LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

cnrs

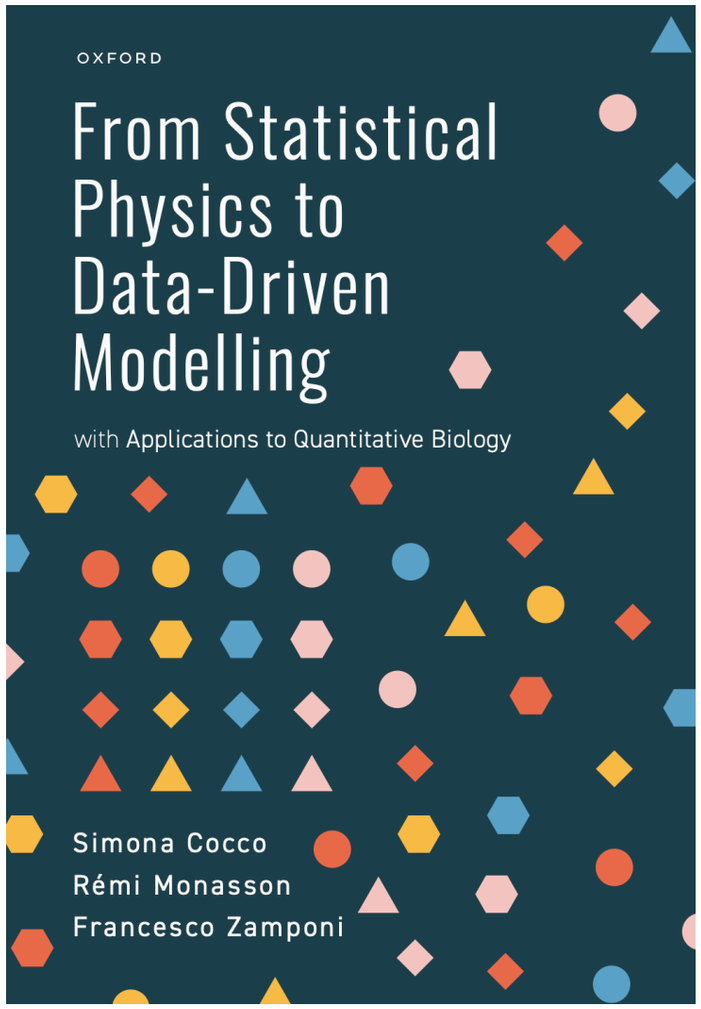
Rough fitness landscapes: from protein evolution to protein design

Francesco Zamponi

Gulliver, ESPCI, October 17, 2022

SIMONS FOUNDATION

erc



Background

2016-2021: teaching with S.Cocco and R.Monasson

- Bayesian and high-dimensional inference
- Regularization
- Graphical models
- Supervised and unsupervised learning
- Time series analysis

With coding tutorials using real biological data

Oxford University Press, 2022
Promotion code: ASPROMP8

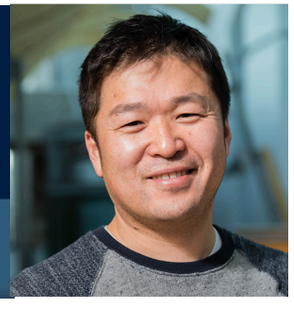
SORBONNE UNIVERSITÉ

Since 2017: collaboration with Martin Weigt's group

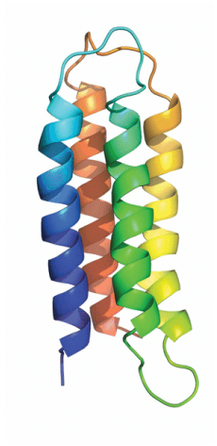
Since 2021: collaboration with Nobuhiko Tokuriki's lab

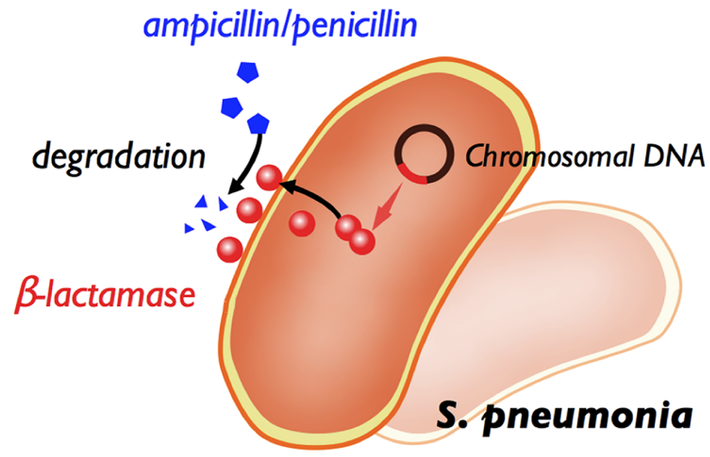
THE UNIVERSITY OF BRITISH COLUMBIA

Tokuriki Lab - Michael Smith Laboratories

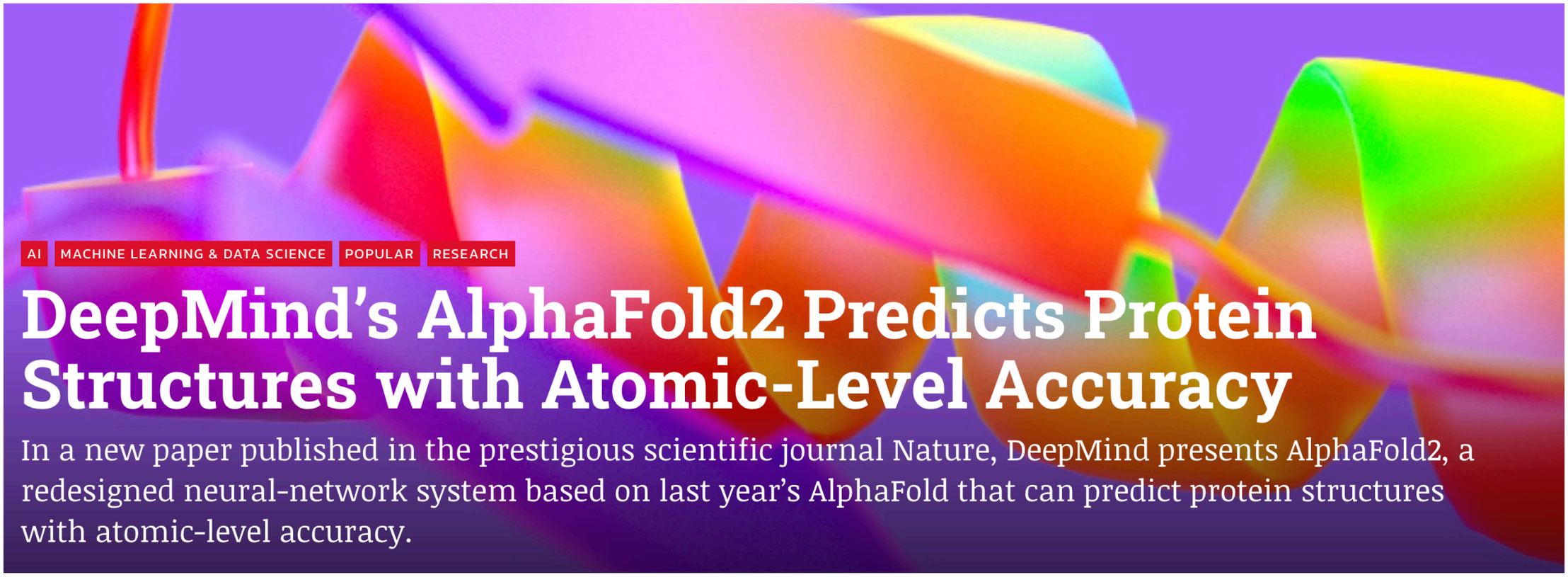


Sequence-to-function paradigm





KPDLKPYSPLRDK

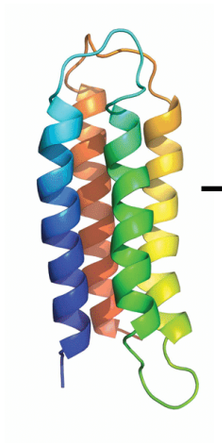


AI MACHINE LEARNING & DATA SCIENCE POPULAR RESEARCH

DeepMind's AlphaFold2 Predicts Protein Structures with Atomic-Level Accuracy

In a new paper published in the prestigious scientific journal Nature, DeepMind presents AlphaFold2, a redesigned neural-network system based on last year's AlphaFold that can predict protein structures with atomic-level accuracy.

Inverse problem: protein design



Desired structure and function

Black box (Generative model)

KPDLKPYSPLRDK

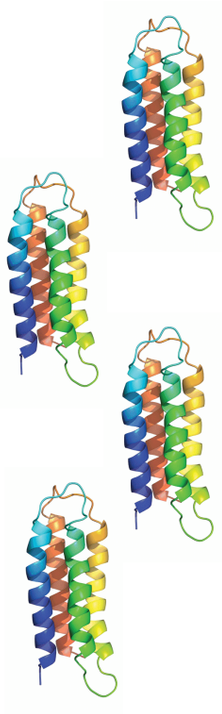
REDLVNYNPITEK

QNDLKEYSPVDEK

AADLVANSPVTGK

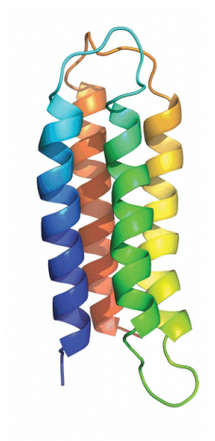
Generated artificial sequences

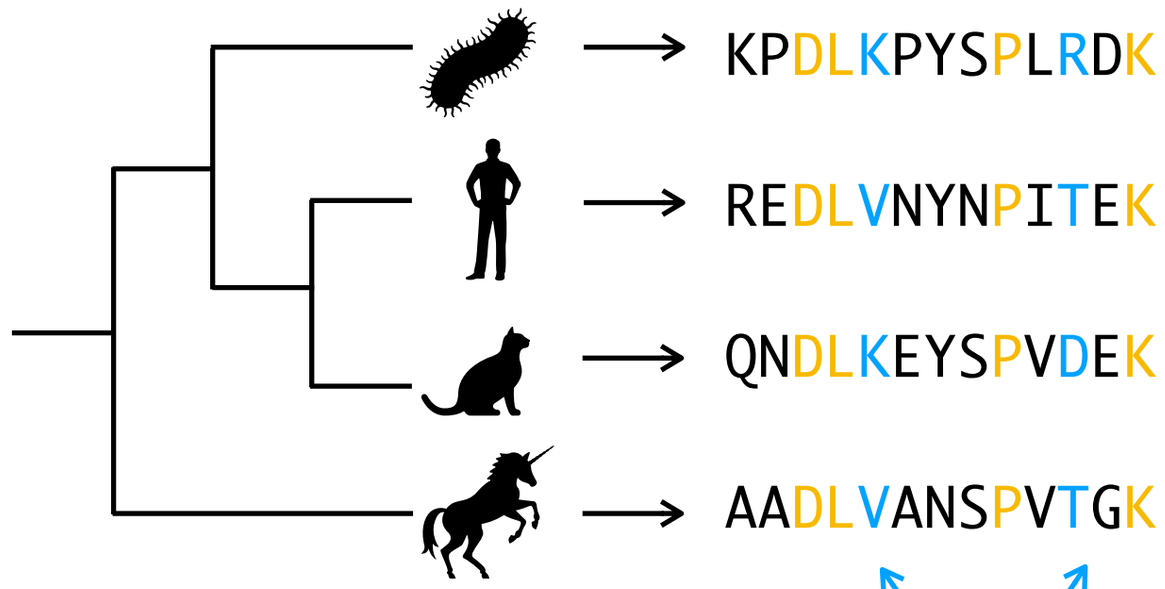
Same structure and function!



Data #1: natural protein sequences

Ancestral protein



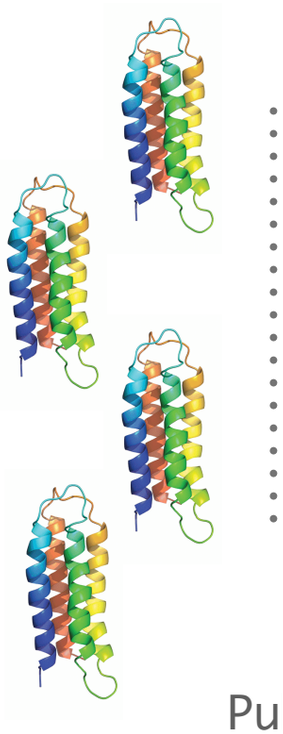


Natural evolution

Conservation

Coevolution

Same structure and function!

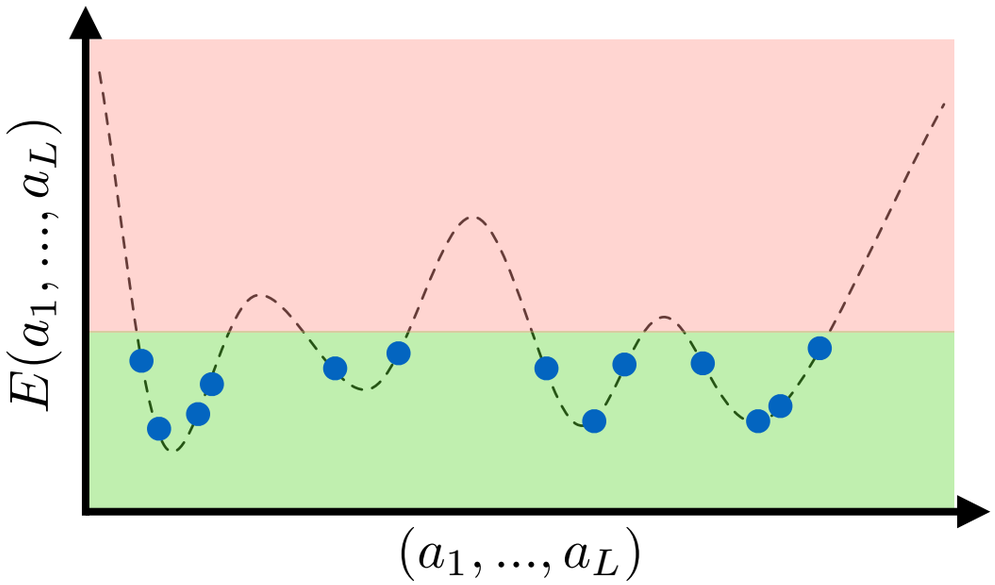


Available sequences: 10^2-10^5

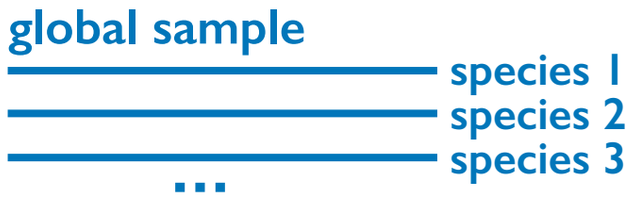
Sequence divergence: 70-80%

Public databases (PFAM, UniProt..)

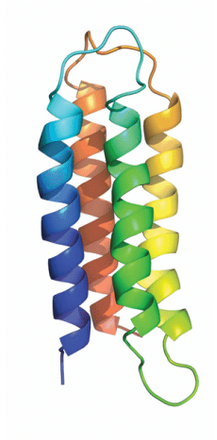
sequence landscape



sequence data



Data #2: deep mutational scan



APDLKPYSPLRDK

KADLKPYSPLRDK

KPALKPYSPLRDK

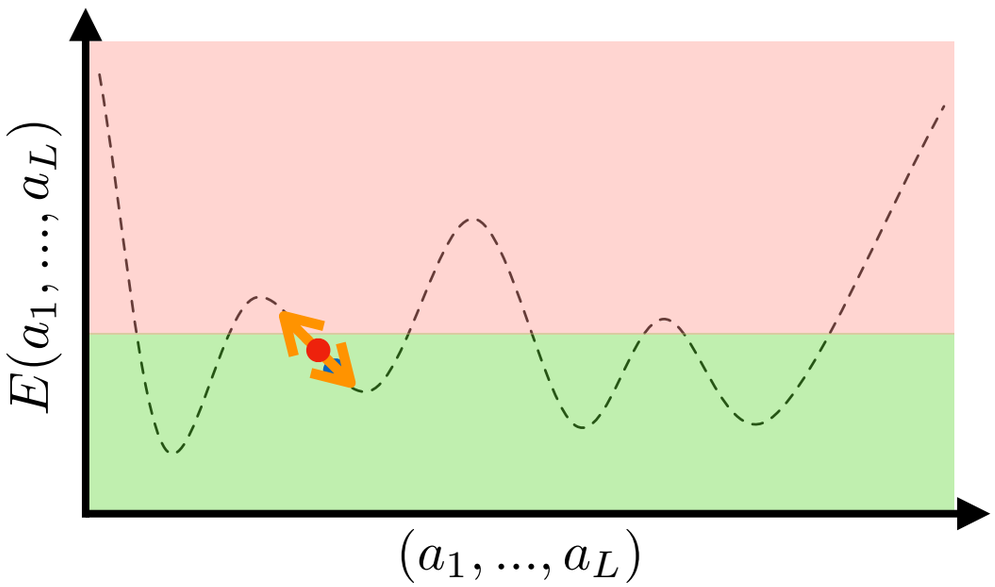
...

KPDLKPYSPLRDA

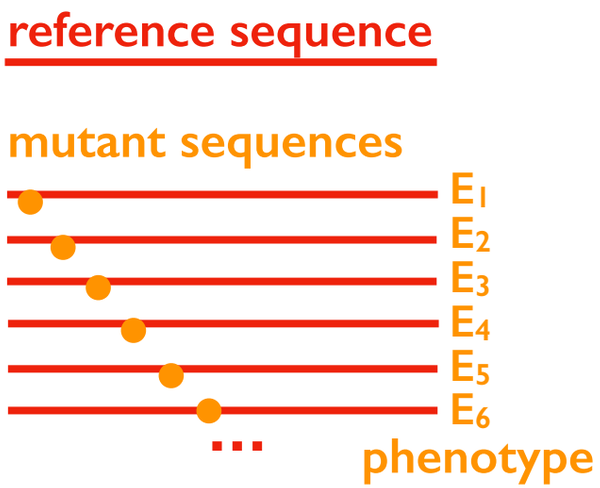
Available sequences: $\sim 10^4$

Sequence divergence: 1

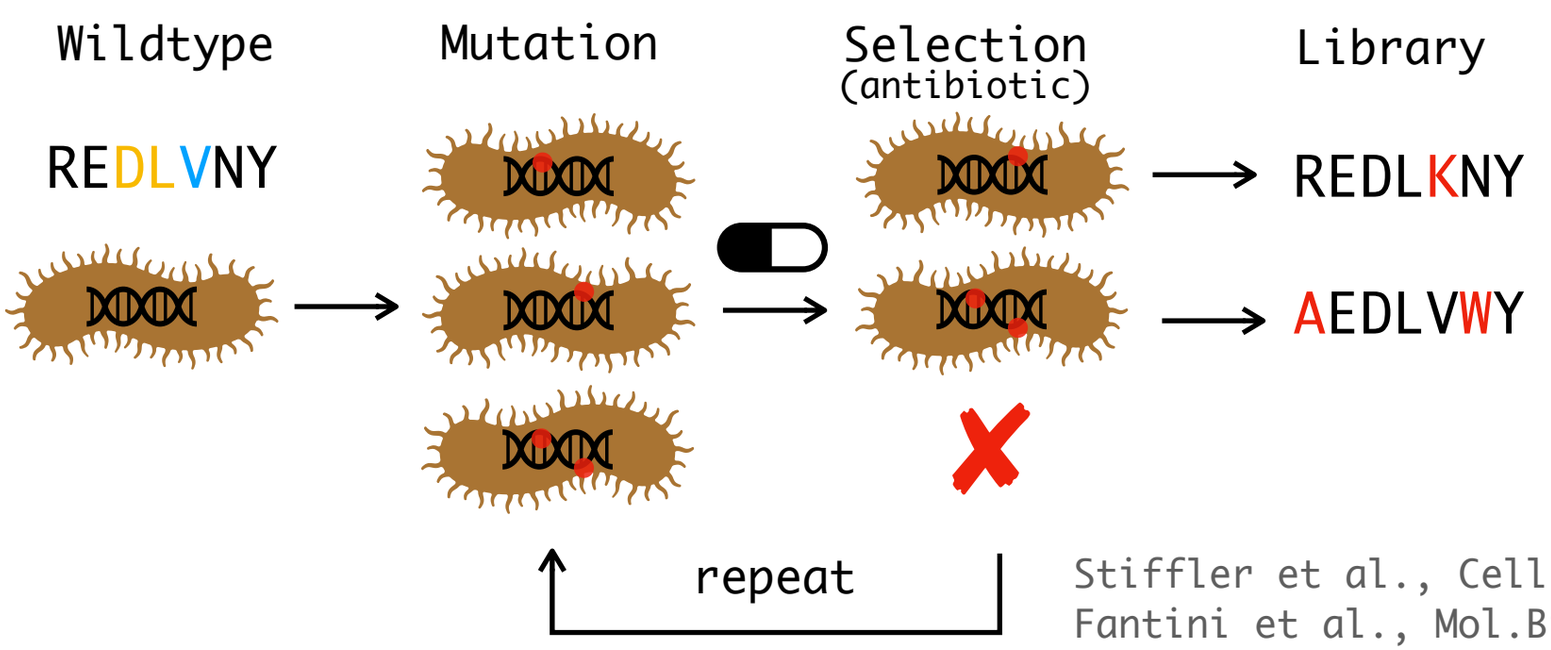
sequence landscape



sequence data



Data #3: in vitro evolution



Available sequences: 10^5-10^6

Sequence divergence: 10-20%

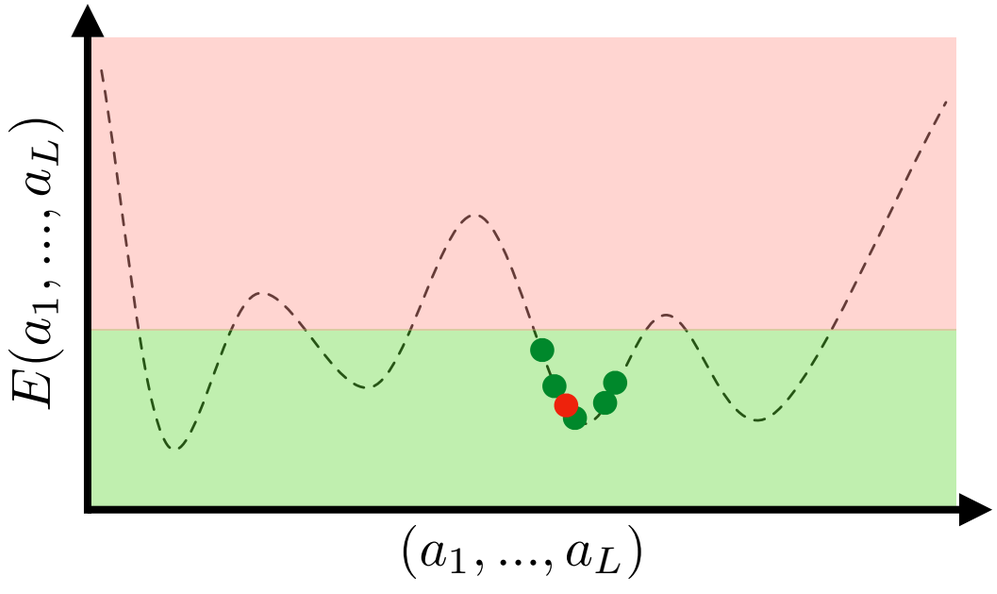
Natural evolution

Available sequences: 10^2-10^5

Sequence divergence: 70-80%

Stiffler et al., Cell Systems (2020)
Fantini et al., Mol.Biology and Evo. (2020)

sequence landscape



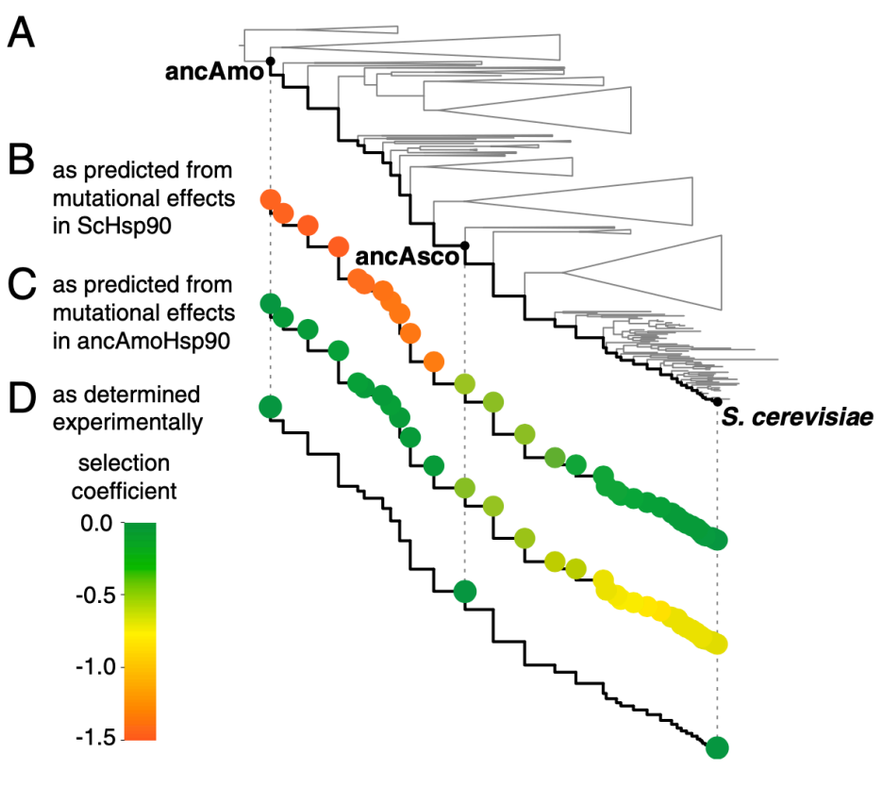
sequence data

reference sequence

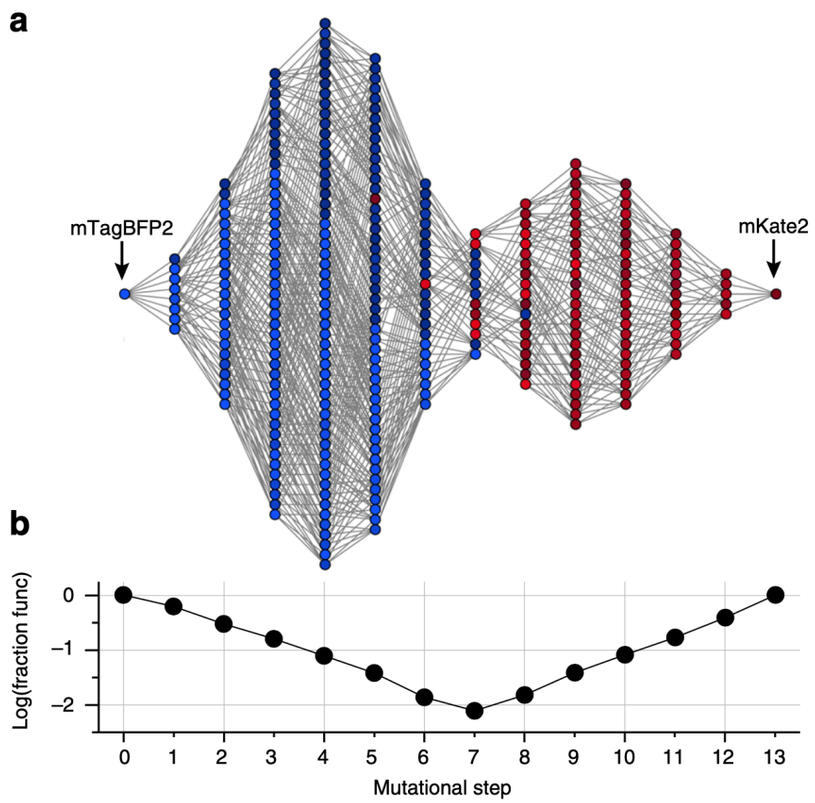
local sample

mutant 1
mutant 2
mutant 3
...

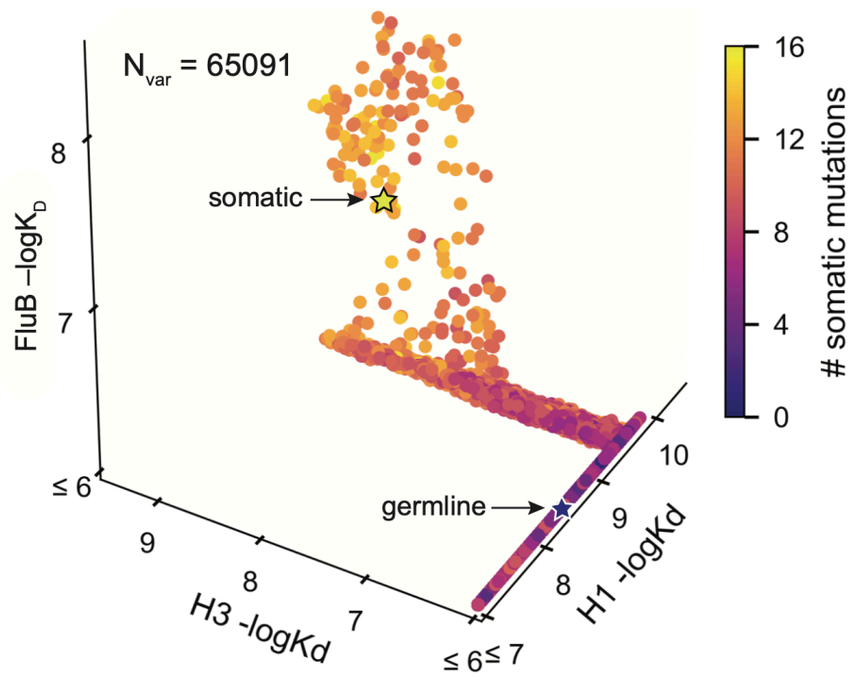
Data #4: path/phylogeny reconstruction



Starr et al. PNAS (2017)

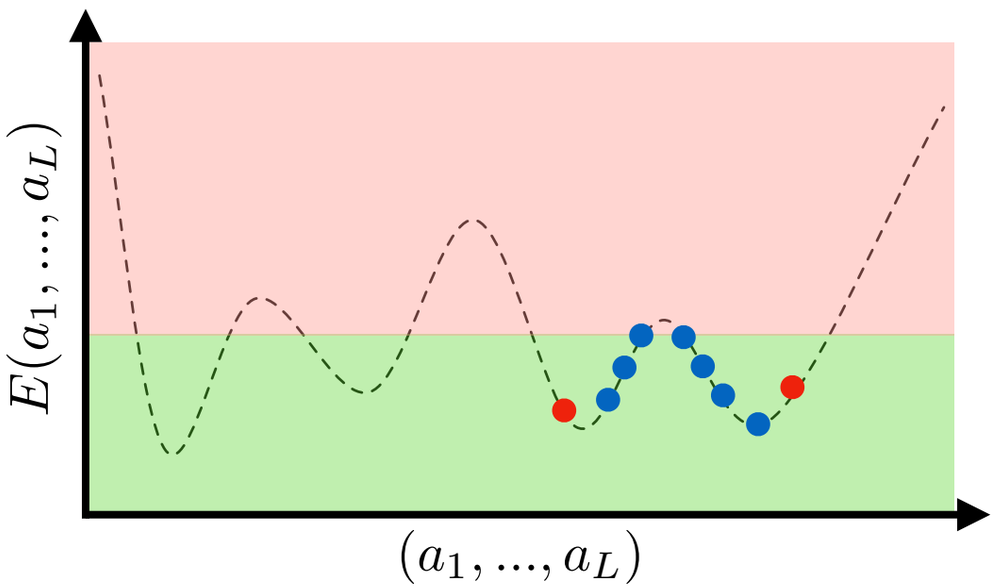


Poelwijk et al. Nat.Comm. (2019)



Phillips et al. eLife (2021)

sequence landscape



sequence data

reference sequence

local sample

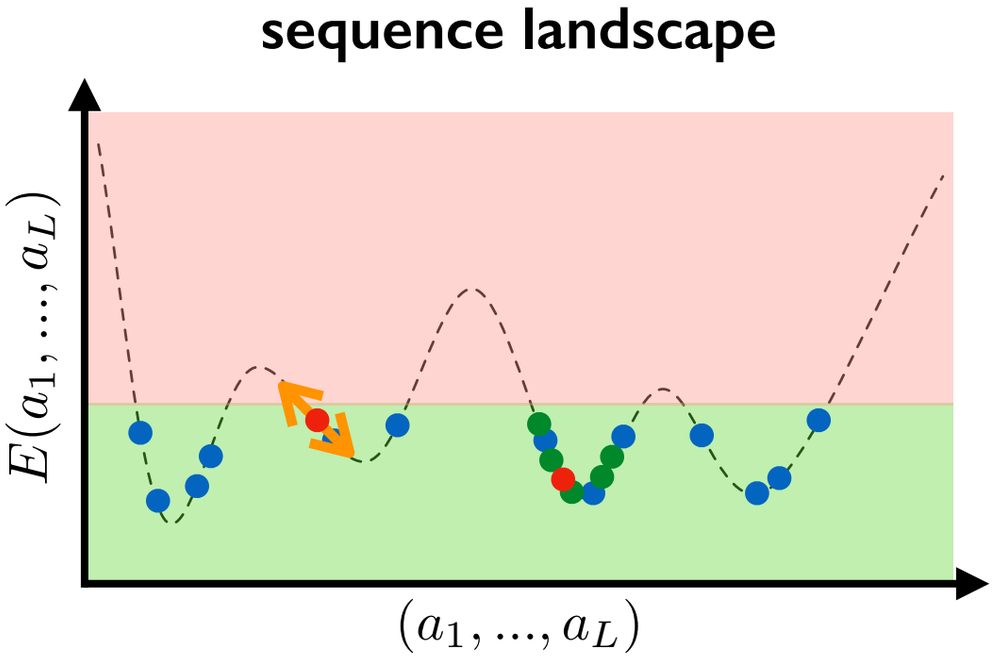
path 1

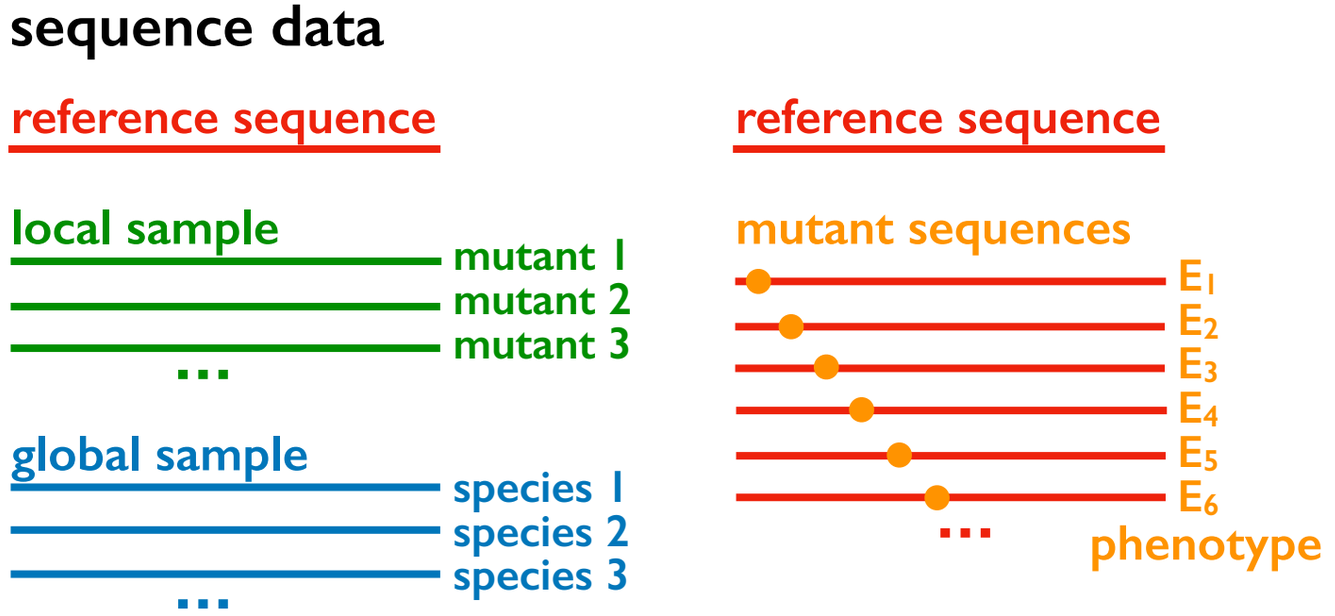
path 2

path 3

...

Massive amount of data!





Big data-driven questions:

- Generate artificial sequences
- Generate artificial sequences with desired properties
  (binding affinity, fluorescence, antibiotic resistance, catalytic activity…)
- Model in vitro evolution (controlled environment)
- Optimize evolution protocols (antibiotic concentration, vaccination protocols…)
- Understand natural evolution (phylogeny, fluctuating environment)
- Generate paths of artificial sequences connecting two natural ones

Generative modeling

Many images of human Faces

Black box
(Generative model)
$P(image)$

this-person-does-not-exist.com

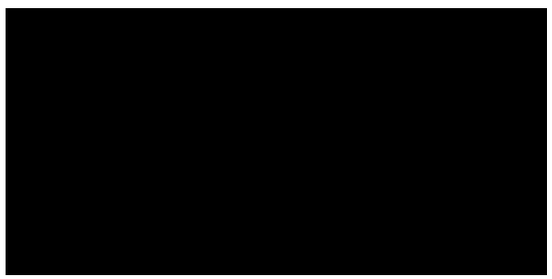
KPDLKPYSPLRDK

REDLVNYNPITEK

QNDLKEYSPVDEK

AADLVANSPVTGK

Alignment of $10^2 - 10^5$ natural proteins



$P(a_1, a_2, \cdots, a_L)$

CYDLVGWEPATAK

This protein does not exist in nature, but is functional!

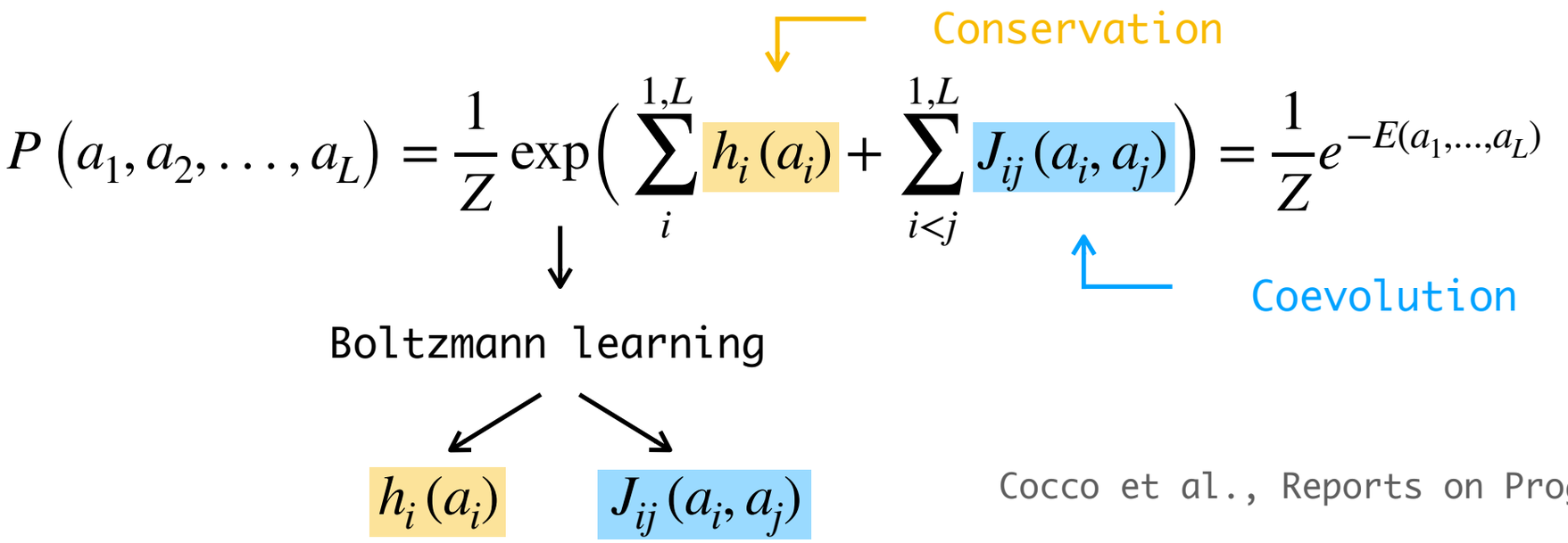
Russ et al. Science (2020)
Repack et al. Nature Machine Intelligence (2021)
Hawkins-Hooker et al. PLoS Comp. Bio. (2021)

Typical protein with $L = 100$

$20^L \sim 10^{130}$ possible sequences

$S[P] \sim 1.5 \sim \log(q)/2 \rightarrow 10^{65}$ functional sequences

Generative modeling



Cocco et al., Reports on Progress in Physics (2018)

Statistical physics approach: maximum likelihood, disordered Potts models

Many other different models (GAN, VAE, deep or not)

Coevolution is at the basis of all structure prediction methods (DCA, AlphaFold, etc.)

Some contributions from our group:

- Increase alignment accuracy Muntoni et al. PRE (2020)
- Simple and computationally efficient architectures Trinquier et al. Nat.Comm. (2021)
- Information-based procedure for parameter reduction Barrat-Charlaix et al. PRE (2021)
- Two publicly available packages: arDCA and bmDCA Muntoni et al. BMC Bioinformatics (2021)

Local sampling ~ evolution?

$$P\left(a_1, a_2, \ldots, a_L\right) = \frac{1}{Z}\exp\left(\sum_{i}^{1,L} h_i\left(a_i\right) + \sum_{i<j}^{1,L} J_{ij}\left(a_i, a_j\right)\right) = \frac{1}{Z}e^{-E(a_1,\ldots,a_L)}$$

Conservation

Coevolution

Boltzmann learning

$h_i\left(a_i\right)$ $J_{ij}\left(a_i, a_j\right)$

Wildtype

REDLVNY

Gibbs MCMC

draw i

$a_i{}^* \sim P\left(a_i \,|\, a_{j\neq i}\right) \propto e^{-\beta E(a_i|a_{j\neq i})}$

repeat

Library

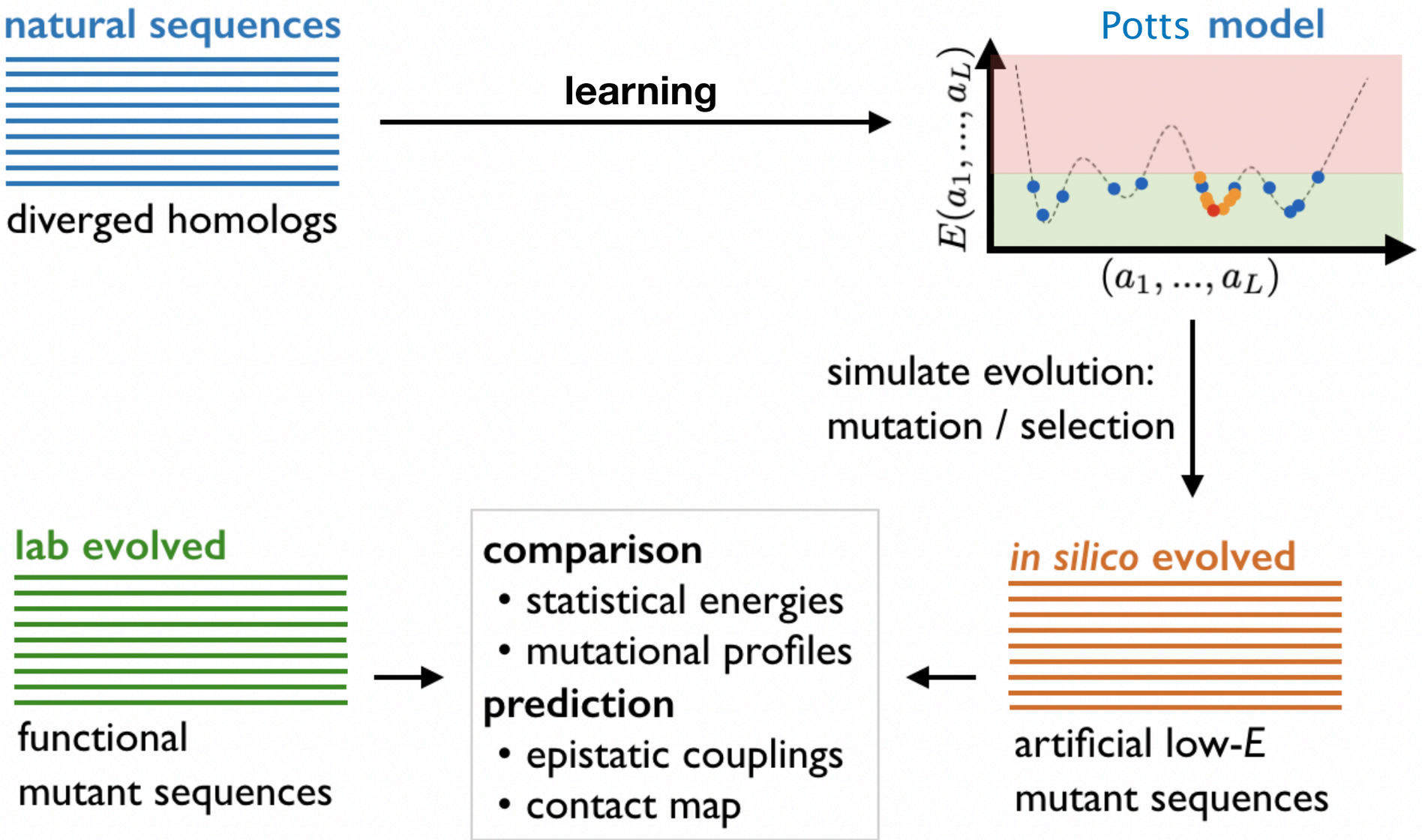
REDLANY
AEDLWNY
QNDLVWY

Independent chains

* a_i is drawn from the set of amino acids at distance 1 in terms of nucleotides (DNA sequence)

De la Paz et al., PNAS (2020)

Local sampling ~ evolution?

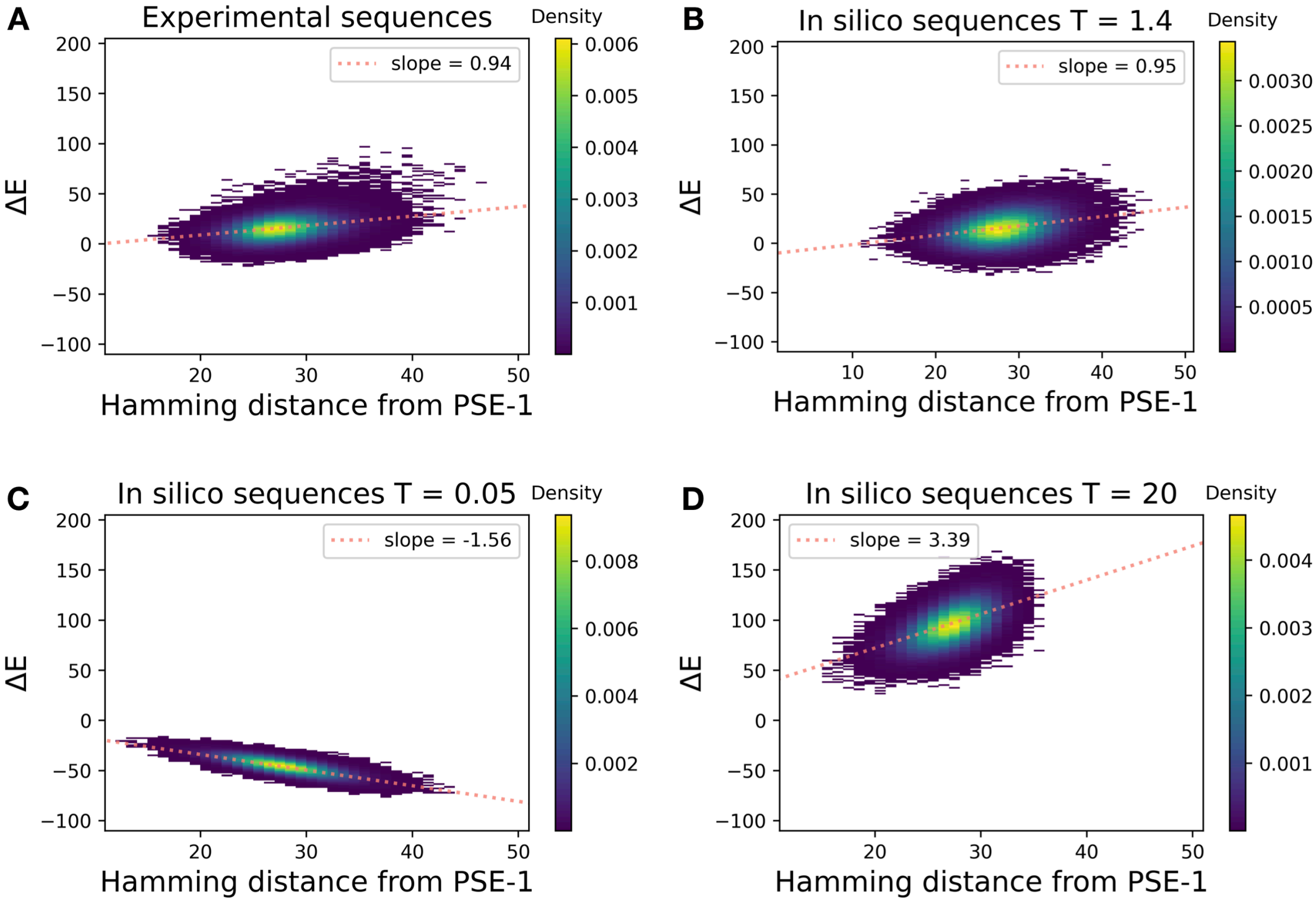


Global learning ≠ Local sampling

Natural evolution ≠ In-vitro evolution

Phylogenetic effects ≠ Independent chains (star phylogeny)

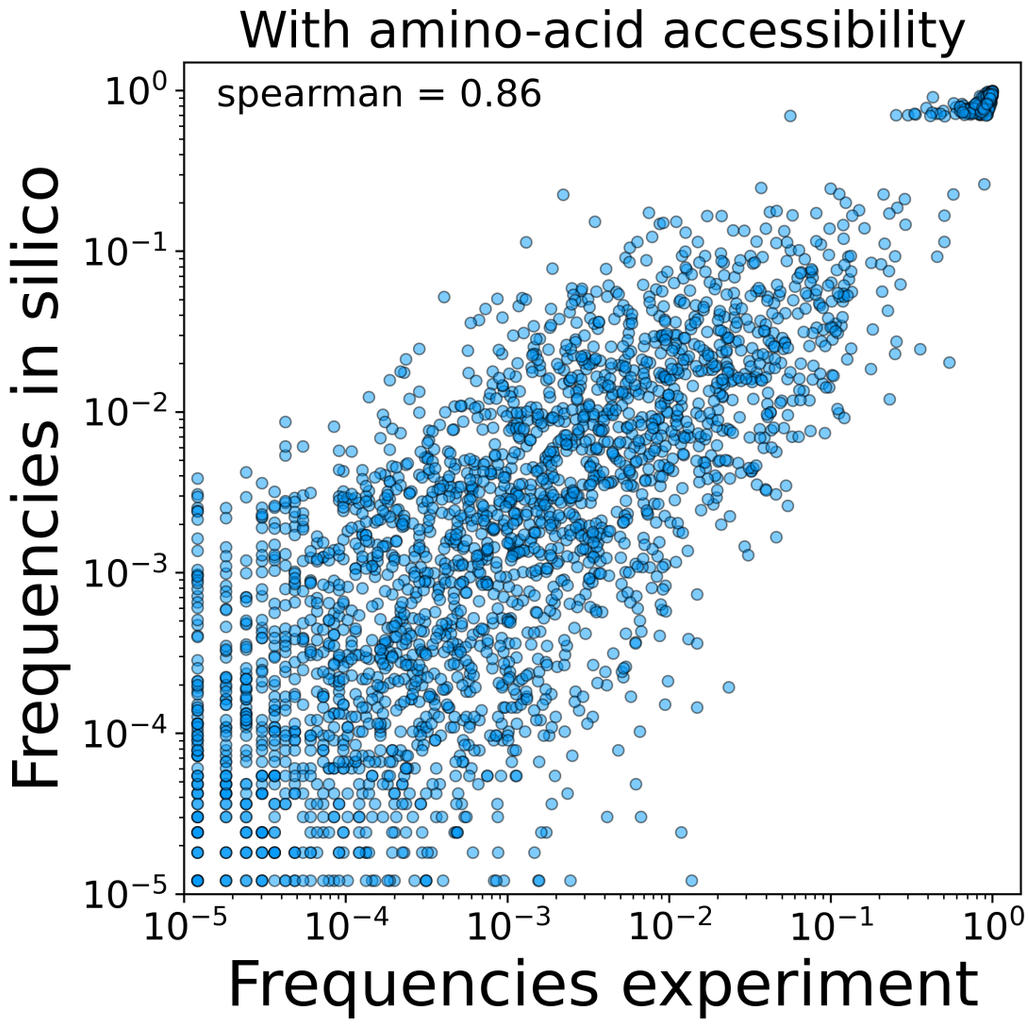
Local sampling ~ evolution?



Artificial temperature T can model selection strength

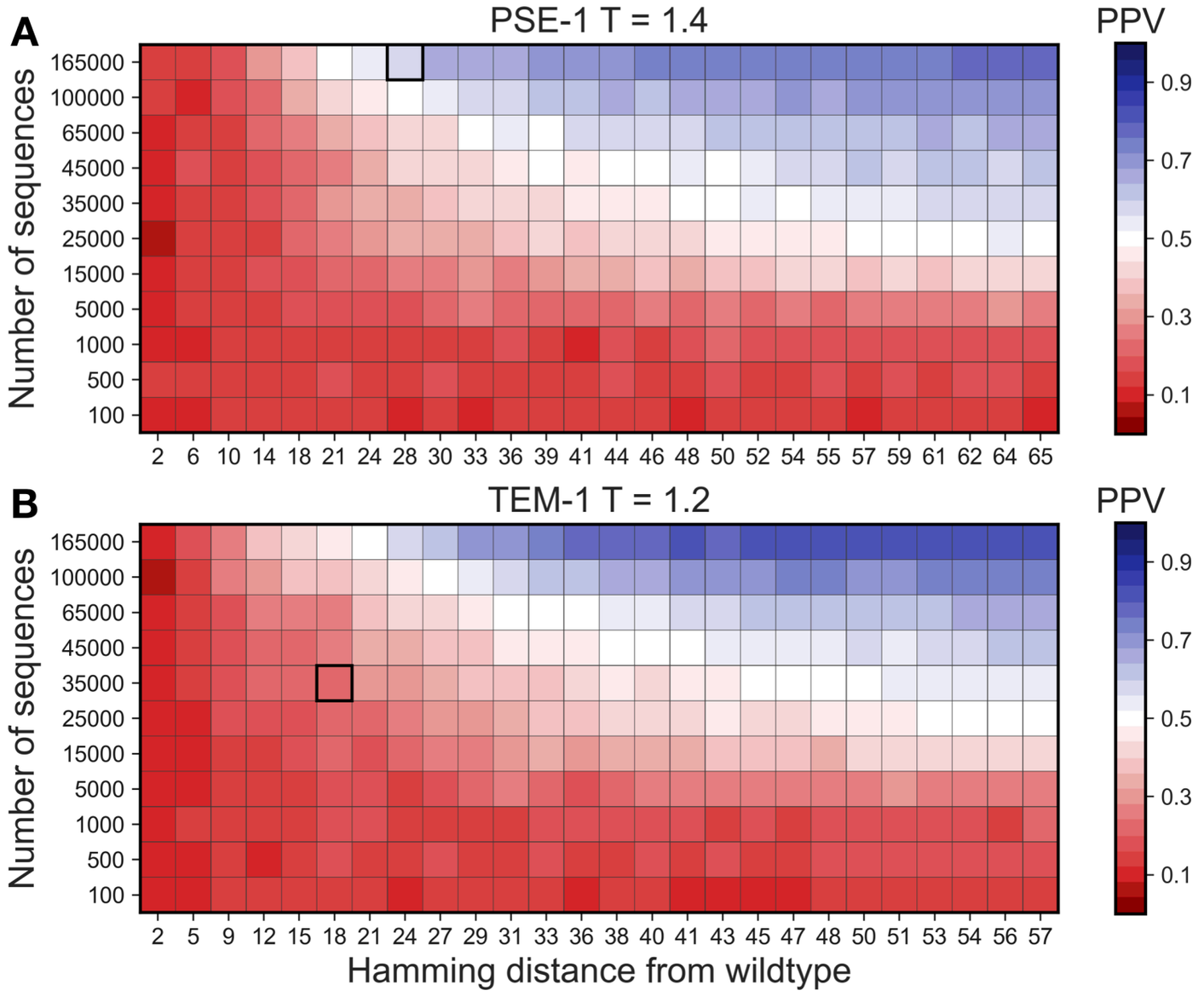
Local sampling ~ evolution?

The model can quantitatively reproduce statistical features of experiments:
e.g. high correlation (86%) of amino acid frequencies



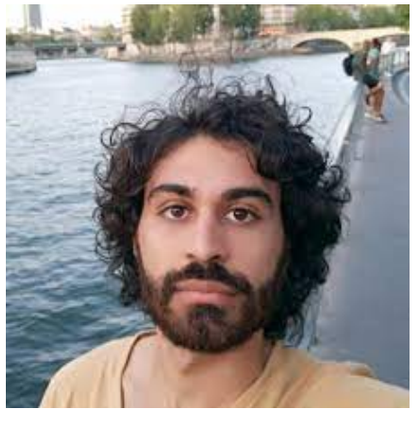
Experimental library	In silico library
REDLKNY	REDLVNY
EEDLVNY	AEDLKWY
QNDLVWY	RNDLVWY
$f_i^{exp}(a_i)$	$f_i^{sil}(a_i)$

In silico evolution: design new experiments



Stiffler et al., Cell Systems (2020)

Fantini et al., Mol.Bio.Evo. (2020)



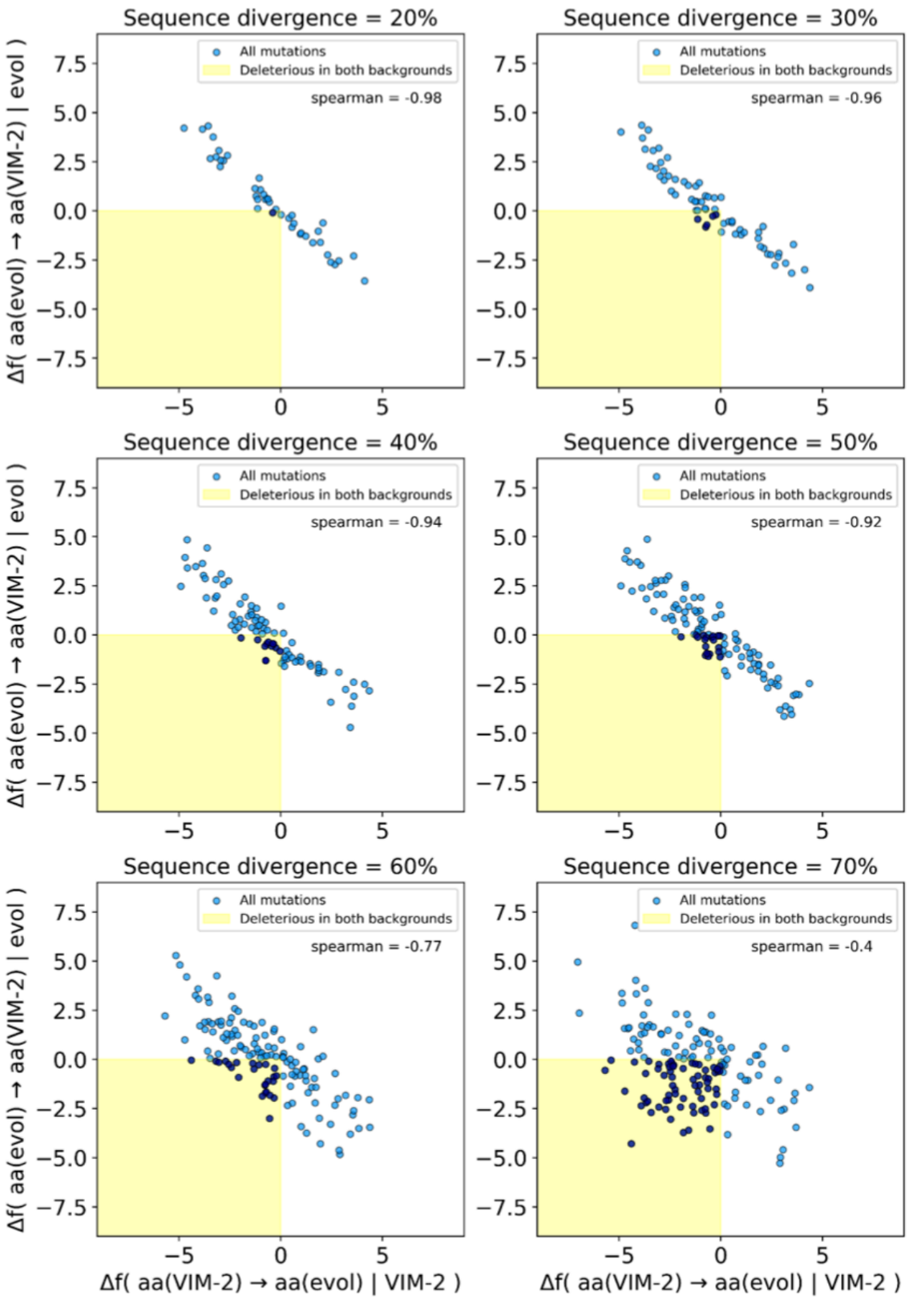
Artificial temperature T can model selection strength

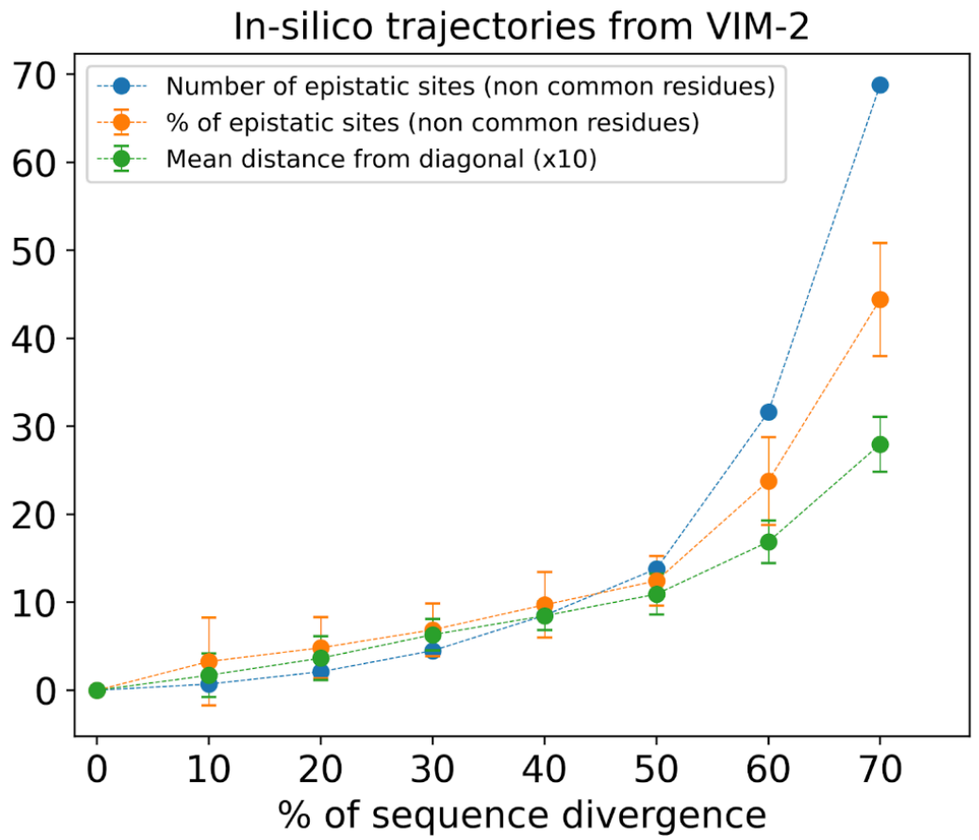
Emergence of coevolution as a function of distance and number of sequences

Optimize and design new experiments

Bisardi, Rodriguez-Rivas, FZ, Weigt, Mol.Bio.Evo. (2022)

In silico evolution: emergence of coevolution



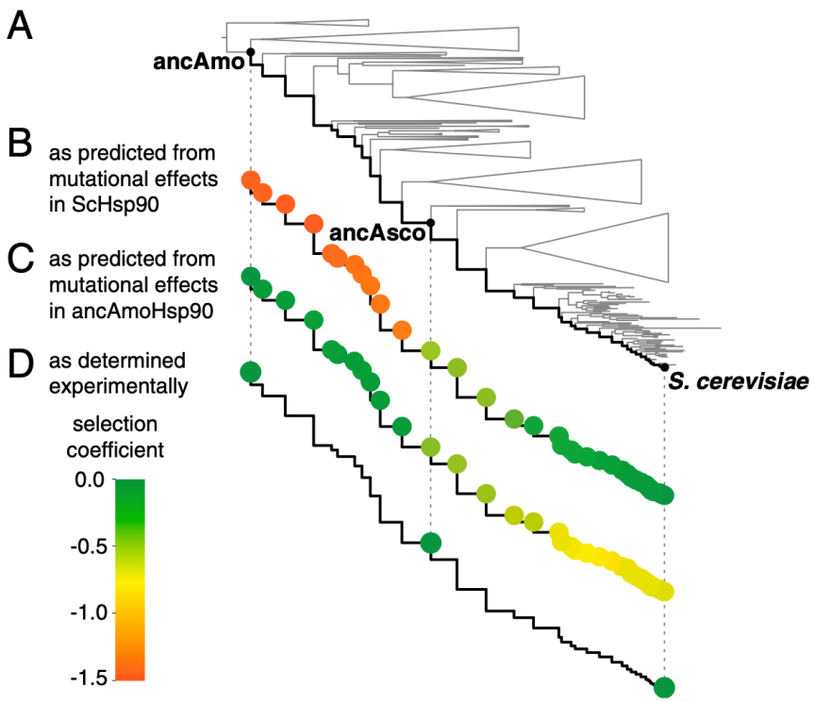


Emergence of coevolution (aka epistasis)

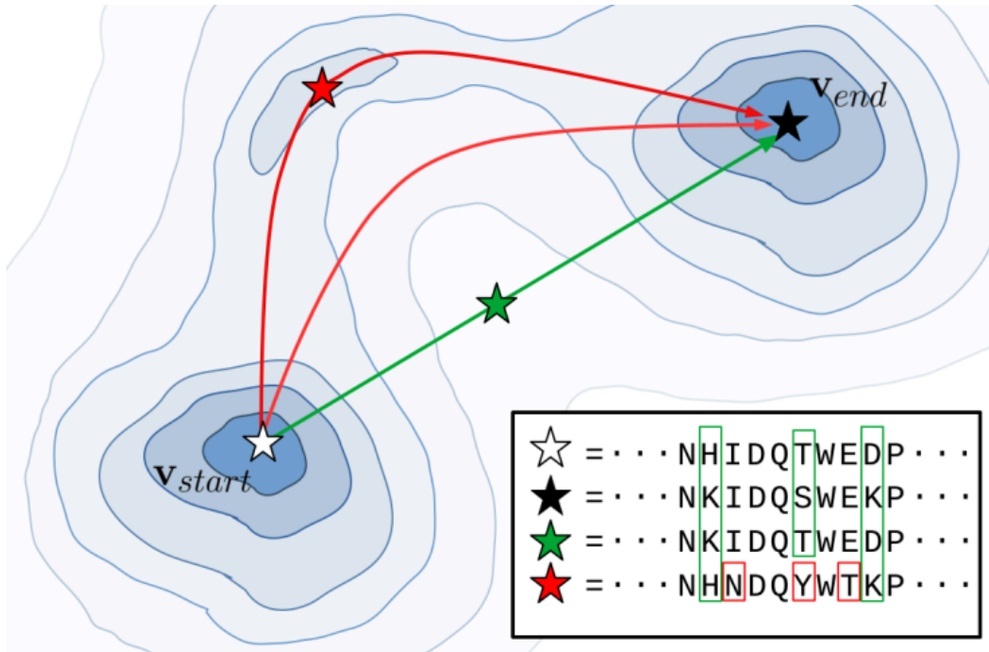
along in silico evolutionary trajectories

Bisardi, Cotogno, Weigt, FZ
+ Tokuriki lab (in preparation)

Path sampling



Starr et al. PNAS (2017)



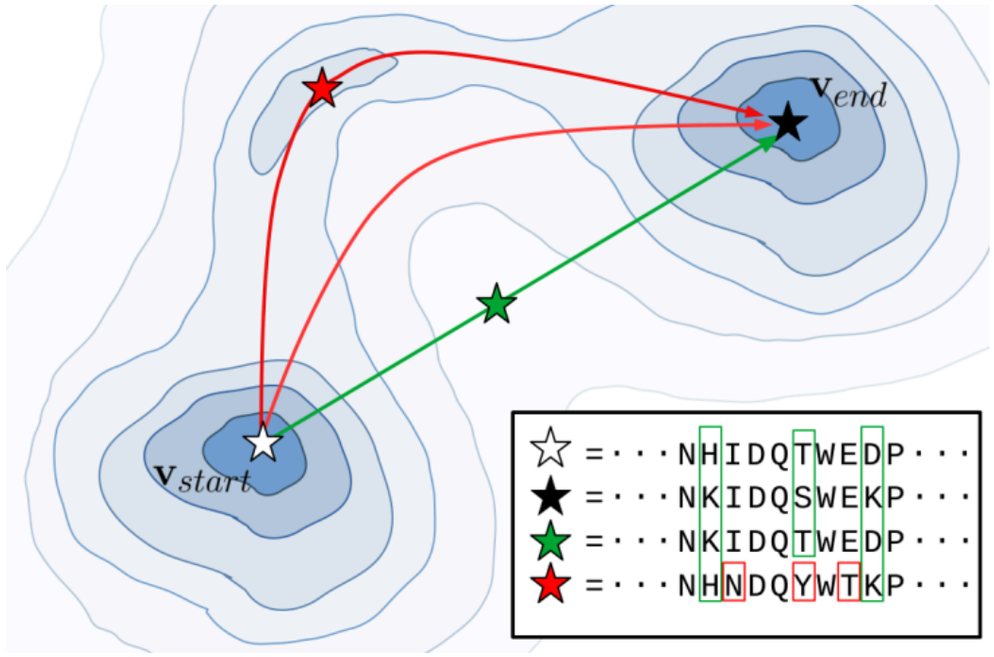
Mauri, Cocco, Monasson, preprint (2022)

- Navigability of the sequence landscape
- Evolutionary paths connecting two wild types
- Intermediate sequences between two wild types with desired properties
- Direct versus wandering paths
- Well defined computational framework: transition path sampling

Dellago et al. JCP (1998), Mora-Walczak-FZ PRE (2012)

$$P(\mathbf{a}_1, \cdots, \mathbf{a}_T) = P(\mathbf{a}_1)P(\mathbf{a}_1 \to \mathbf{a}_2)P(\mathbf{a}_2 \to \mathbf{a}_3)\cdots P(\mathbf{a}_{T-1} \to \mathbf{a}_T)$$

Path sampling



Metallo β-lactamases enzymes confer broad antibiotic resistance in bacteria

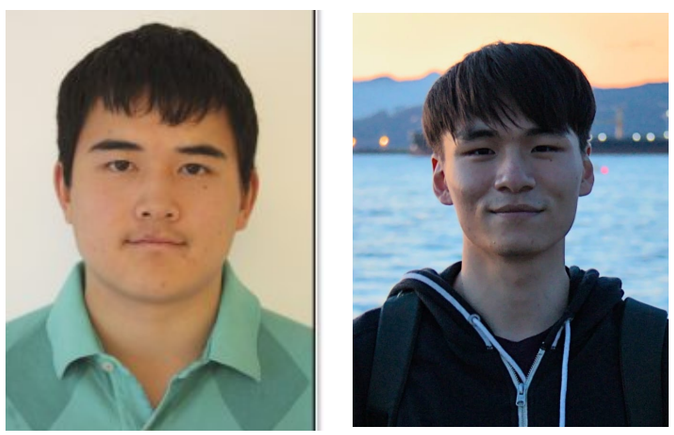
VIM-2 and NDM-1: well studied “superbugs” in this class
~64% sequence divergence

Tokuriki’s lab characterized all single-mutants of both enzymes (Deep Mutational Scanning, DMS)

We constructed a model based on an alignment of natural sequences and the integration of the experimental DMS

Generate intermediate artificial sequences along a mutational path VIM-2 → NDM-1
Express them in bacteria and measure their fitness

Path sampling



Bisardi, Cotogno, Chen, Lee
Work in progress

Preliminary result: a functional path

Sequence	Position model	Position VIM	Old AA	New AA	EC50 µg/mL AMP	log2(ΔEC50)	ΔFitness DMS VIM	-ΔE integrated model	-ΔE non-integ. model	Residue distance
VIM2-LE					406.7					
MeanE1 1	144	181	T	S	37.86	-3.43	0.05	-1.39	0.59	
MeanE1 2	30	67	Y	V	101.18	1.42	1.27	1.14	-3.23	19.4
MeanE1 3	27	ins	I	F	36.01	-1.49	missing	0.49	1.22	
MeanE1 4	205	242	L	A	22.4	-0.68	0.83	1.13	0.79	
MeanE1 5	20	58	A	S	123.52	2.46	1.14	2.57	3.73	10.4
MeanE1 6	220	257	K	D	1.56	-6.31	-1.08	-1.82	-1.01	25.2
MeanE1 7	108	145	V	Q	53.61	5.1	missing	-0.06	1.55	26.3
MeanE1 8	19	57	I	T	22.79	-1.23	1.25	1.4	2.25	26
MeanE1 9	65	102	Q	E	43.59	0.94	0.55	1.15	1.11	9.1
MeanE1 10	101	138	S	L	157.57	1.85	0.24	-0.29	3	31.4
MeanE1 11	17	55	S	Q	163.89	0.06	0.95	0.59	-0.3	23.8
MeanE1 12	192	229	H	A	315.52	0.95	0.57	1.03	1.61	20.3
MeanE1 13	53	90	K	D	125.61	-1.33	0.87	0.62	1.22	33
MeanE1 14	209	246	L	R	113.19	-0.15	0.62	1.27	1.37	30.9
MeanE1 15	143	180	S	T	190.9	0.75	1.09	1.17	1.05	17.9
MeanE1 16	141	178	A	G	153.31	-0.32	-0.46	1.04	2.58	3
MeanE1 17	135	172	L	V	16.04	-3.26	-0.14	0.46	4.87	13.8
MeanE1 18	162	199	A	L	299.84	4.22	-0.6	-0.86	-2.89	12.3
MeanE1 19	148	185	I	T	123.13	-1.28	-0.32	-1.8	0.19	5
MeanE1 20	164	201	Y	K	11.46	-3.43	-3.05	-0.96	6.57	9.7
MeanE1 21	158	195	Y	F	77.75	2.76	0.38	-0.47	1.74	9.4
MeanE1 22	23	61	S	D	1.74	-5.48	-0.54	-1.67	1.24	16.6
MeanE1 23	61	98	E	W	1.74	0	missing	-0.3	0.96	18.2
MeanE1 24	153	190	S	G	2.24	0.36	0.15	0.03	0.54	20.1
MeanE1 25	213	250	K	T	1.86	-0.27	0.19	0.39	1.9	21.9
MeanE1 26	124	161	S	G	1.83	-0.02	0.38	0.2	2.69	17.7
MeanE1 27	169	206	T	K	2.04	0.16	0.44	0.56	-0.53	24.9
MeanE1 28	168	205	R	A	17.17	3.07	0.68	1.41	-0.15	1.3

LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

Summary

cnrs

Generative modeling for protein families leads to disordered models with "rough" fitness landscapes
Basins of attraction, transition paths, barriers, topology

Study of the dynamics (evolution) in these landscapes:

- Validate the model against *in vitro* evolution
- Study the emergence of coevolution (epistasis)
- Construct evolutionary paths *in silico* and *in vitro*
- Ultimate goal: design artificial proteins with desired properties

Thank you for your attention!

SIMONS FOUNDATION

erc