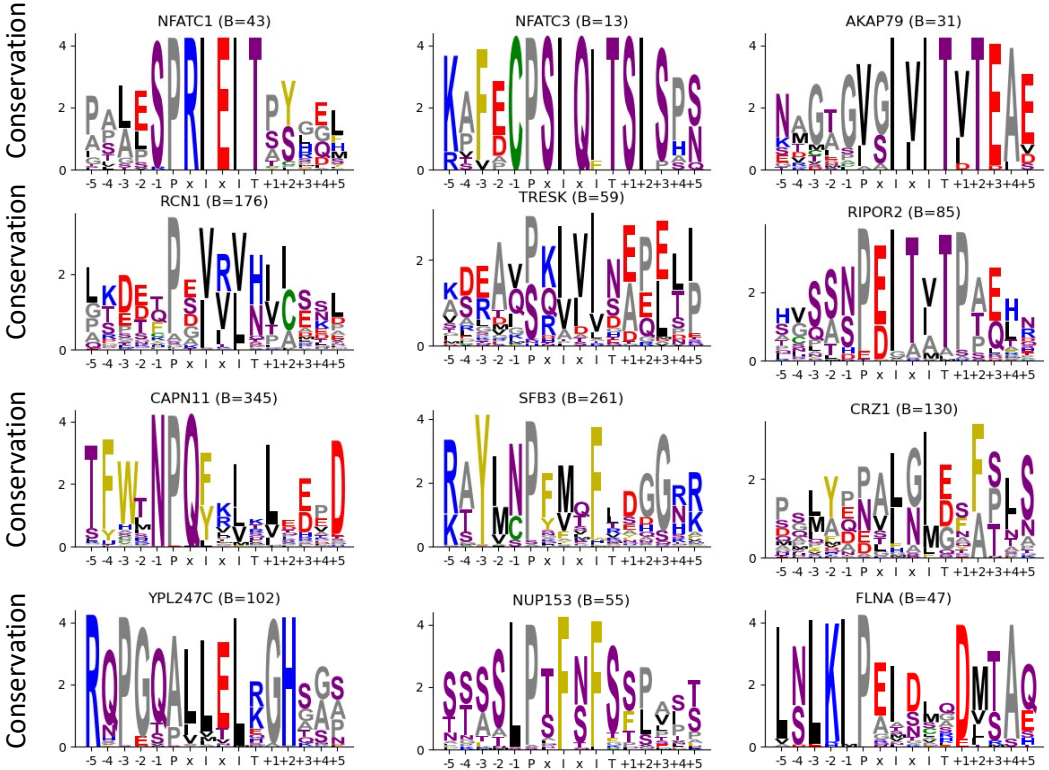
MirrorTree method
(Pazos and Valencia 1994)

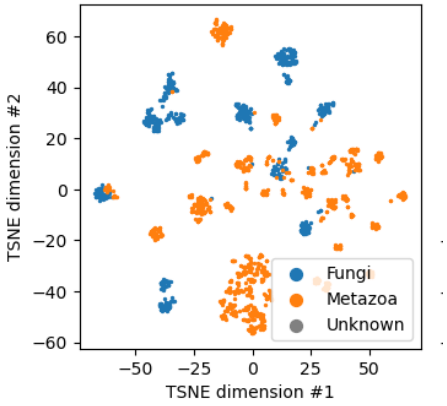
Find pairing that maximizes key fingerprints of interacting proteins:

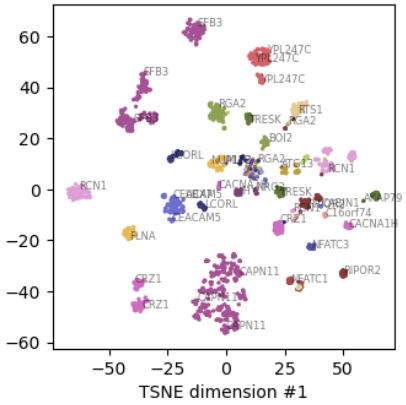
- Interacting partners sequences tend to mutate at similar rates
- Binding sites tend to coevolve

Marmier et al PLOS CB 2019)

Calcineurin-binding fragments are highly diverse







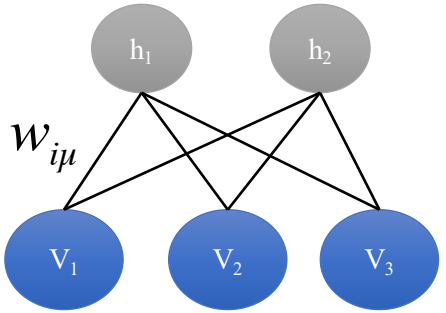
Step 2: Sequence Generative Modeling (compositional Restricted Boltmann Machines)

Graphical model constituted by two coupled sets of random variables

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp\left[-E(\mathbf{v},\mathbf{h})\right]$$

$$E(\mathbf{v},\mathbf{h}) = -\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{N}} \mathbf{g_i}(\mathbf{v_i}) + \sum_{\mu=\mathbf{1}}^{\mathbf{M}} \mathcal{U}_\mu(\mathbf{h}_\mu) - \sum_{\mathbf{i},\mu} \mathbf{w_{i\mu}}(\mathbf{v_i})\mathbf{h}_\mu$$

Hidden (Representation) layer



Visible (Data) layer

Ackley Sejnowski Hinton 1985
Smolensky 1986

The marginal defines a probability distribution over the data space

$$P(v) = \int \prod_\mu dh_\mu P(\mathbf{v},\mathbf{h}) = \frac{\mathbf{1}}{\mathbf{Z}} \exp\left[\sum_{\mathbf{i}} \mathbf{g_i}(\mathbf{v_i}) + \sum_{\mu} \mathbf{\Gamma}_\mu \left(\sum_{\mathbf{i}} \mathbf{w_{i\mu}}(\mathbf{v_i})\right)\right]$$

Trainable, non-quadratic function (generalizes over pairwise models)

Sparse weight matrix (Interpretability, compositionality)

The conditional distribution defines a representation of data

$$\langle \mathbf{h}_\mu | \mathbf{v} \rangle = \Gamma'_\mu \left[\sum_i w_{i\mu}(\mathbf{v}_i)\right]$$

Tubiana, Monasson PRL 2017
Tubiana, Cocco, Monasson eLife 2019
Tubiana, Cocco, Monasson Neur. Comp. 2019

Training algorithm for RBM

Data set: $\{\mathbf{v}^1, \mathbf{v}^2, \ldots \mathbf{v}^B\}$

Want to maximize log-likelihood: $\mathcal{L} = \sum_b \log P\left(\mathbf{v}^b | \{w_{i\mu}\}, \{g_i\}, \{\mathcal{U}_\mu\}\right)$

parameters Θ

Stochastic gradient ascent:

$$\log P(\mathbf{v}) = -E_{\text{eff}}(\mathbf{v}) - \log Z$$

$$\frac{\partial \mathcal{L}}{\partial \Theta_a} = \left\langle \frac{\partial E_{\text{eff}}(\mathbf{v}|\boldsymbol{\Theta})}{\partial \Theta_a} \right\rangle_{\mathbf{v}\sim\text{RBM}} - \left\langle \frac{\partial E_{\text{eff}}(\mathbf{v}|\boldsymbol{\Theta})}{\partial \Theta_a} \right\rangle_{\mathbf{v}\sim\text{Data}}$$

Moment Matching Equations

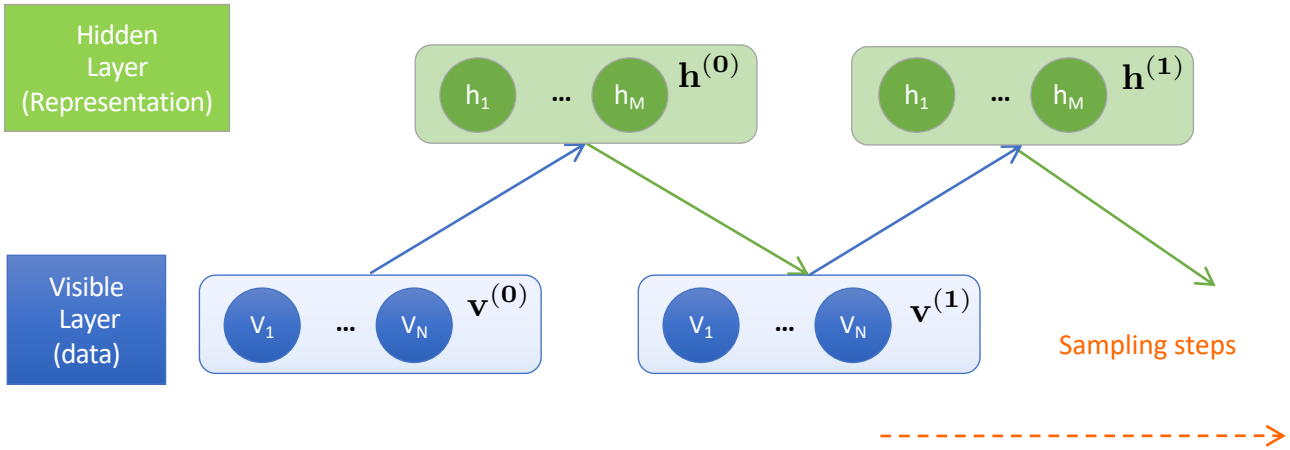
Requires MCMC sampling

Computed directly from data

Learning algorithms : Boltzmann Machine Learning (Ackley Hinton Sejnowski 1985), PCD (Tieleman Hinton 2008)

Sampling from RBM

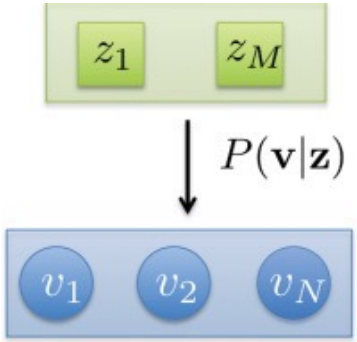
- Compute Hidden units Inputs $I_\mu = \sum_i w_{i\mu} v_i$
- Sample each hidden unit independently $P(h_\mu|I_\mu) \propto \exp\left[-\mathcal{U}_\mu(h_\mu) + h_\mu I_\mu\right]$
- Compute the visible layer inputs $I_i = \sum_\mu w_{i\mu} h_\mu$
- Sample each visible unit independently $P(v_i|I_i) \propto \exp\left[(g_i + I_i)\, v_i\right]$



Directed Latent variables model

- PCA, ICA
- Sparse dictionaries
- Variational Autoencoders

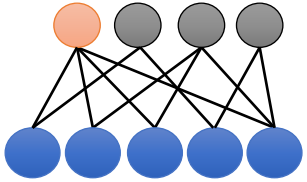
$$P(\mathbf{z}) = \prod_\mu P_\mu(z_\mu)$$



Compositional Restricted Boltzmann Machines

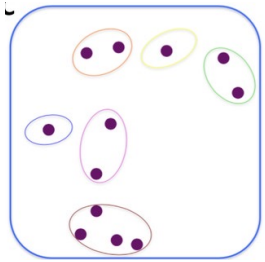
For latent variable generative models, hidden unit distribution guides **weight interpretation** & **extrapolation regime**. For RBMs, it is unspecified.

P(**h**)



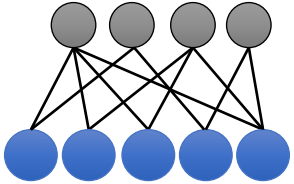
$$\langle \mathbf{v}|\mathbf{h}\rangle = \mathrm{Softmax}(w^T h)$$

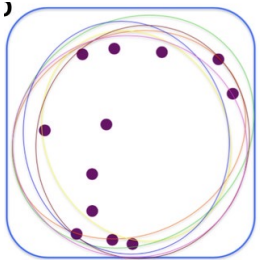
P(**v**)



Prototype (Ferromagnetic) representation

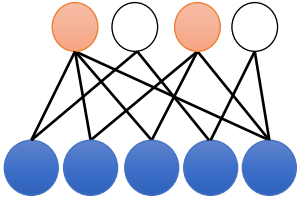
Weights are prototype configurations

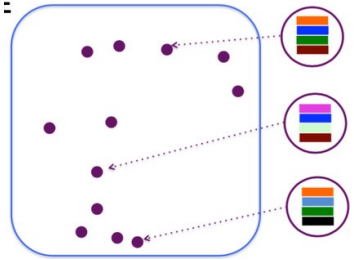




Intricate (Spin-glass) representation

No clear weight interpretation





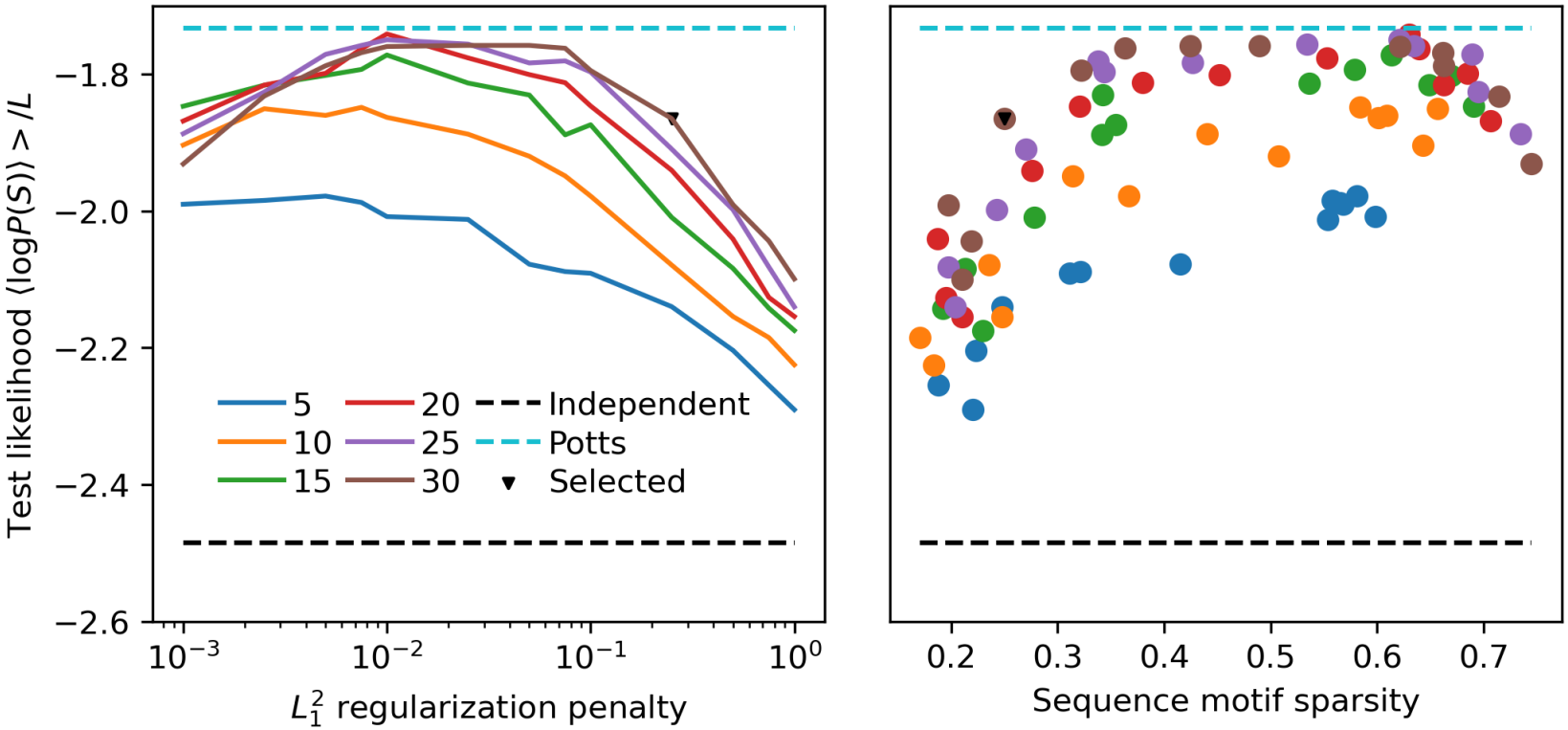
Compositional Representation

Weights are parts of configurations

Sparse weights + unbounded, non-quadratic potentials
⇒ compositional representation

Tubiana, Monasson PRL 2017
Tubiana, Cocco, Monasson Neur. Comp. 2019

Learning cRBM: the interpretability-accuracy trade-off

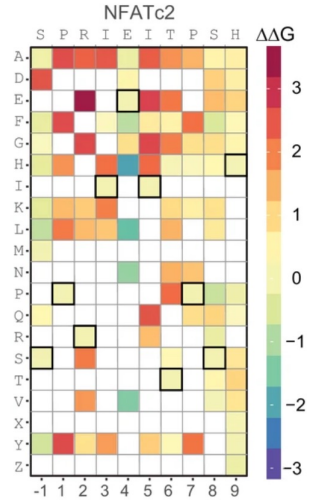


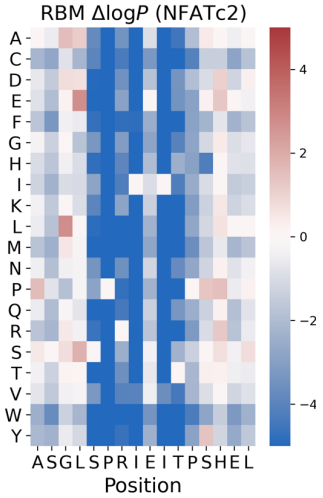
Practical considerations for learning cRBMs

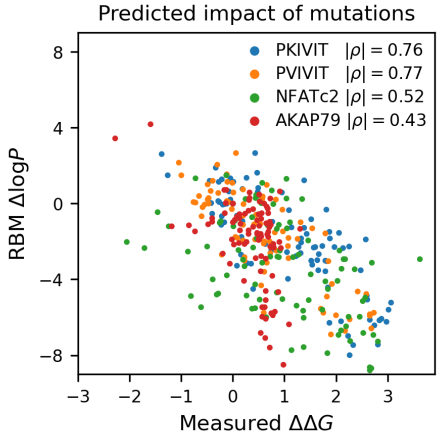
https://github.com/jertubiana/PGM

- Objective function: Maximum likelihood + regularization penalties
- Sampling algorithm: MCMC, PCD.
- Optimizer: RMSprop (adaptive learning rates, improves convergence rates).
- Hidden unit potential: dReLU (adaptive non-linearity, for fitting non-gaussian distributions)
- Parameterization: Batch normalization (improves hessian conditioning, promotes homogeneity).
- Regularization: L_2 on fields, L_1^2 on weights (promotes sparsity+homogeneity).
- Partition function estimation: Annealed Importance Sampling.

SGM predicts binding affinity changes upon mutation

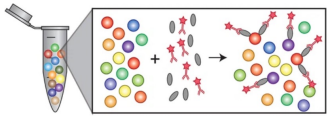




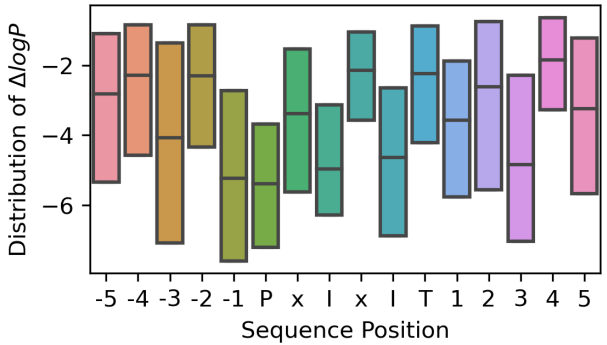


Model	PKIVIT	PVIVIT	NFATc2	AKAP79
RBM	**0.76**	**0.77**	0.52	**0.43**
Independent	0.74	0.72	0.39	0.13
Rosetta flex_ddG	NA	-0.2	NA	0.21
FoldX ddG	NA	0.21	NA	0.39
ProteinMPNN	0.53	0.45	**0.53**	0.36

SGM better predicts ΔΔG measurements than baseline methods



$$\Delta \log P = \log P(Mut.) - \log P(WT)$$

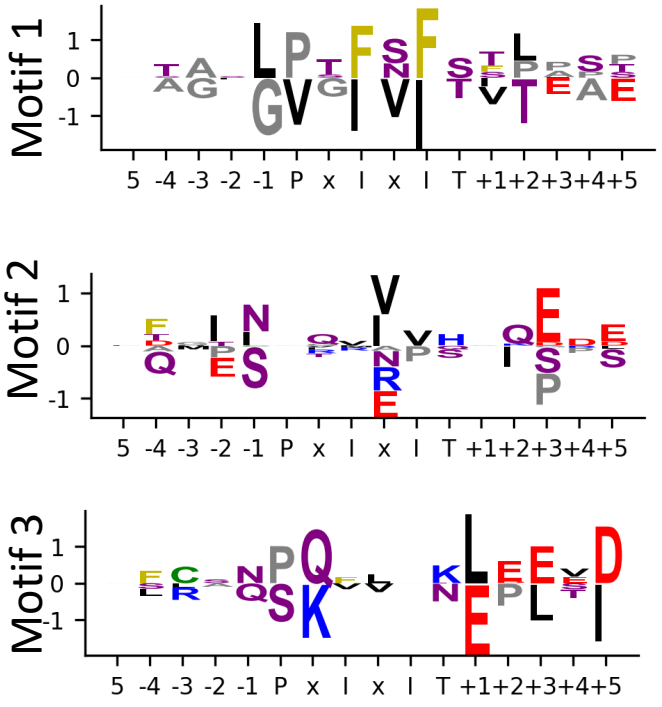


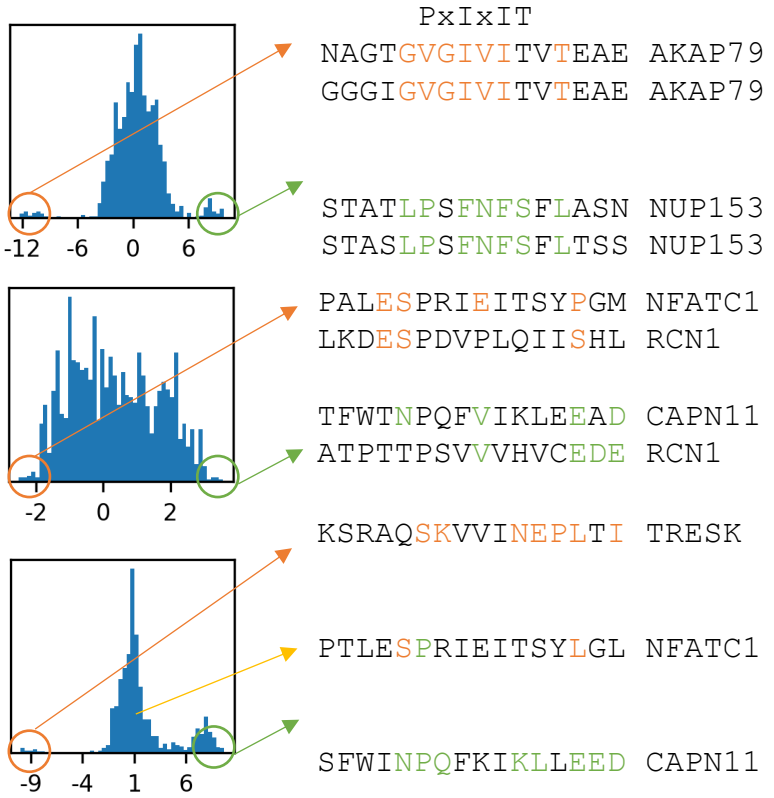
Quantitative mapping of protein-peptide affinity landscapes using spectrally-encoded beads
Nguyen et al eLife 2019 (Fordyce & Cyert labs)

ΔΔG predictions for flanking residues reveals important positions

SGM learns sequence motifs shared between evolutionary-unrelated binders

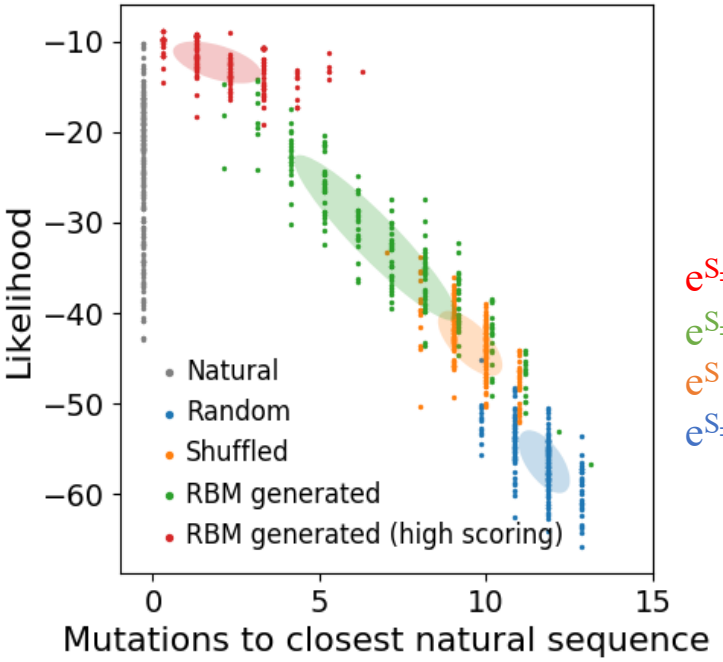
$$P(S) = \frac{1}{Z} exp[\sum_{i=1}^{N} g_i(s_i) + \sum_{\mu=1}^{M} \Gamma_\mu(\sum_{i=1}^{N} w_{i\mu}(s_i))]$$





Sequence motif w Activity distribution Top-activating fragments

Peptide Library generation



$e^S = 10^{2.8}$

$e^S = 10^{13.7}$

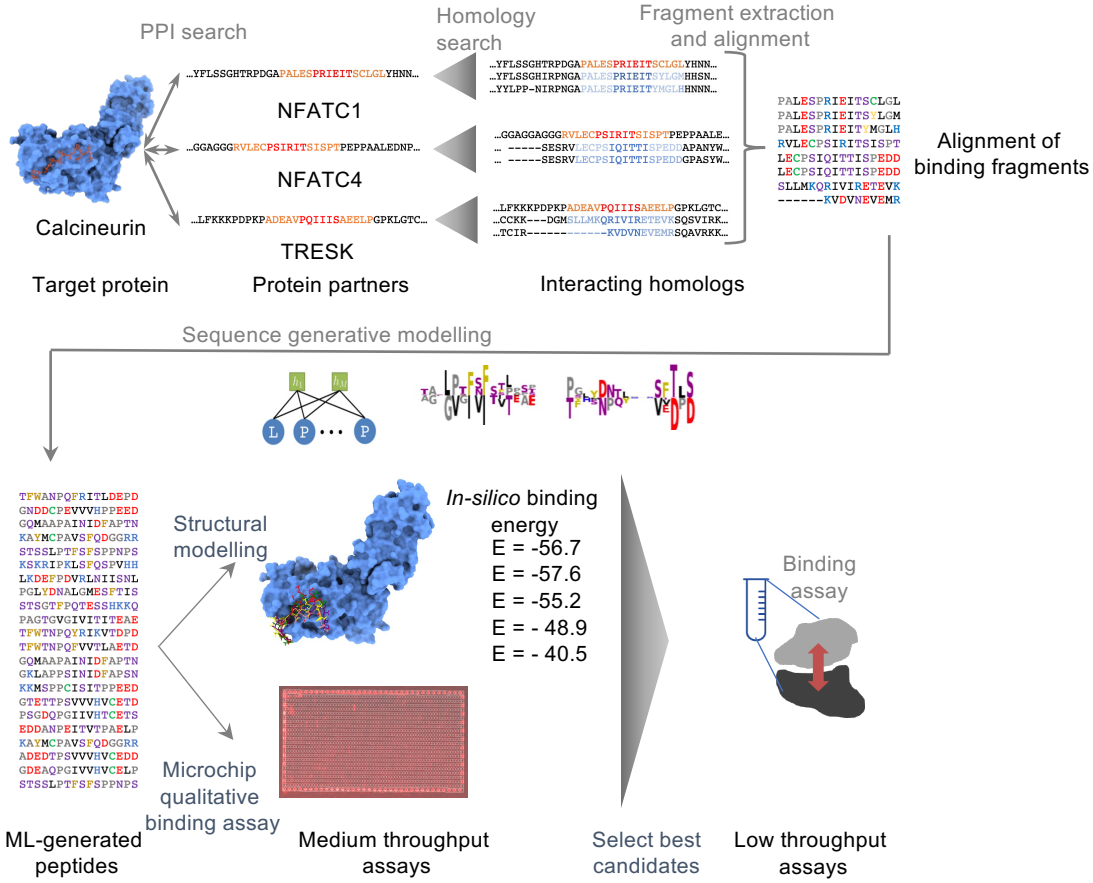
$e^S = 10^{17.3}$

$e^S = 10^{20.8}$

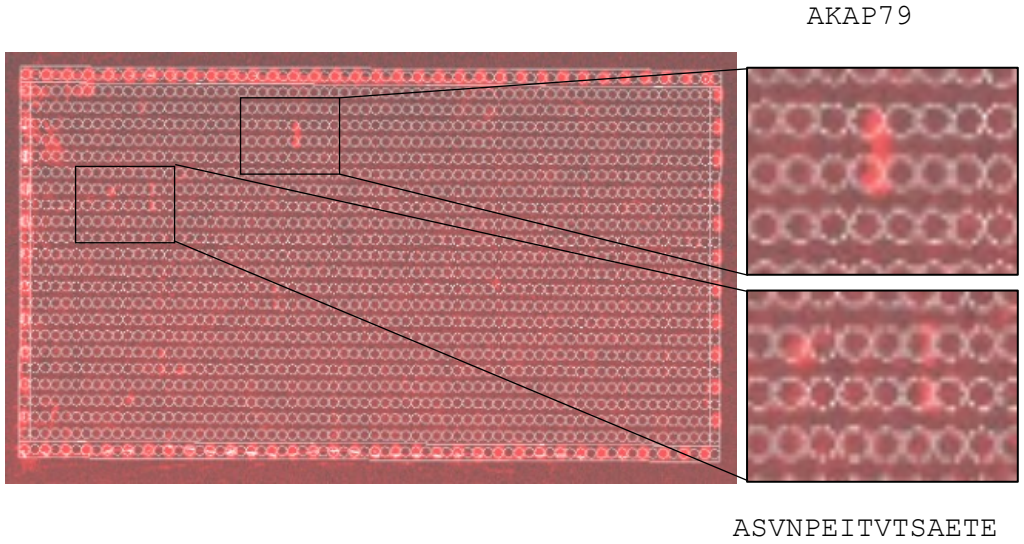
Source	Role	#Num sequences
Random peptides	Negative control	36
Literature designs	Positive control	2
Natural peptides	Positive (?)	75
Independent (PSSM)	Baseline design	72
cRBM, $\beta = 1$	Design	180
cRBM, $\beta = 2$	Design	361

Novelty/Diversity-Quality trade-off

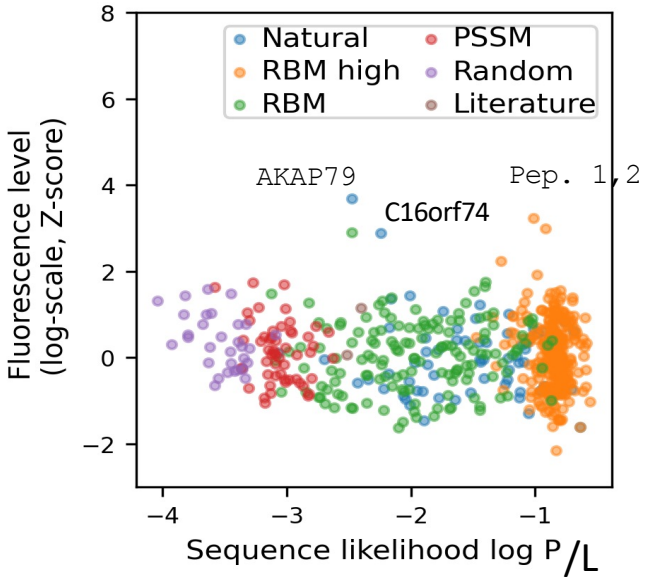
An integrative PPI inhibitor peptide design protocol



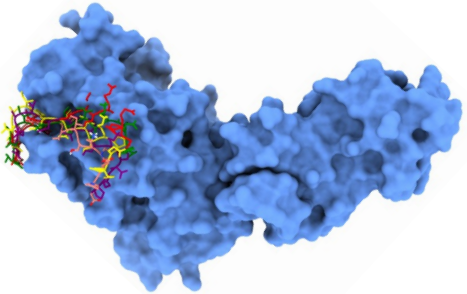
Step 3: Library filtering by microarray screening experiment

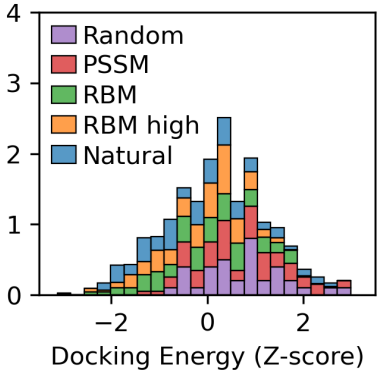


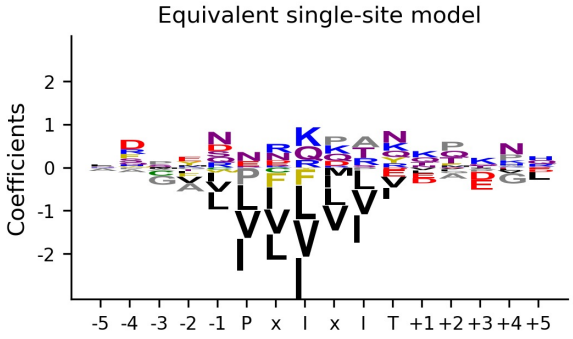
PepPerChip Microarray screening

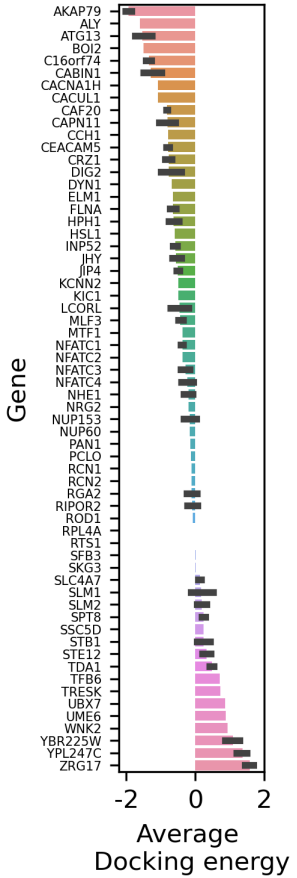


Step 3': Library filtering by molecular docking







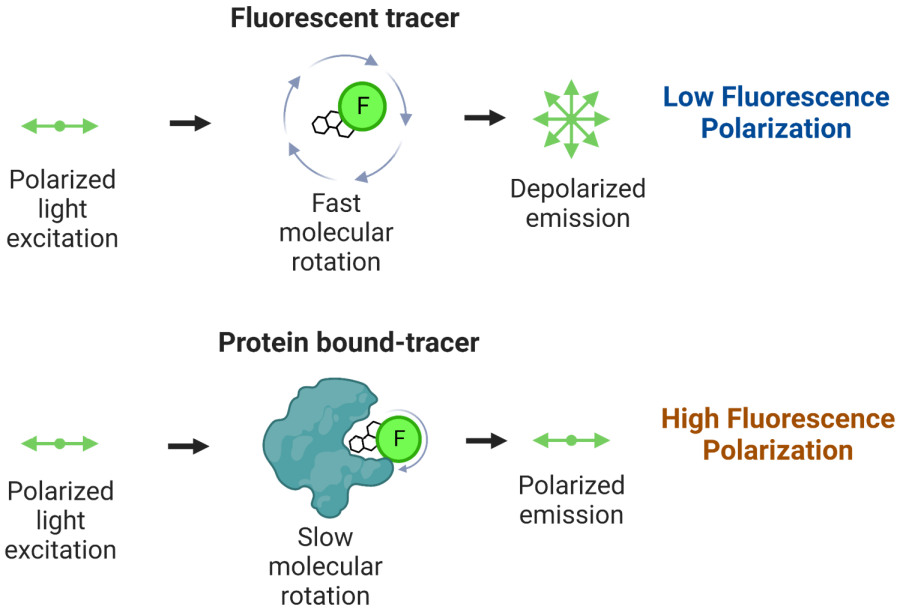


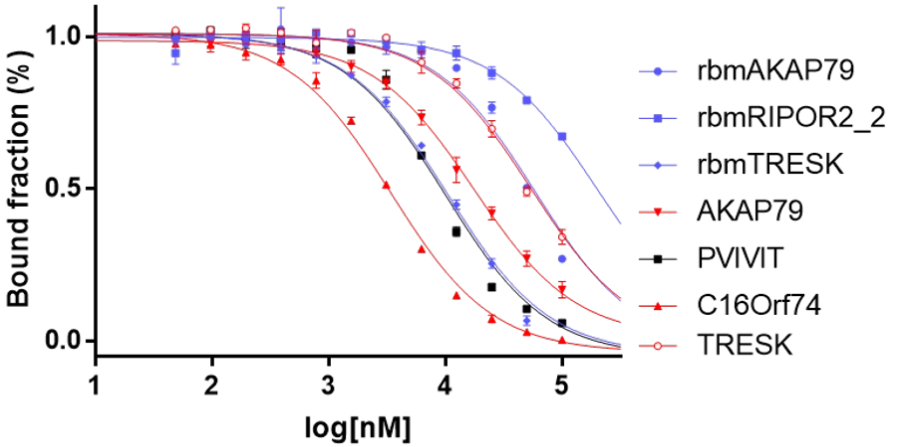
In-silico docking energy score:

- Ensemble of five bound crystal structures as templates
- Threading with Modeller
- Flexible refinement and scoring using PepCrawler (average of minimum energies).

PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors
Donsky and Wolfson Bioinformatics 2011

Step 4: Experimental validation by FP binding assay





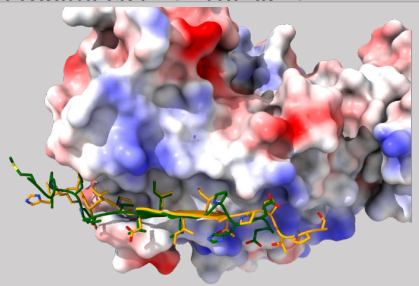
Credit: https://bpsbioscience.com/product-types/biochemical-assay-kits-by-format-type/fluorescence-polarization

Step 4: Experimental validation by FP binding assay

Name	Sequence	IC50 (uM)	Source	Closest natu sequence			
C16Orf74	KHLDVPDIIITPPTPT	**1.17**	Natural	KHLDV DIIITPI			
PVIVIT	MAGPHPVIVITGPHEE	10.2	Positive control	/		/	/
rbmTRESK	ADEAIPEIVISKPEEP	**14**	Designed (low T)	ADEAVPQIIISA			
AKAP79	KRMEPIAIIITDTEIS	17.5	Natural	KRMEPIAIIITD			
TRESK	ADEAVPQIIISAEELP	54	Natural	ADEAV QIIISAEELP	Homo sapiens	TRESK	0
rbmAKAP79	AAGAGVGIVITVTEAE	57	Designed (low T)	AAGAG VGIVITVTEAE	Pelecanus crispus	AKAP79	2
rbmTRESK_2	ADEAIPEITITSAELP	60	Designed (low T)	ADEAIPQITITA	Propithecus	TRESK	
rbmAKAP79_2	ADGAGVGIVITVTEAE	69	Designed (low T)	ADGAGVGIVITV			
rbmRIPOR2	ASVSNPEITVTSAETE	79	Designed	QSQSNPEITVTP			
rbmRIPOR2_2	HVSSSPRITITPTQHR	200	Designed (low T)	HVSSSPDITATPT			

Motif recombination:
ADEAIPEIVISKPEEP [Design]
ADEAIPQIVIDAGADE [TRESK, 50uM]
SPSNP**PEIVIS**SREDN [KCNN3]
TFWTNPQFKIYL**PEE**D [CAPN11]

Lack PxIxIT SLIM:
AAGAG**VGIVIT**VTEAE
ADGAG**VGIVIT**VTEAE



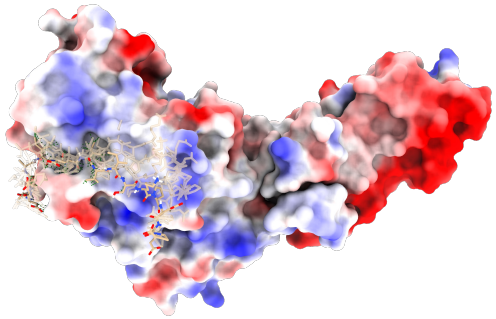
C-terminal polyproline helix expands interaction surface at no entropic cost

7/10 designs
3/4 natural compete with binding of substrate to Cn

Tubiana, Adriana-Lifshi* *23*

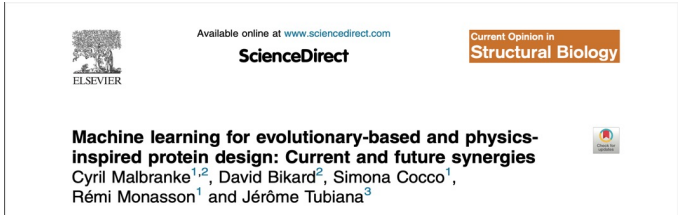
Summary and future directions

- Peptides are attractive candidates for PPI inhibitor design, but design is challenging.
- We proposed and validated an integrative design protocol based on a Sequence Generative Model trained from native binders of the target protein.
- The SGM captures key sequence patterns important for binding, and recombines them to generate novel and diverse peptide binders.
- Flexible molecular docking efficiently complements SGMs by differentiating between weak and strong native binders.



Next steps for Calcineurin:

- Cellular assays
- Display experiments to further optimize binding affinity, multivalent constructs
- Pharmacophore-based drug discovery / HTS via competition assays



Acknowledgements

Tel Aviv University
Haim Wolfson
Mark Rozanov
Naama Hurwitz
Michael Nissan
Yoav Lotem

Maayan Gal
Lucia Lifshits
Daniel Bar

TAU CS system team
Sonia Lichtenzveig Sela

The Hebrew University of Jerusalem
Dina Schneidman-Duhovny
Merav Breitbart
Lirane Bitton
Matan Halfon
Shon Cohen
Tomer Cohen
Edan Patt

Ecole Normale Supérieure
Rémi Monasson
Simona Cocco

Icahn School of Medicine Mount Sinai NY
Yi Shi
Yufei Xiang
Zhe Sang

University of Pittsburgh
Kong Chen
Li Fan