SORBONNE UNIVERSITÉ

SPQB

STATISTICAL PHYSICS FOR QUANTITATIVE BIOLOGY

GitHub

SAPIENZA UNIVERSITÀ DI ROMA

Emergent time scales of epistasis in protein evolution

Francesco Zamponi

Mathematical Physics Webinar, Rutgers University,
February 19, 2025

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

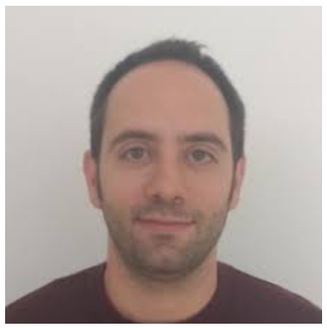
Di Bari, Bisardi, Cotogno, Weigt, FZ, PNAS (2024)

Rossi, Di Bari, Weigt, FZ, arXiv:2412.01969 (2024)

Martin Weigt

Matteo Bisardi

Sabrina Cotogno



Juan Rodriguez -Rivas

Leonardo Di Bari

Saverio Rossi

Goal

Sequence-to-structure-to-function paradigm

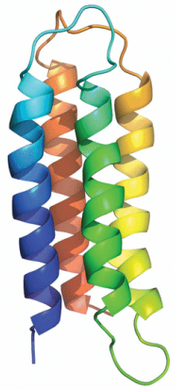
Sequence to function
Lack of function data, partially unsupervised
Less defined: what is function?
Context dependent

Sequence
~10^8 sequences available
genome sequencing

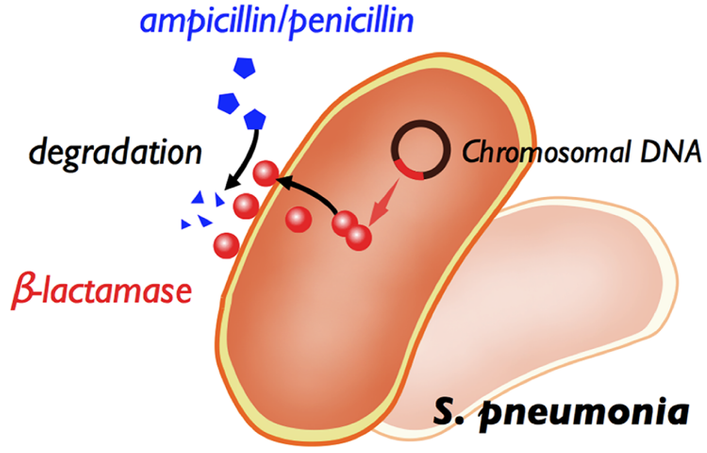
Structure
~10^5 structures available
X-ray crystallography, NMR or electron microscopy

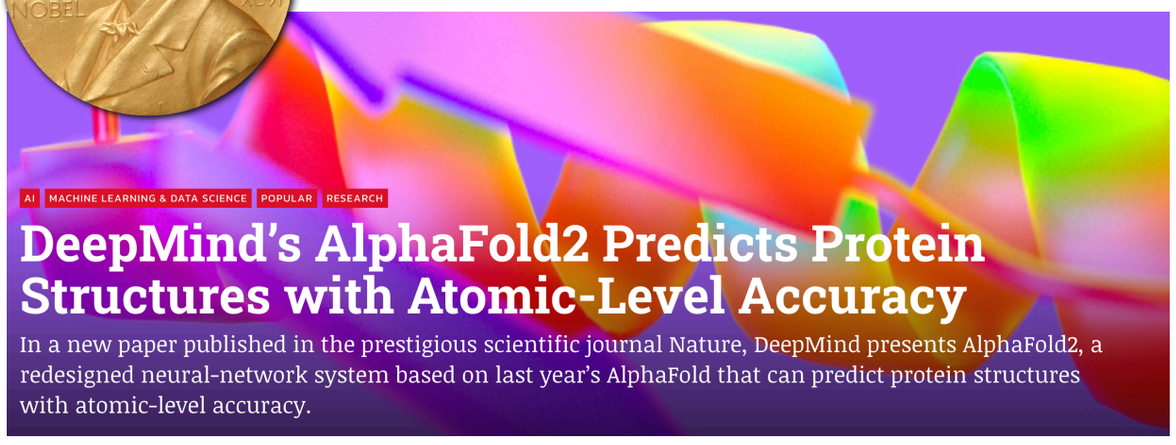
Function
Severe lack of reliable functional annotations

KPDLKPYSPLRDK



Sequence to structure
Supervised problem
Mathematically well defined
Deep neural networks





Long term goals

- Learn something about the sequence-to-function relationship
- Design artificial protein sequences with desired functions
- Understand how nature does it by evolution
- Understand and optimize protocols to do it in the lab (directed in vitro evolution)

Take home: we use a data-driven approach and statistical physics modelling

SPQB

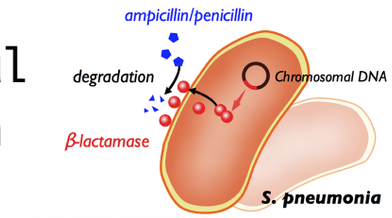
STATISTICAL PHYSICS FOR QUANTITATIVE BIOLOGY

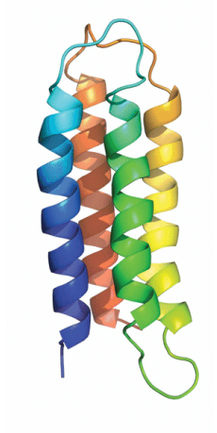
GitHub

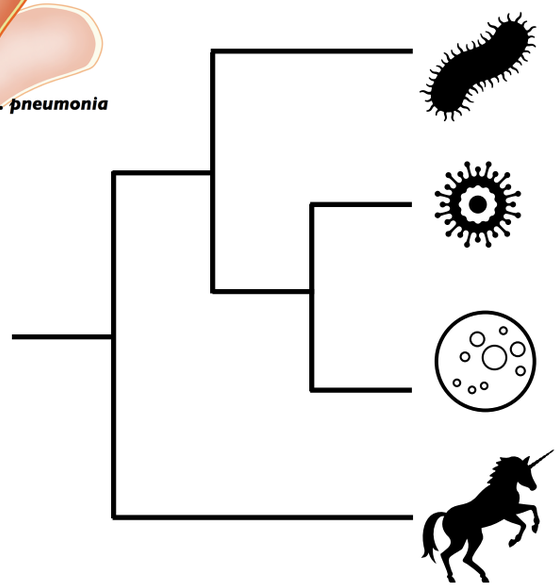

Data

Data #1: natural protein sequences

Ancestral protein







Natural evolution

Conservation

KPDLKPYSPLRDK

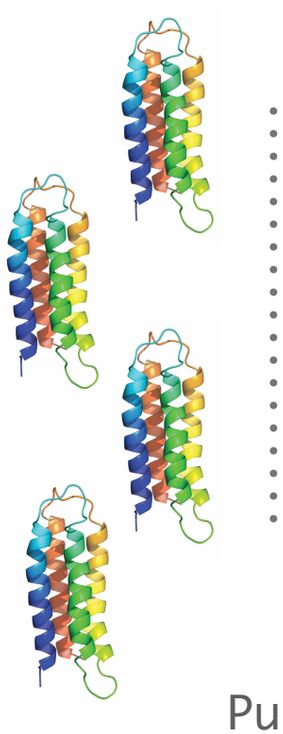
-EDLVNYNPITEK

QNDLKE-SPVDEK

AADLVANSPVTGK

Coevolution

Same structure and function!

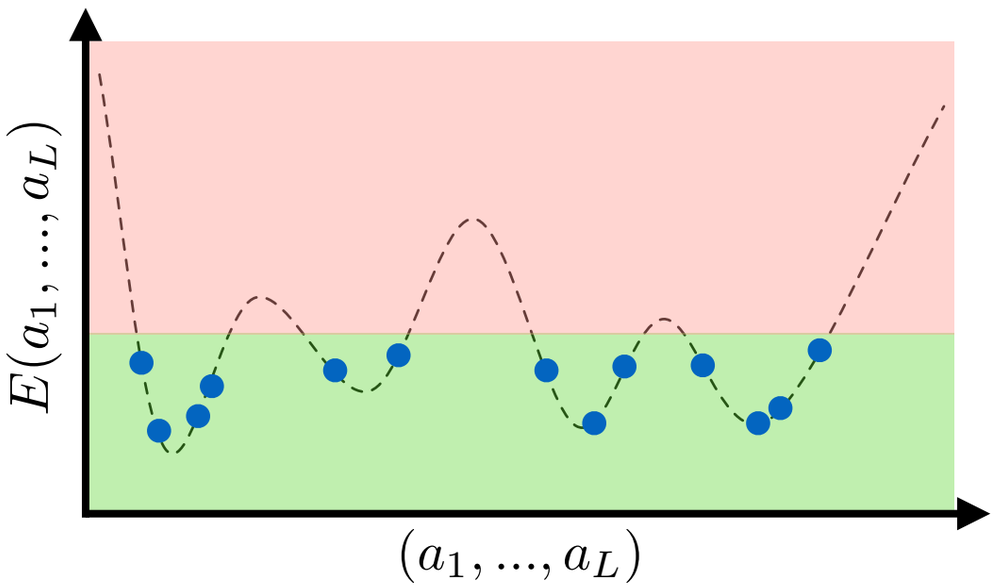


Available sequences: 10^2-10^5

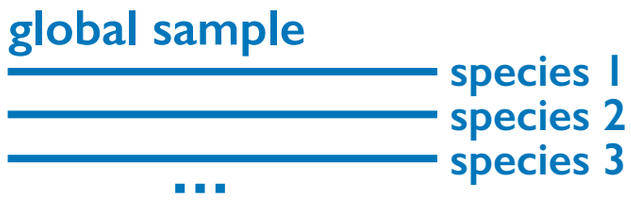
Sequence divergence: 70-80%

Public databases (PFAM, UniProt..)

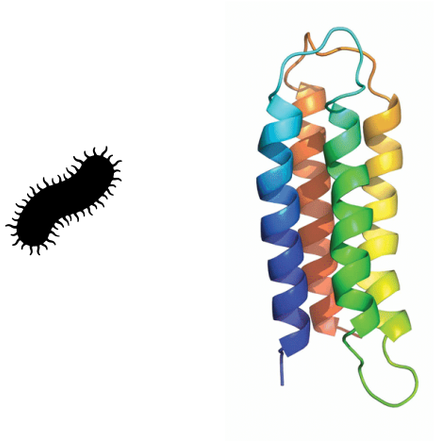
sequence landscape



sequence data



Data #2: deep mutational scans



APDLKPYSPLRDK

KADLKPYSPLRDK

KPALKPYSPLRDK

...

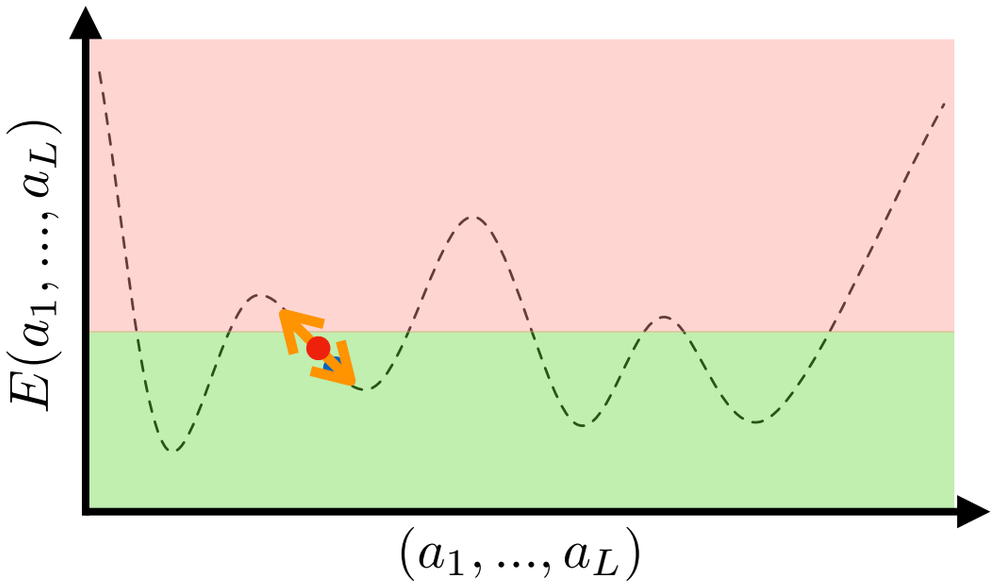
KPDLKPYSPLRDA

Available sequences: $\sim 10^4$

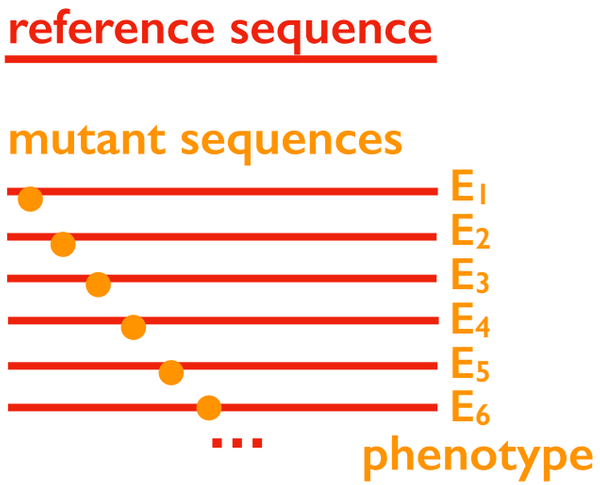
Sequence divergence: 1

Fowler and Fields, Nature Methods (2014)

sequence landscape

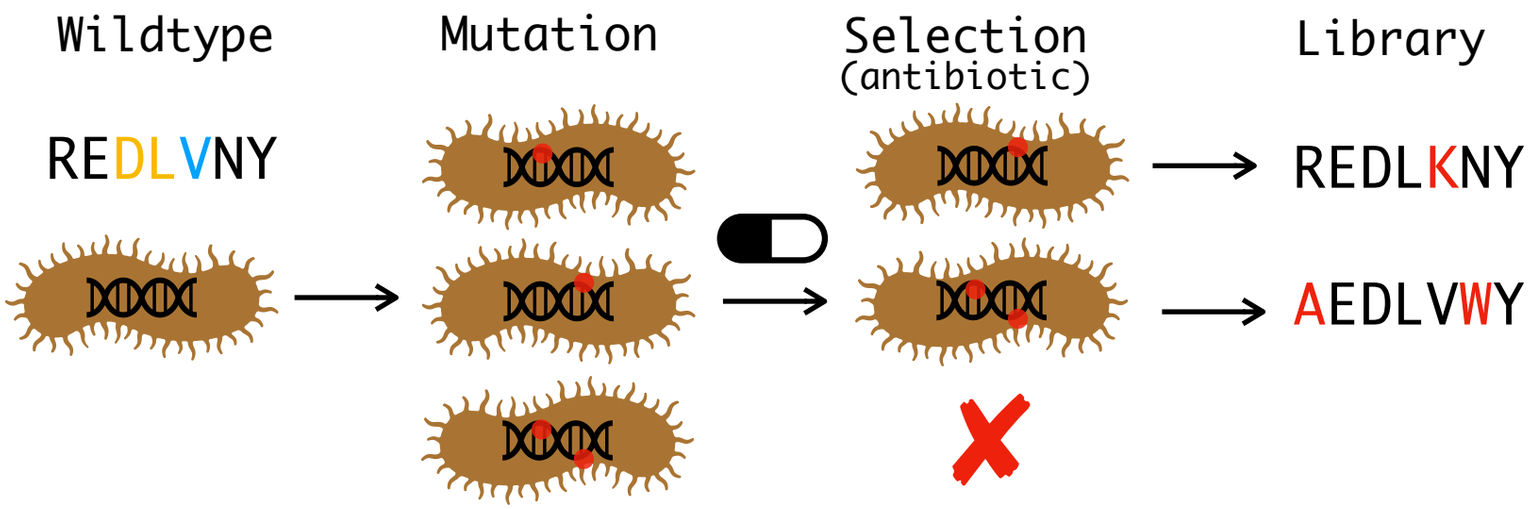


sequence data



Data #3: in vitro evolution

(weak selection)



Available sequences: 10^5-10^6

Sequence divergence: 10-20%

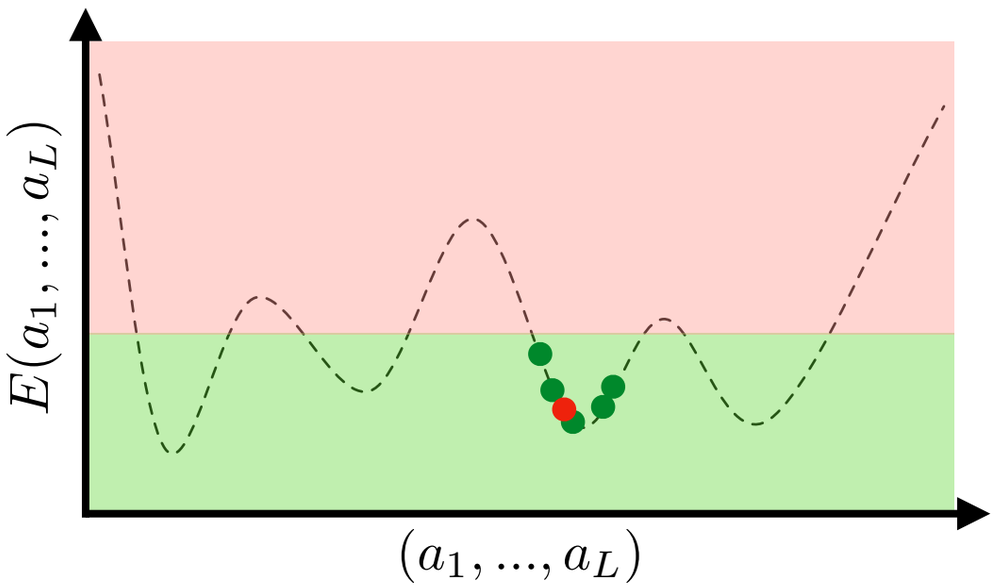
Stiffler et al., Cell Systems (2020)
Fantini et al., Mol.Biology and Evo. (2020)
Erdogan et al., bioRxiv (2023)

Natural evolution

Available sequences: 10^2-10^5

Sequence divergence: 70-80%

sequence landscape



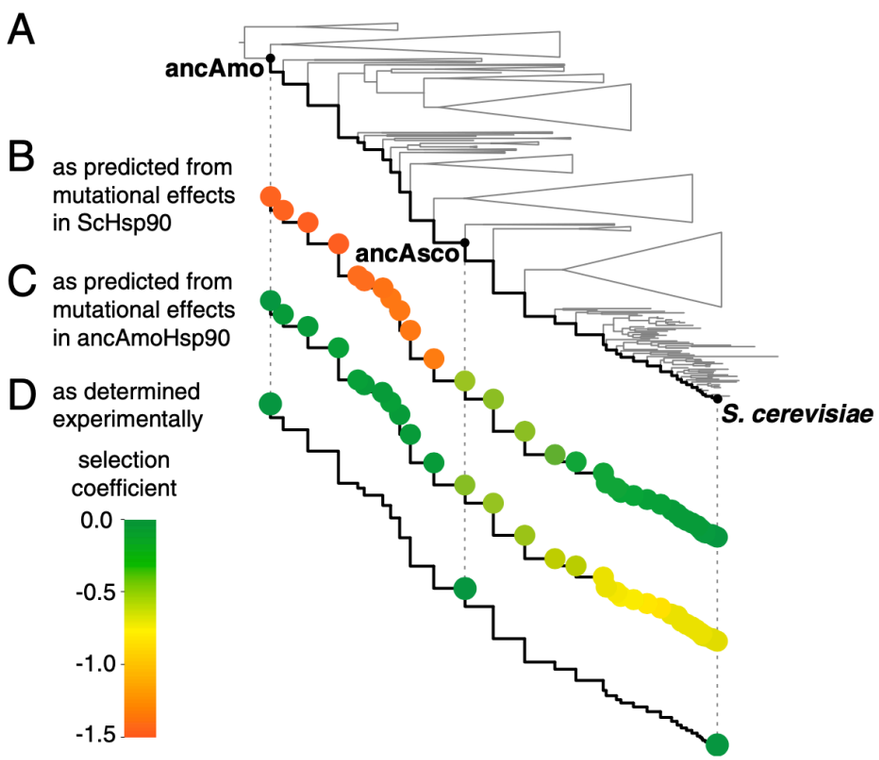
sequence data

reference sequence

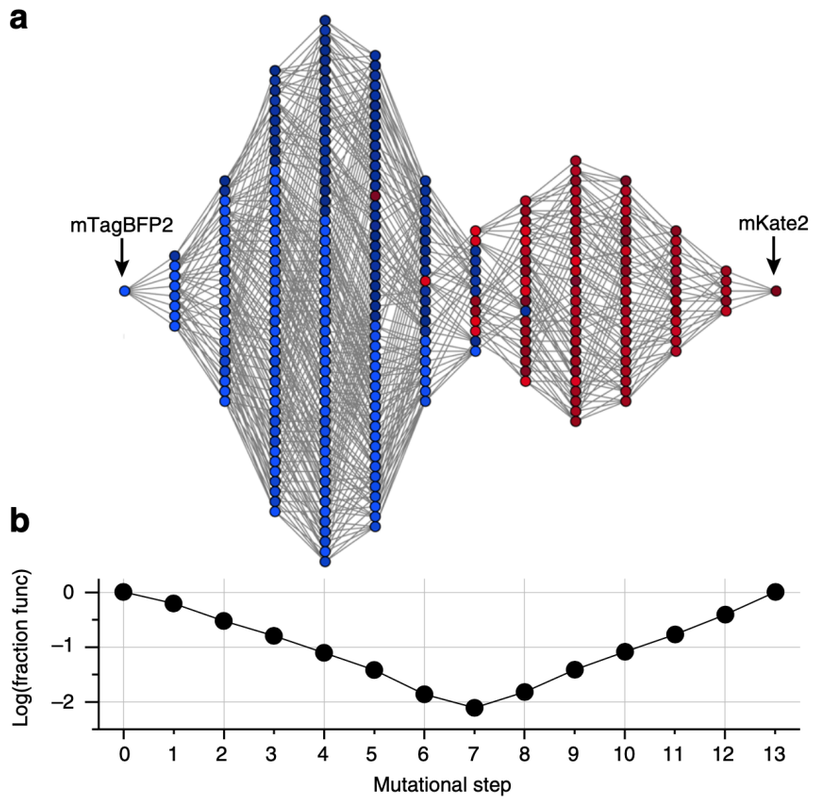
local sample

mutant 1
mutant 2
mutant 3
...

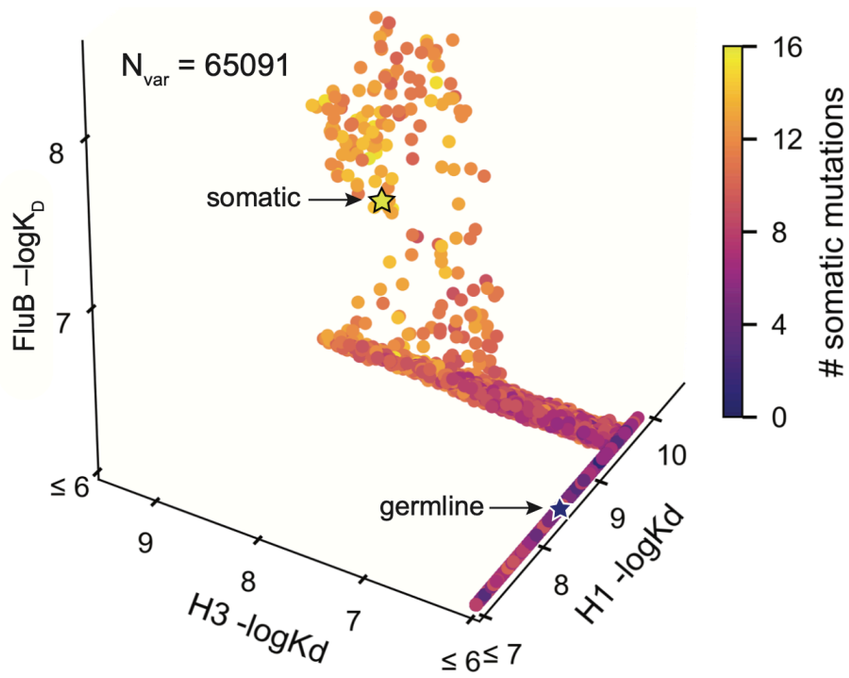
Data #4: path/phylogeny reconstruction



Starr et al. PNAS (2017)

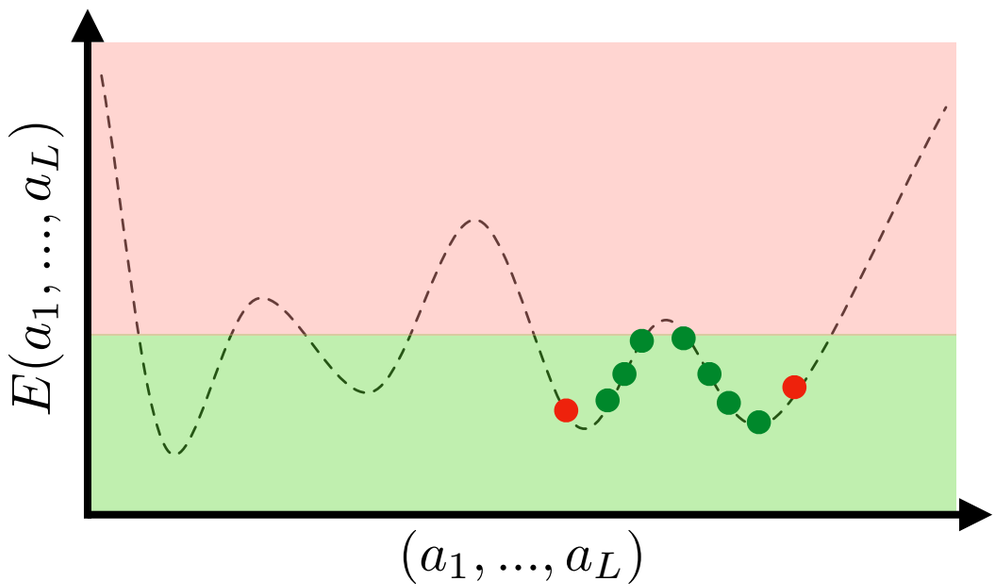


Poelwijk et al. Nat.Comm. (2019)



Phillips et al. eLife (2021)

sequence landscape



sequence data

reference sequence

local sample

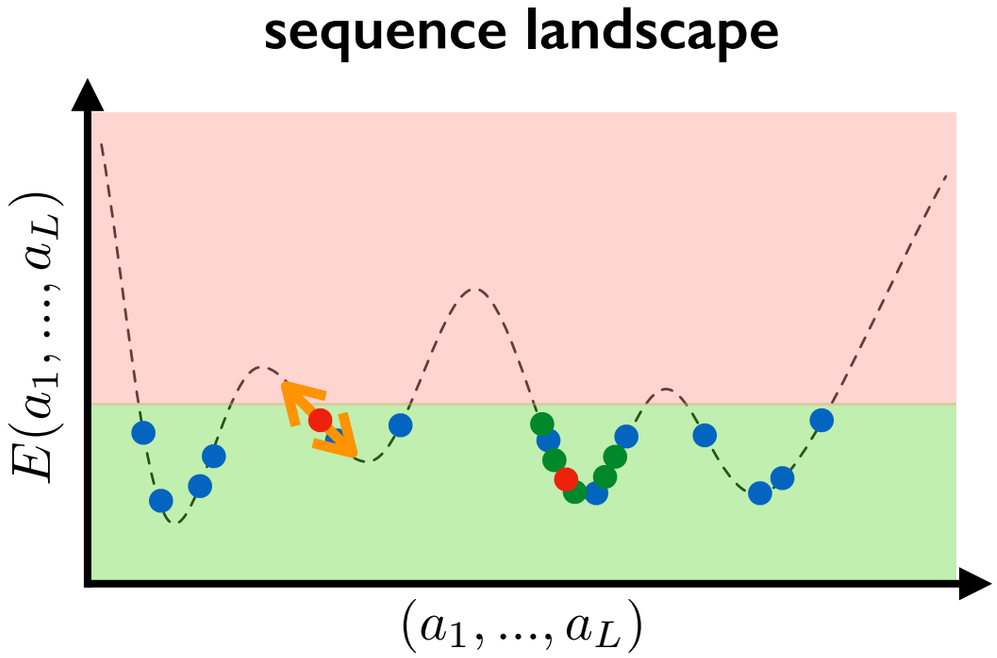
path 1

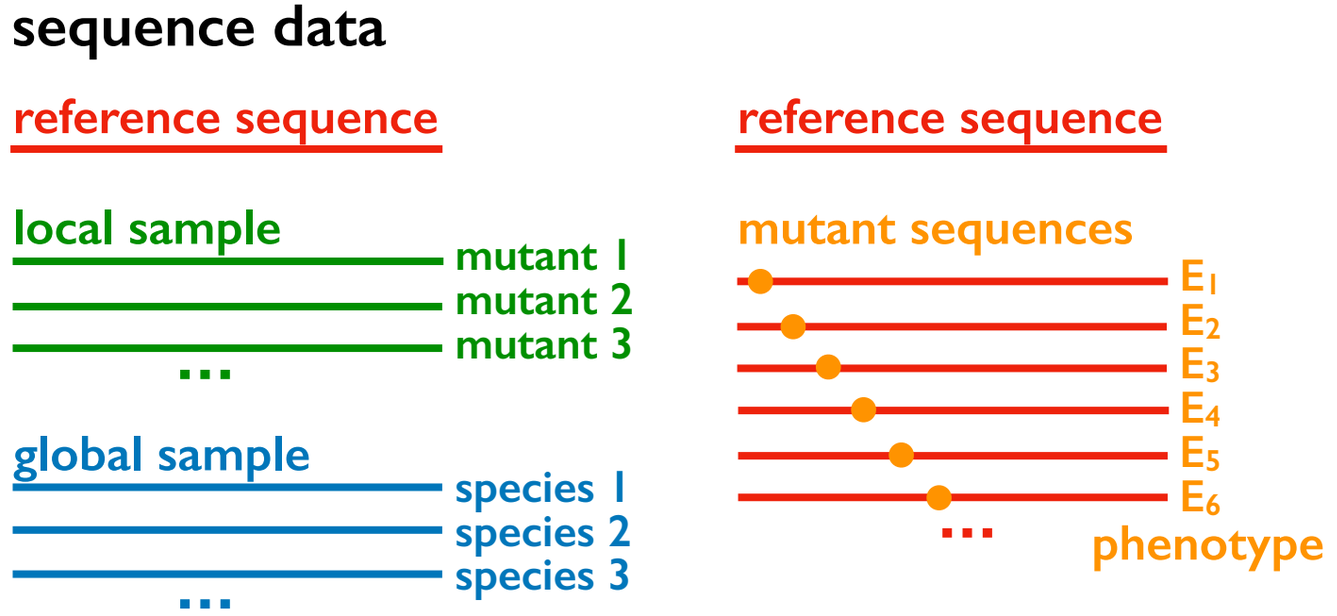
path 2

path 3

...

Massive amount of data!





Characterize the dynamics in the sequence landscape:

- Model in vitro evolution (controlled environment)
- Model natural evolution (phylogeny, fluctuating environment)
- Optimize evolution protocols (antibiotic concentration, vaccination protocols…)
- Generate paths of artificial sequences connecting two natural ones

Take home: there are many data, and many more to come

Generative modelling

Generative modelling (unsupervised)

Many images of human Faces

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

this-person-does-not-exist.com

A lot of text

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

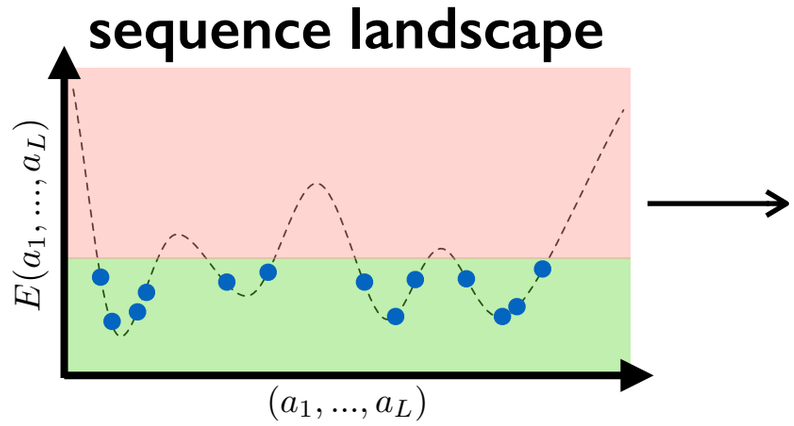
Fish grew some legs,
Said, "I'm done with the sea!"
Now we pay rent.

chatgpt.com

Generative modelling of biological sequences

KPDLKPYSPLRDK
-EDLVNYNPITEK
QNDLKE-SPVDEK
AADLVANSPVTGK

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



CYDLVGWEPATAK

This protein does not exist in nature

Conservation

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

Coevolution

→ Boltzmann learning → $h_i(a_i)$, $J_{ij}(a_i, a_j)$

Statistical physics approach: maximum likelihood, disordered Potts models

Coevolution is at the basis of all structure prediction methods (including AlphaFold)

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

Simpler models
Because we have less data, but…
More interpretable
Energy efficient
Accessible to everyone

Trump Announces $100 Billion A.I. Initiative

OpenAI, Oracle and SoftBank formed a new joint venture called Stargate to invest in data centers, building on major U.S. investments in the technology.

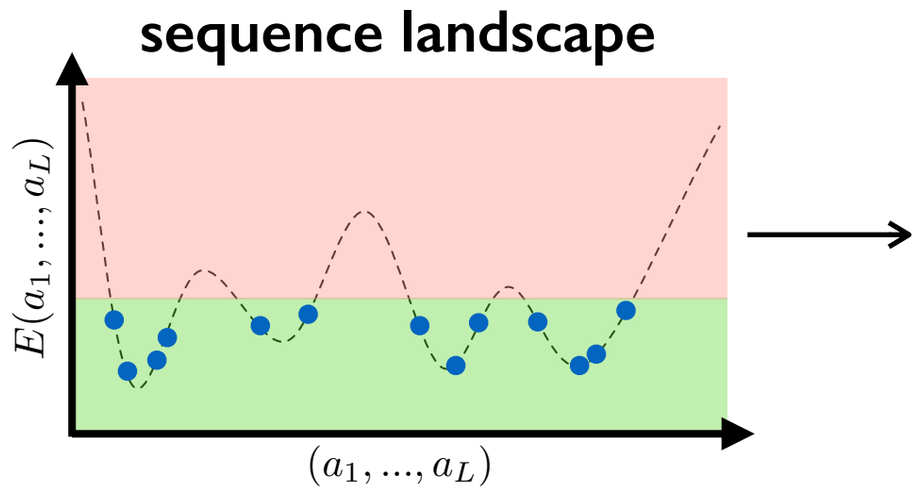
By Cecilia Kang and Cade Metz
Reporting from Washington and S

The New York Times

Jan. 21, 2025

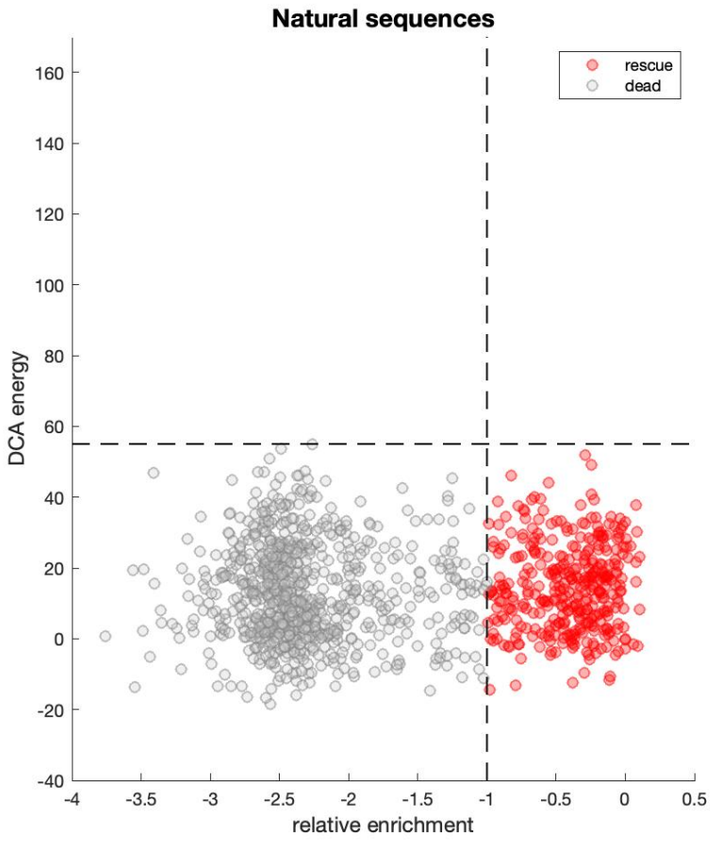
Generative modelling of biological sequences

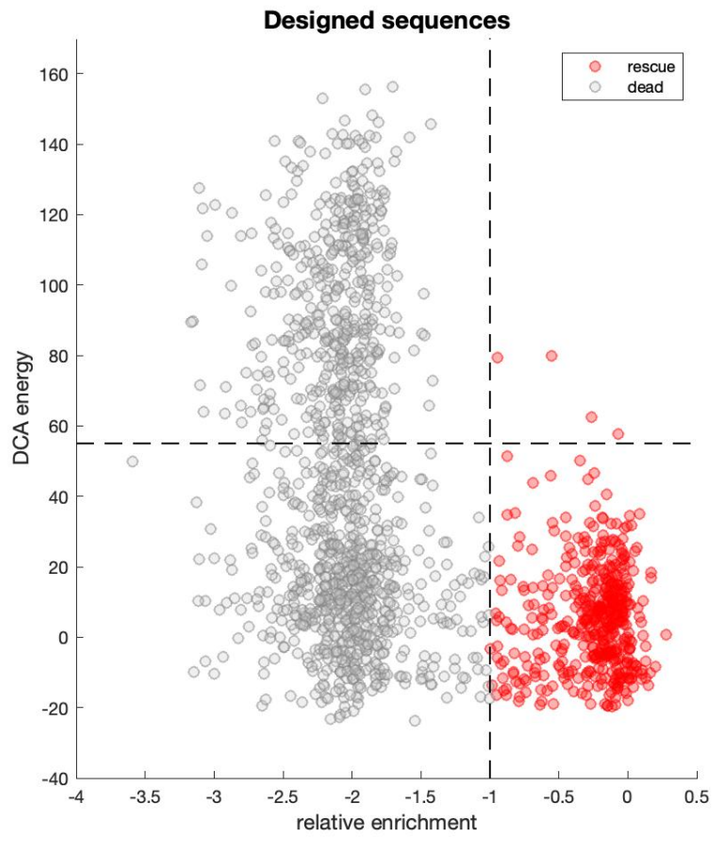
KPDLKPYSPLRDK
-EDLVNYNPITEK
QNDLKE-SPVDEK
AADLVANSPVTGK

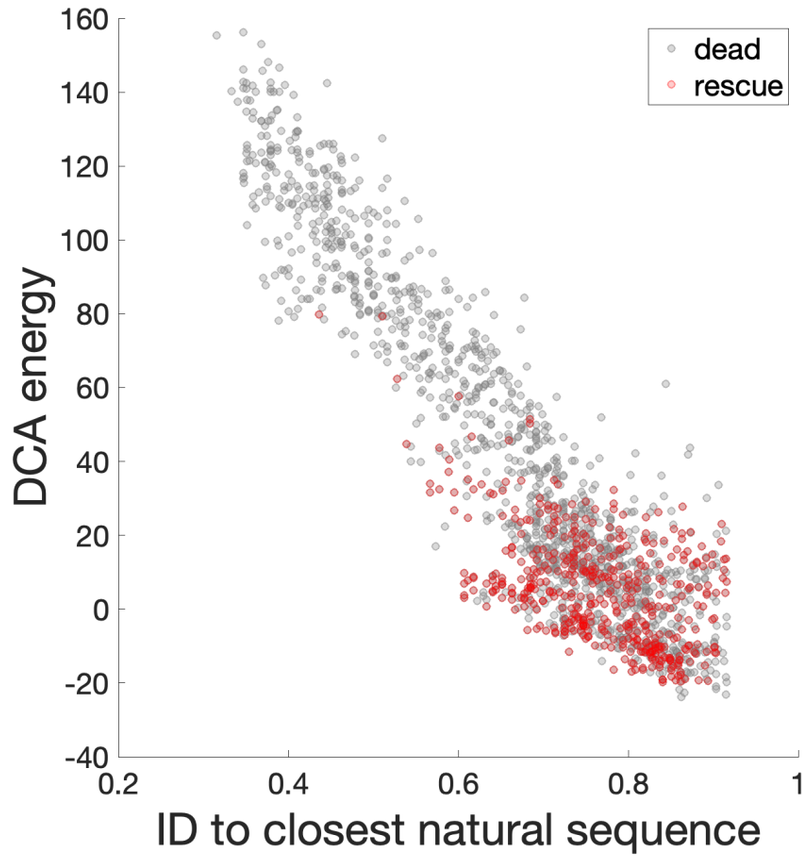


CYDLVGWEPATAK

$$P(a_1, a_2, \cdots, a_L) \sim \exp[-E(a_1, \cdots, a_L)]$$







Typical protein with $L = 100$ (Chorismate Mutase expressed in E.Coli)

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

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

Russ et al. Science (2020)

Generative modelling of biological sequences

Generative models have been experimentally validated!

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

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

Two public packages that everyone can use: adabmDCA and arDCA

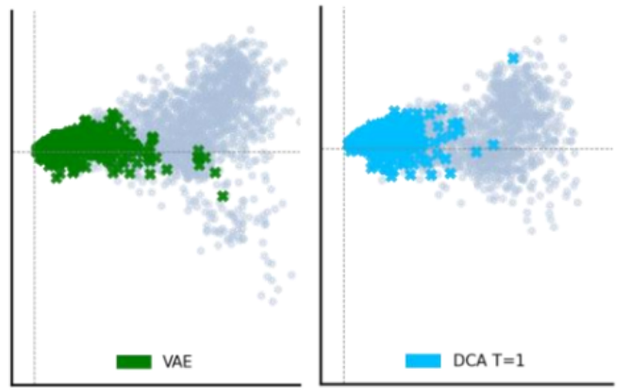
SPQB

STATISTICAL PHYSICS FOR QUANTITATIVE BIOLOGY

GitHub

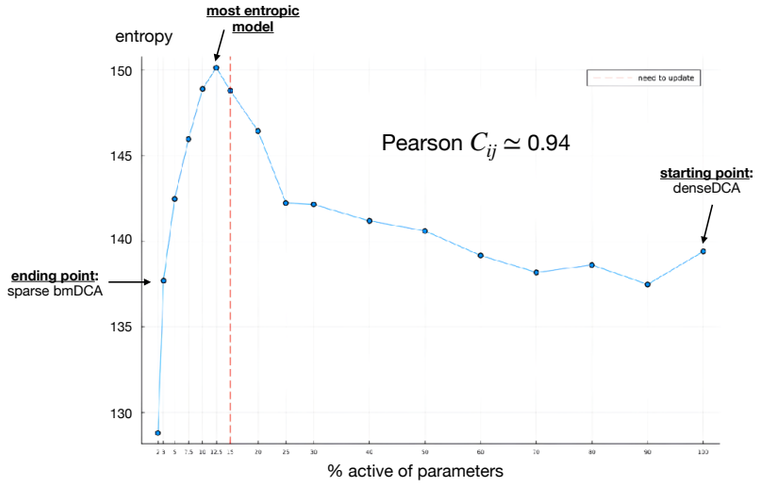
Trinquier et al. Nat. Comm. (2021)
Muntoni et al. BMC Bioinformatics (2021)
Rosset et al. Springer Protocol (to appear)

The model is simple and interpretable, and not generically outperformed by other methods



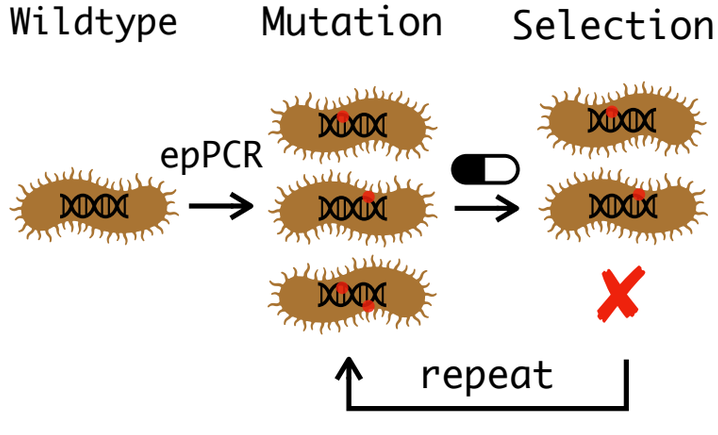
Calvanese et al. Nucleid Acids Research (2024)
Lambert, Opuu, Calvanese et al. bioRxiv (2024)

Information-based parameter reduction and entropy maximization

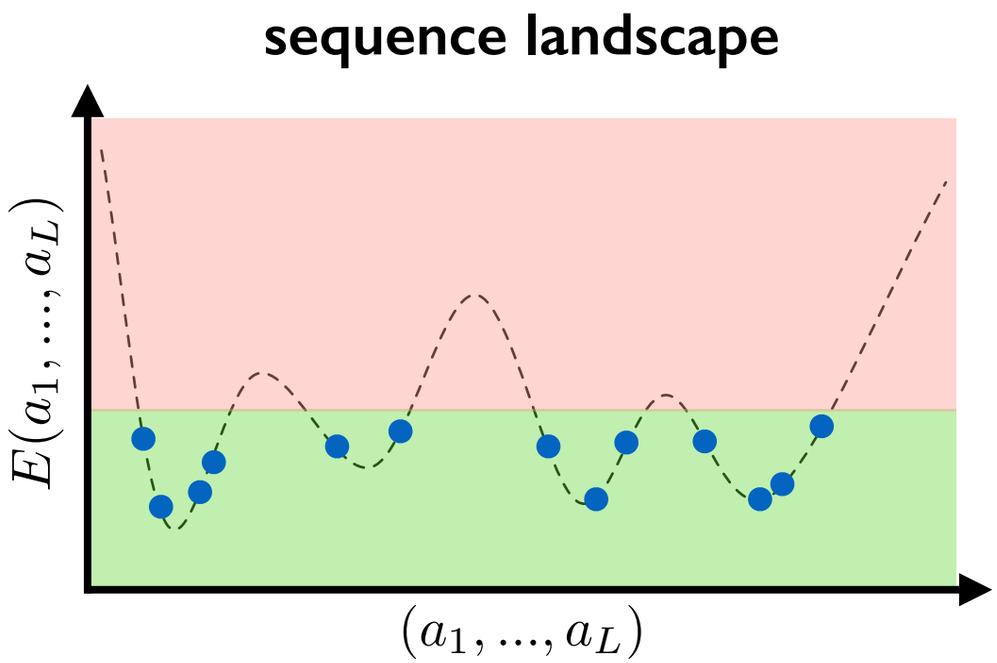


Barrat-Charlaix et al. PRE (2021)
Ongoing experimental test with Ranganathan's group

Model and direct evolution (today's main topic)



What comes next?



Equilibrium sampling

$$P(a_1, a_2, \cdots, a_L) \sim \exp[-E(a_1, \cdots, a_L)]$$

Generate artificial sequences ✓

Tune their properties ✓

Take home: simple generative models can design artificial proteins (equilibrium sampling)

Next: characterize the shape of the sequence landscape and the dynamics

Dynamics

Local sampling ~ evolution?

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

Conservation

Coevolution

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

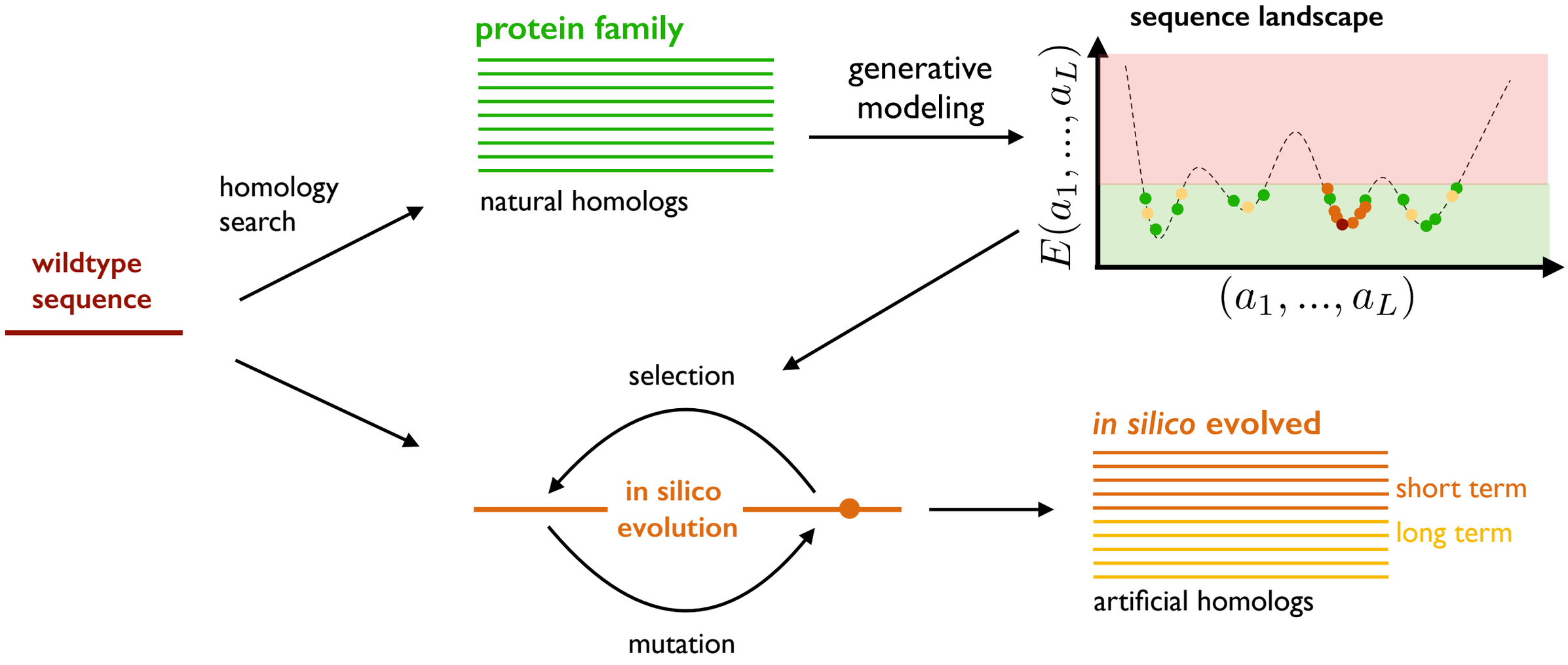
De la Paz et al., PNAS (2020)

a_i can be drawn:

- From the set of all 20 amino acids
- From the set of amino acids corresponding to codons with one mutation
- With or without gaps

Bisardi, Rodriguez-Rivas, FZ, Weigt, Mol.Bio.Evo. (2022)
Di Bari, Bisardi, Cotogno, Weigt, FZ, PNAS (2024)

Local sampling ~ evolution?

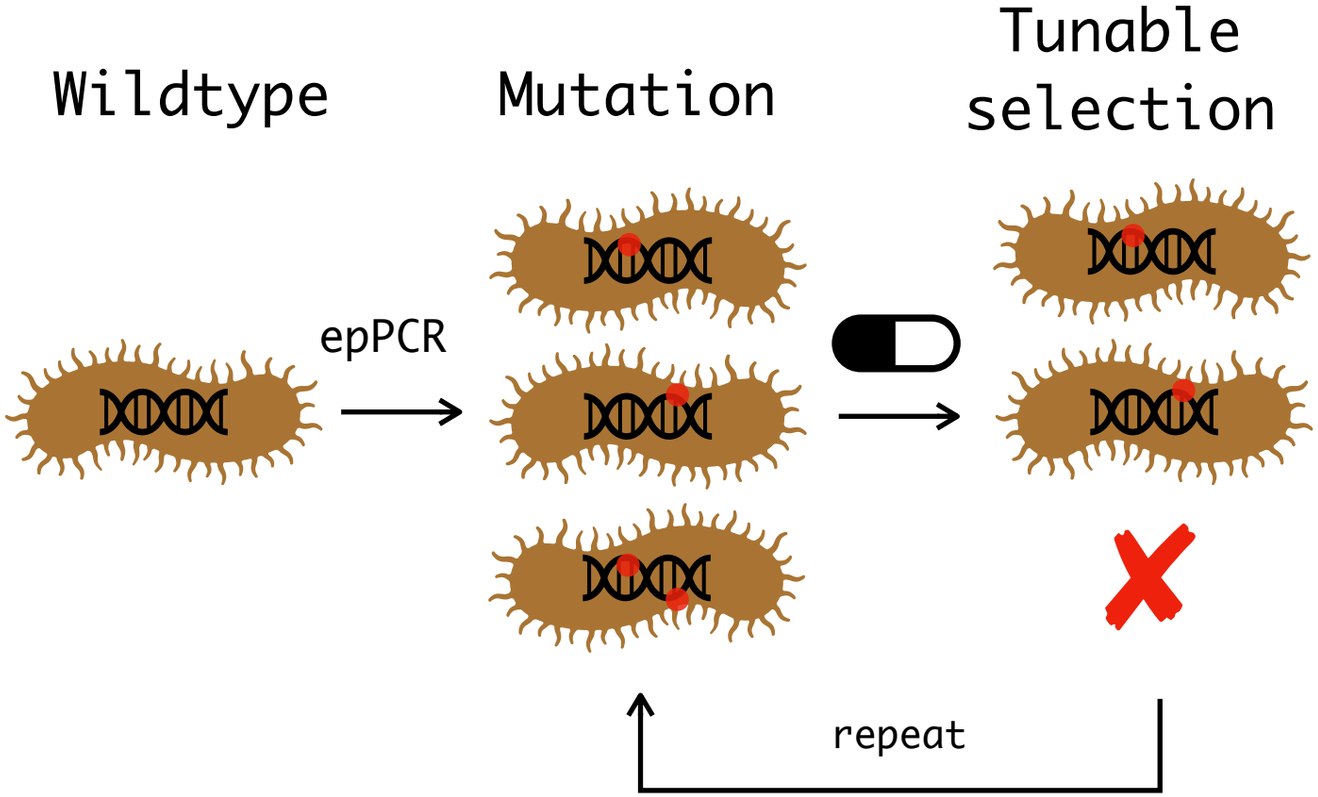


Natural evolution ≠ In-vitro evolution

Phylogenetic effects ≠ Independent chains (star phylogeny)

Bisardi, Rodriguez-Rivas, FZ, Weigt, Mol.Bio.Evo. (2022)
Di Bari, Bisardi, Cotogno, Weigt, FZ, PNAS (2024)

Experimental data



TCEDVAN

TCWDVAN
CCEDVAN
TCEDDDN

TCWDVAN
CCEDVAN

TCWDVN
TKEDDDN
TMIDDNN
TKIDDNN
TCIDVAN

Final library

- Wildtype: PSE-1 (202 aa)
- Sequences: 165k
- Distance from PSE-1: 28

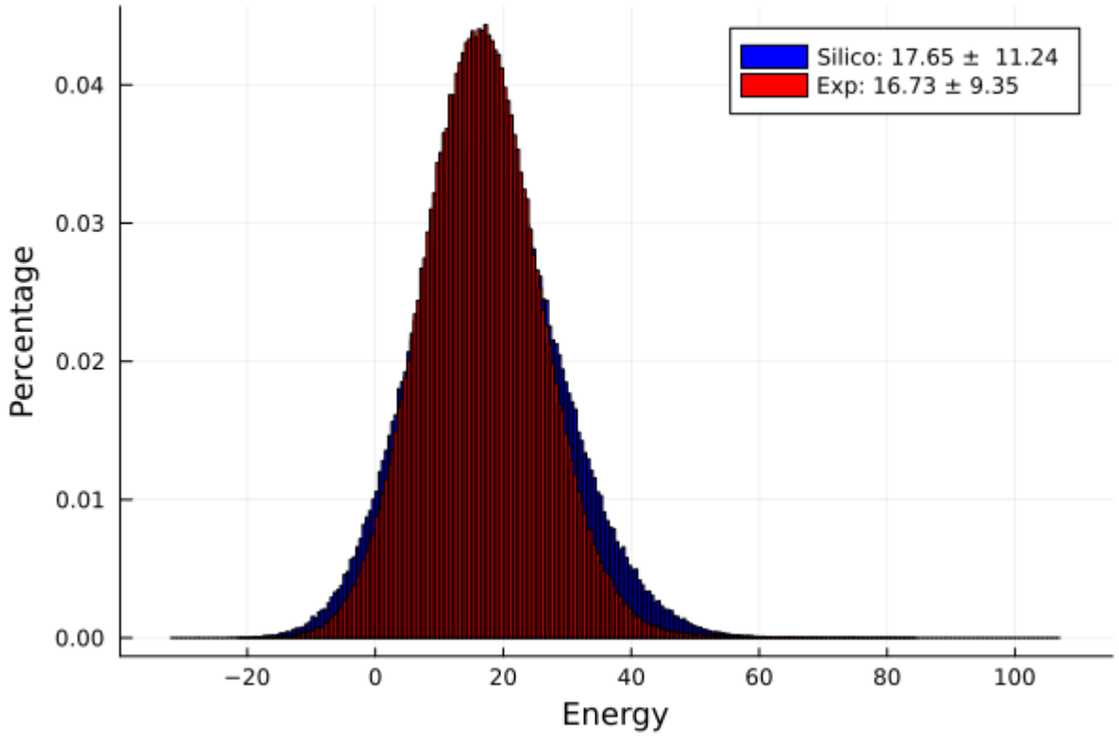
- Wildtype: TEM-1 (202 aa)
- Sequences: 35k
- Distance from TEM-1: 21

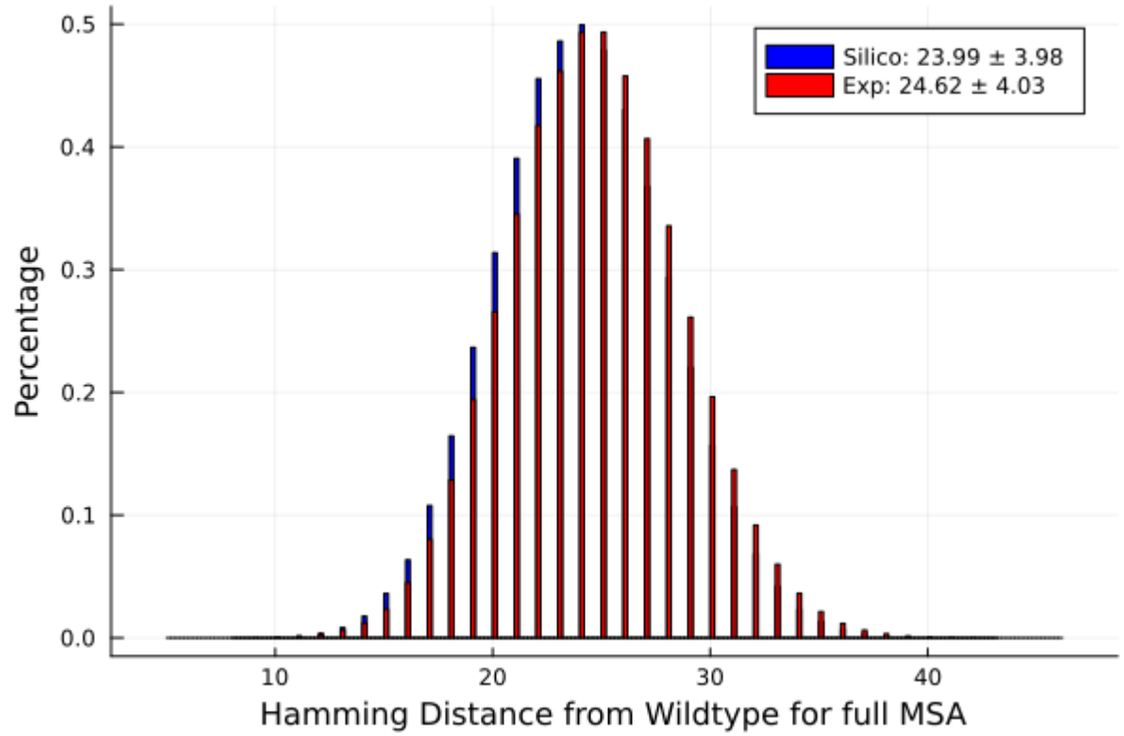
Stiffler et al., Cell Systems (2020)
Fantini et al., Mol.Biology and Evo. (2020)
New data (in progress): Erdogan et al., bioRxiv (2023)

Calibrating the model

Wildtype: PSE-1 Beta-lactamase, 20 generations of in vitro evolution

Stiffler et al., Cell Systems (2020)





Calibrate three parameters:

1. Artificial temperature that models selection strength
2. Number of MCMC steps from reference wild type
3. Number of sequences

Gibbs MCMC

draw i

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

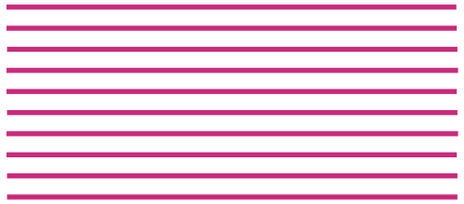
repeat

Short-term evolution

Calibrated:
- temperature
- steps
- number of sequences

in silico evolved short times

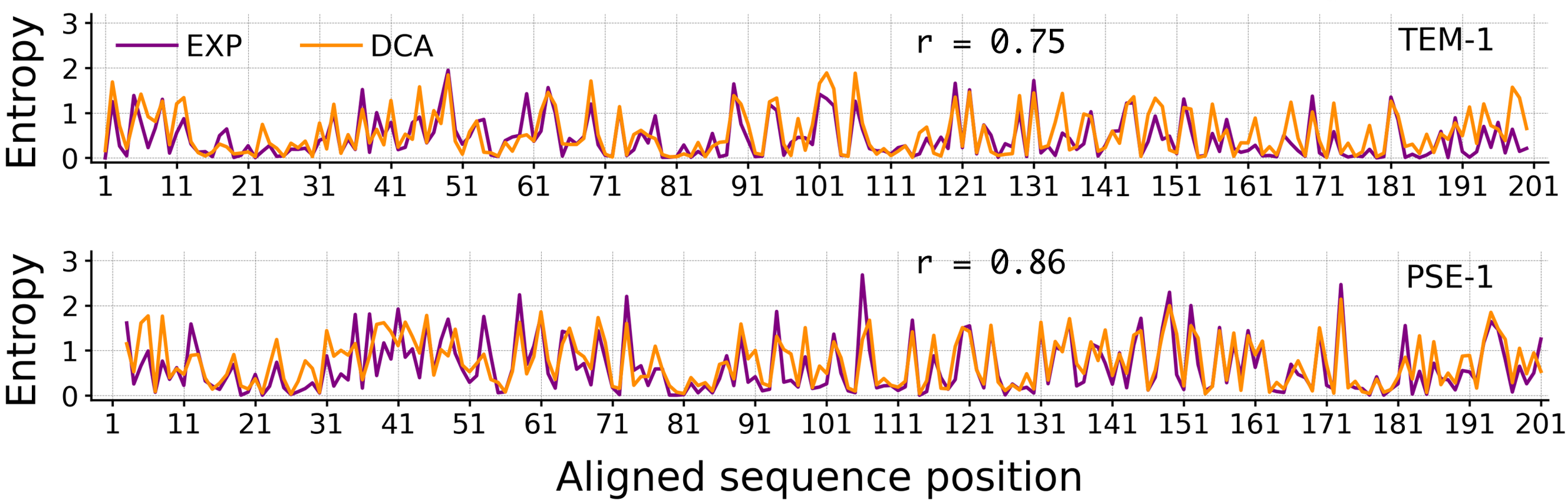
genetic drift experiment



compare

$f_i^{sil}(a_i)$ $f_i^{exp}(a_i)$ ⟶ Entropies

$$S_i = -\sum_a f_i(a) \log_2 f_i(a)$$



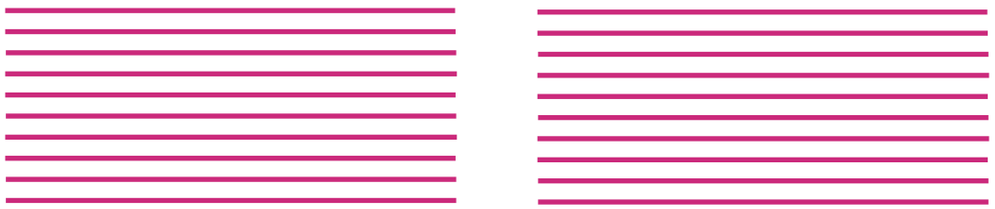
Predict ~400 numbers with 3 fitted parameters

Coevolutionary couplings are needed

genetic drift experiment

TEM-1

PSE-1

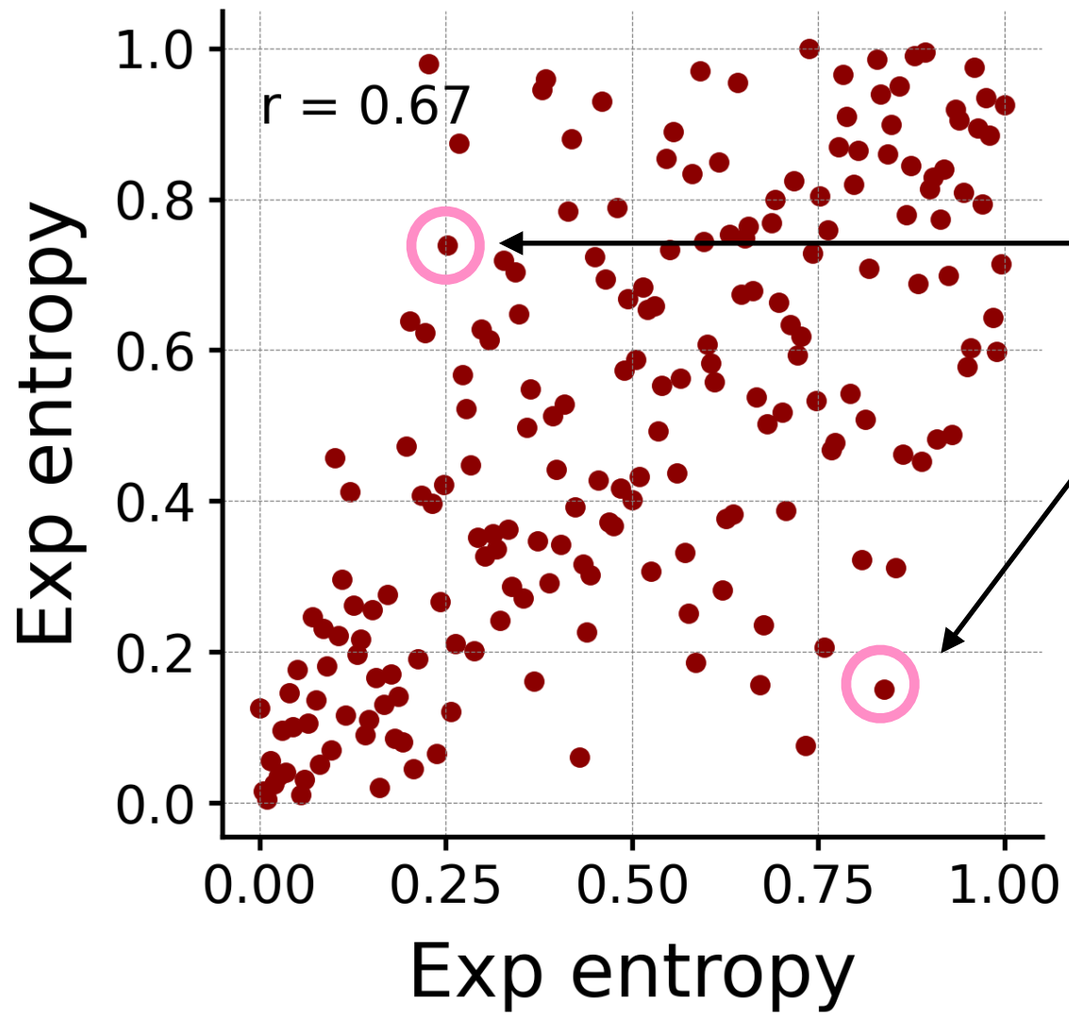


compare

$f_i^{TEM}(a_i)$ $f_i^{PSE}(a_i)$

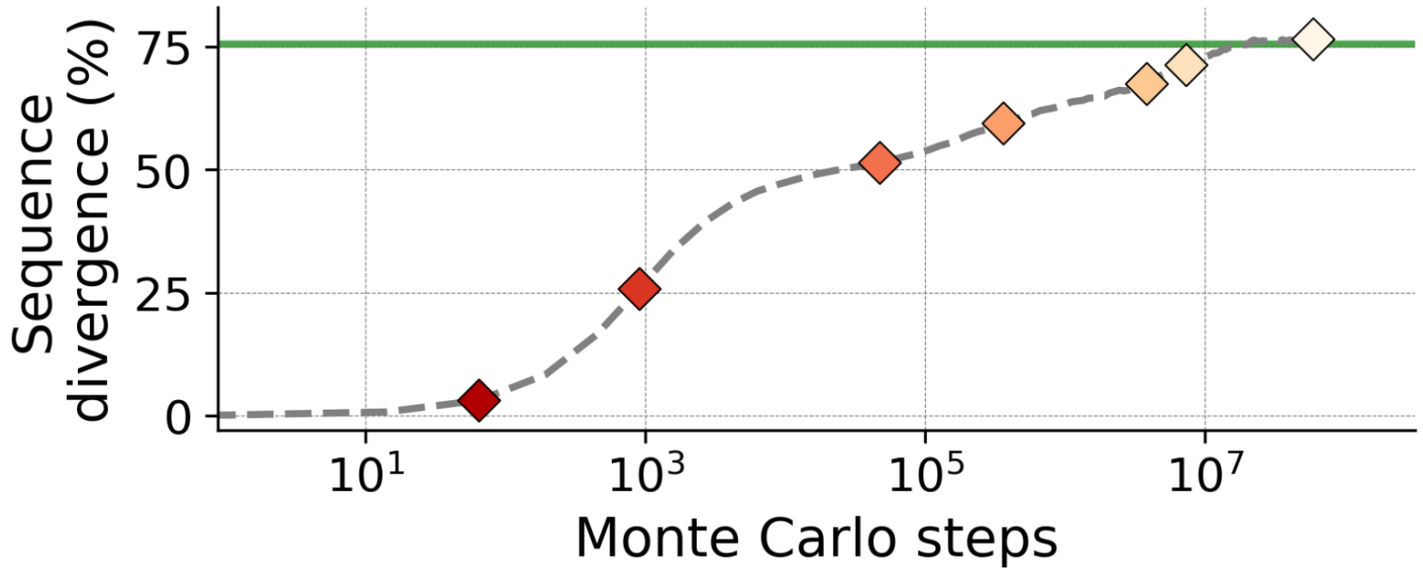
Entropies

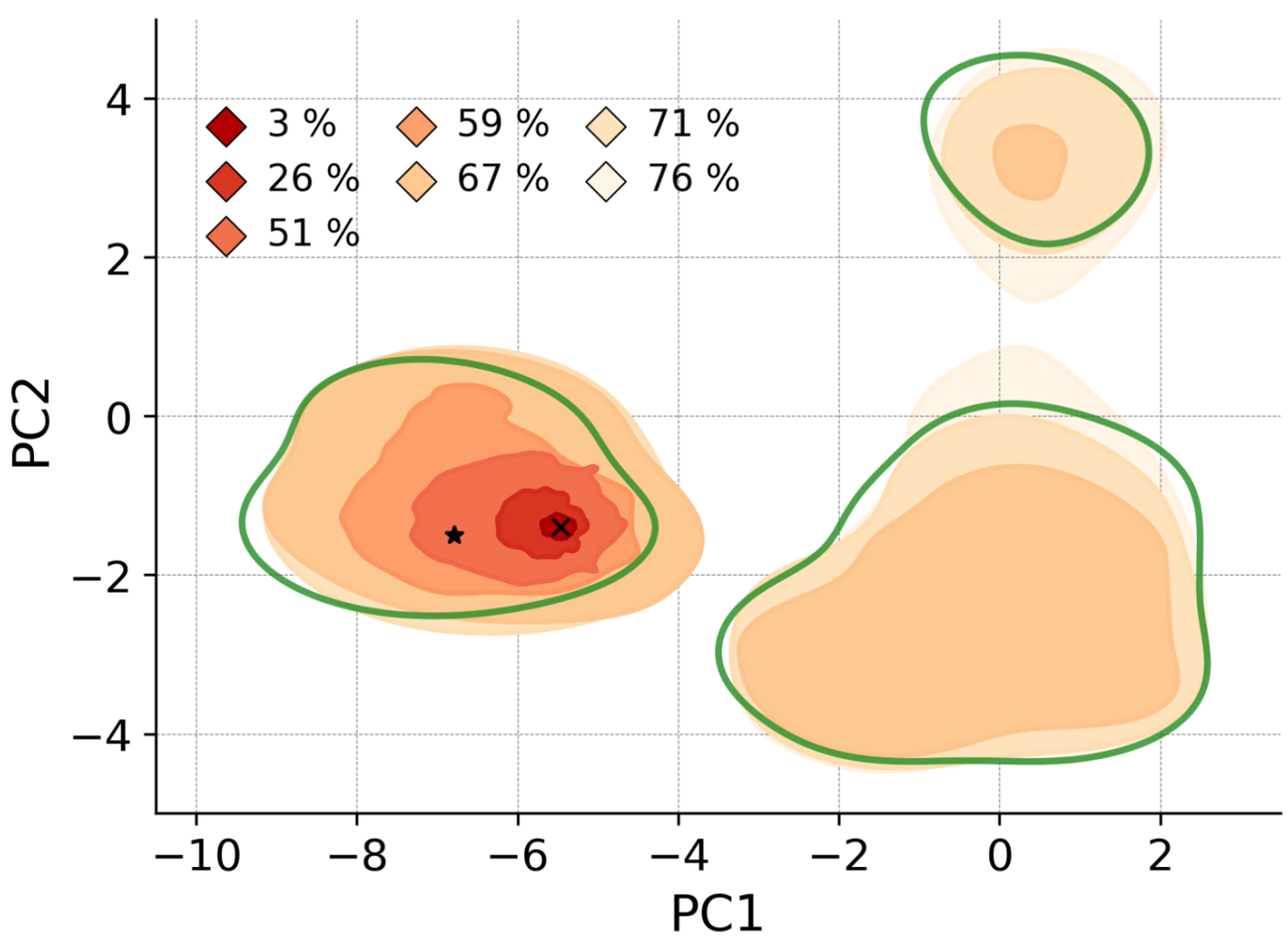
$$S_i = -\sum_a f_i(a)\,\log_2 f_i(a)$$

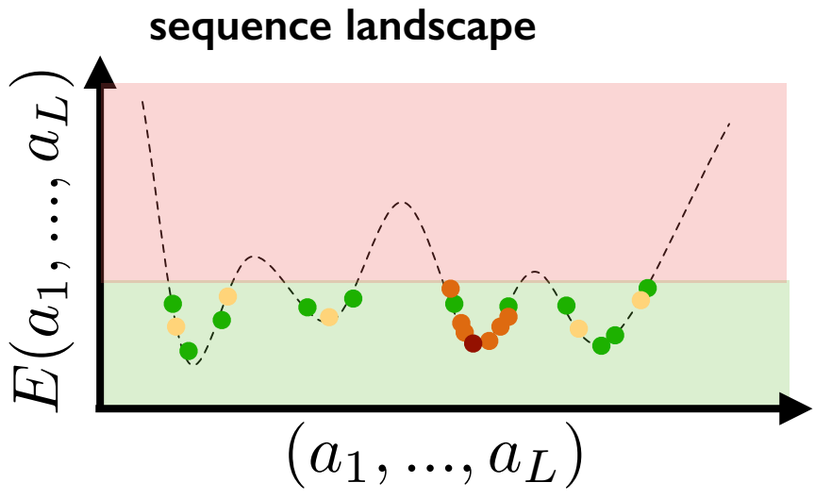


Entropies can be quite different: epistasis

Long-term evolution

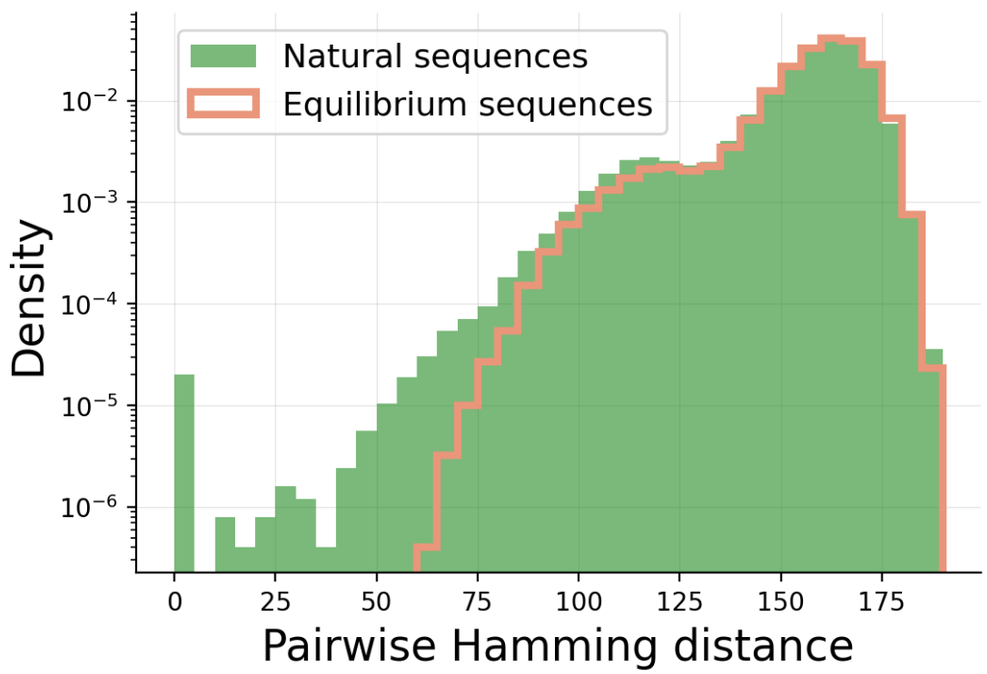






Sampling:

- Start from a wt sequence
- Sample 1000 trajectories
- Save the MSA at various times



Take home:

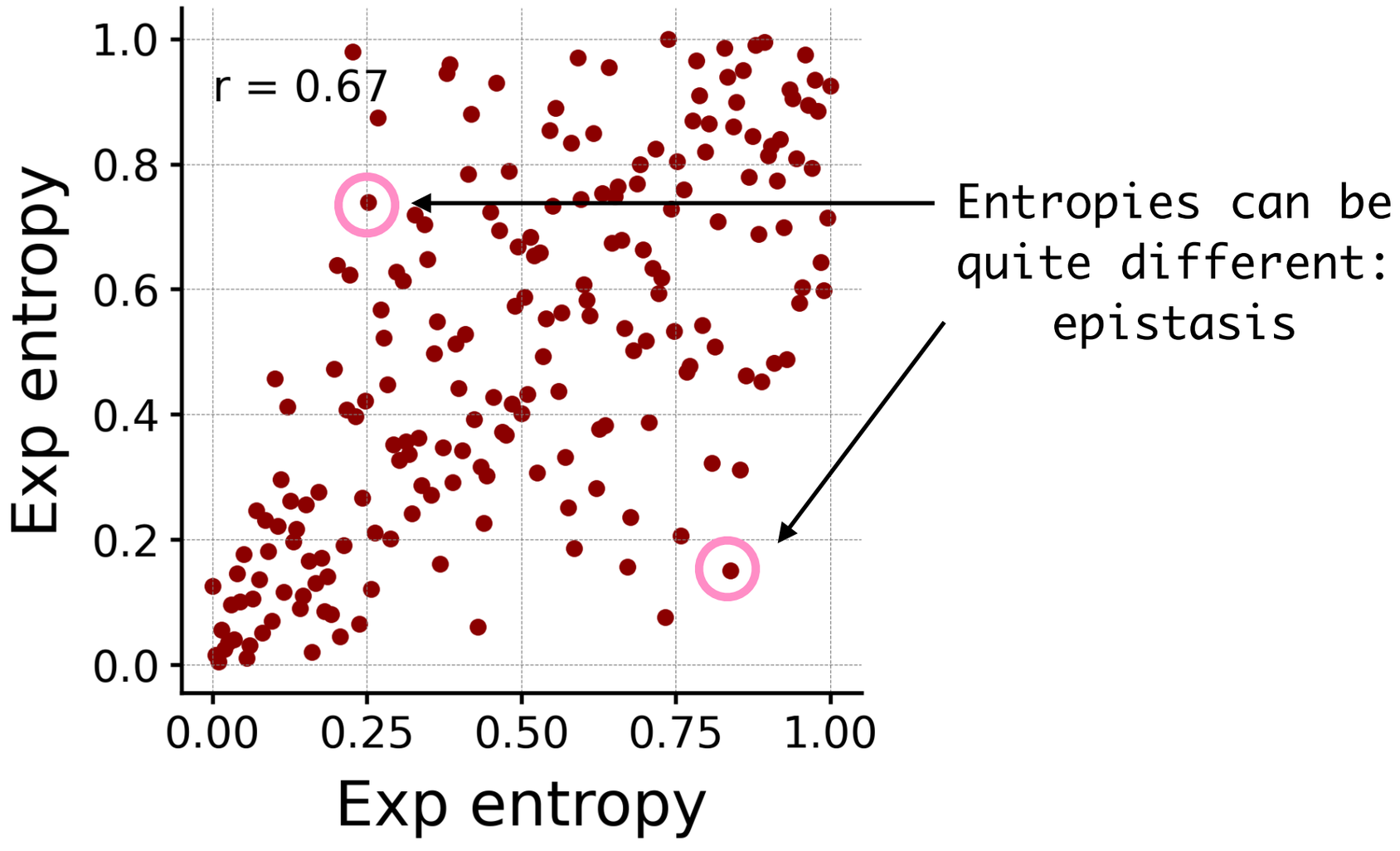
Monte Carlo dynamics of generative models

can accurately describe in vitro evolution (short time, dynamics)

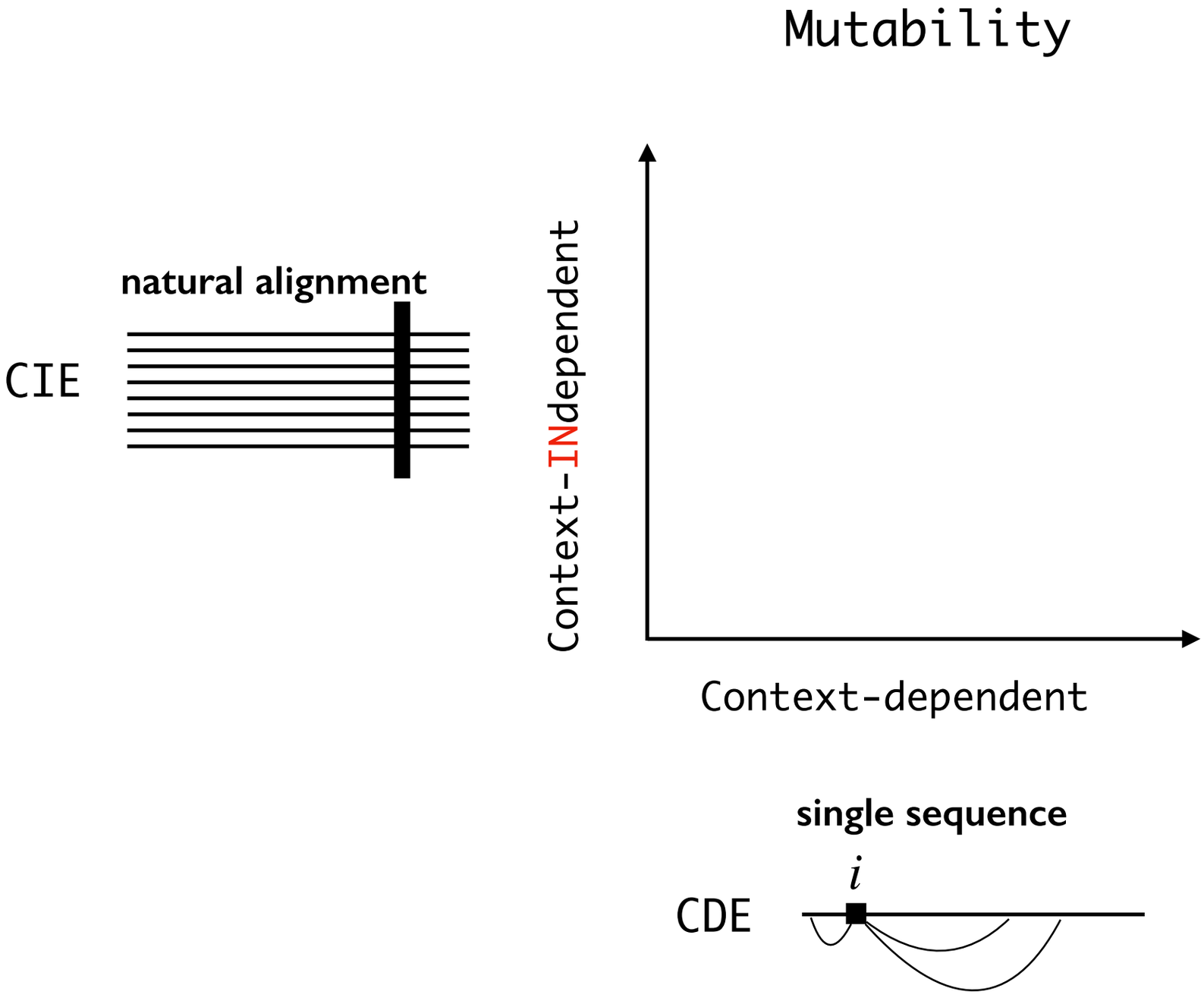
Epistasis

Epistasis

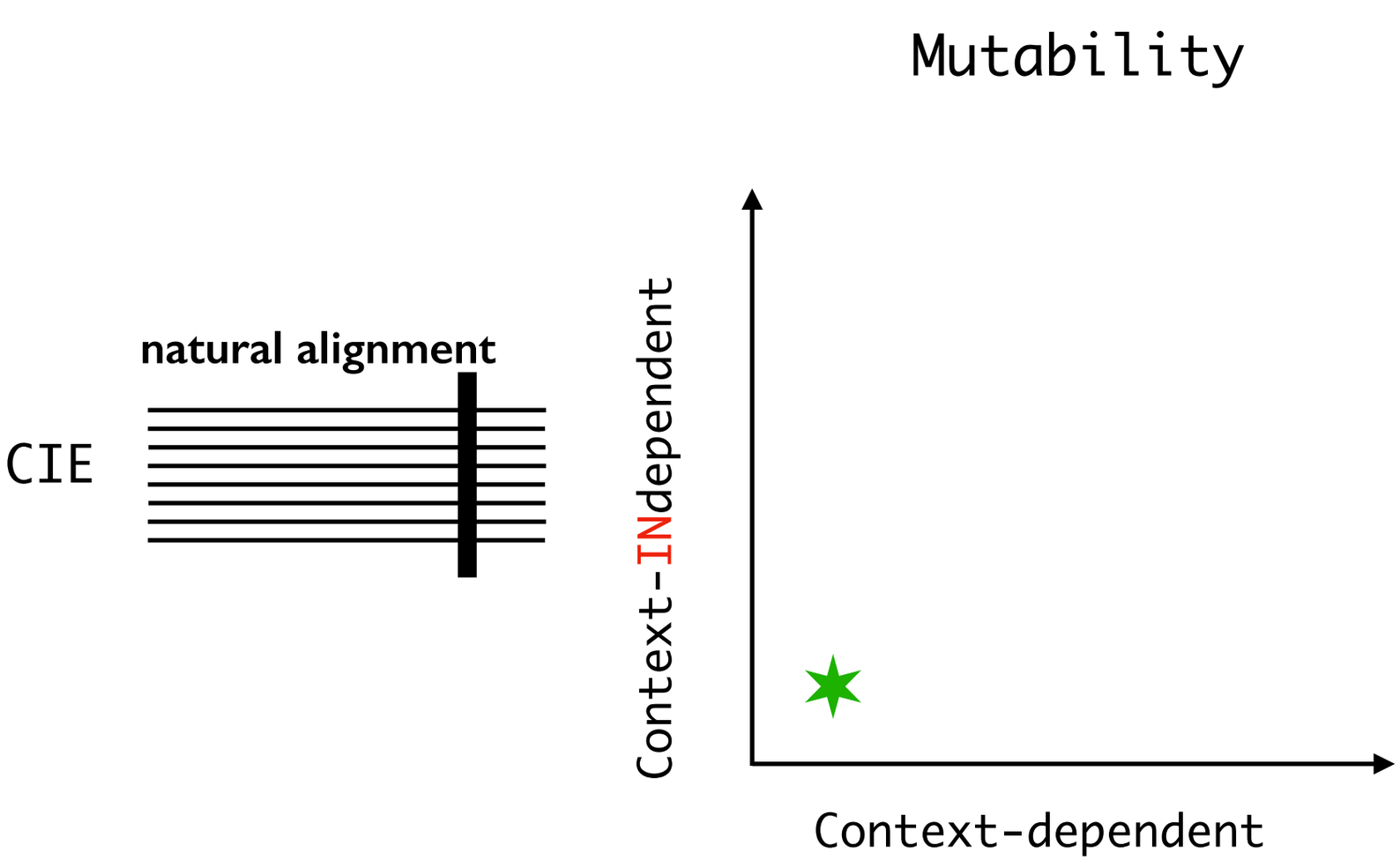
Context dependence of single mutations (related to coevolution)

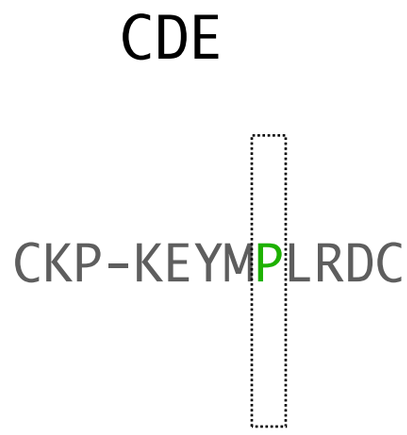


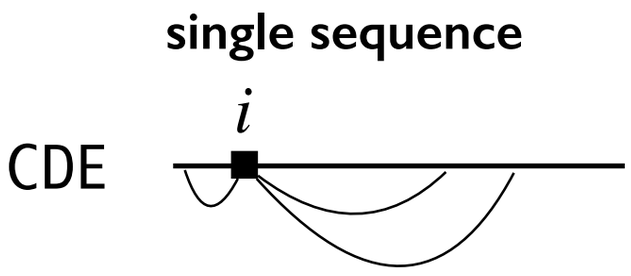
Quantifying epistasis



Quantifying epistasis



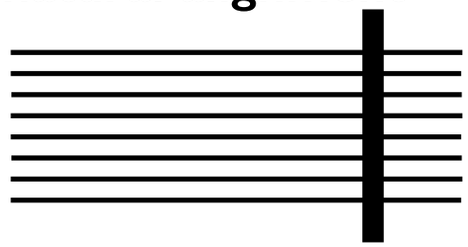




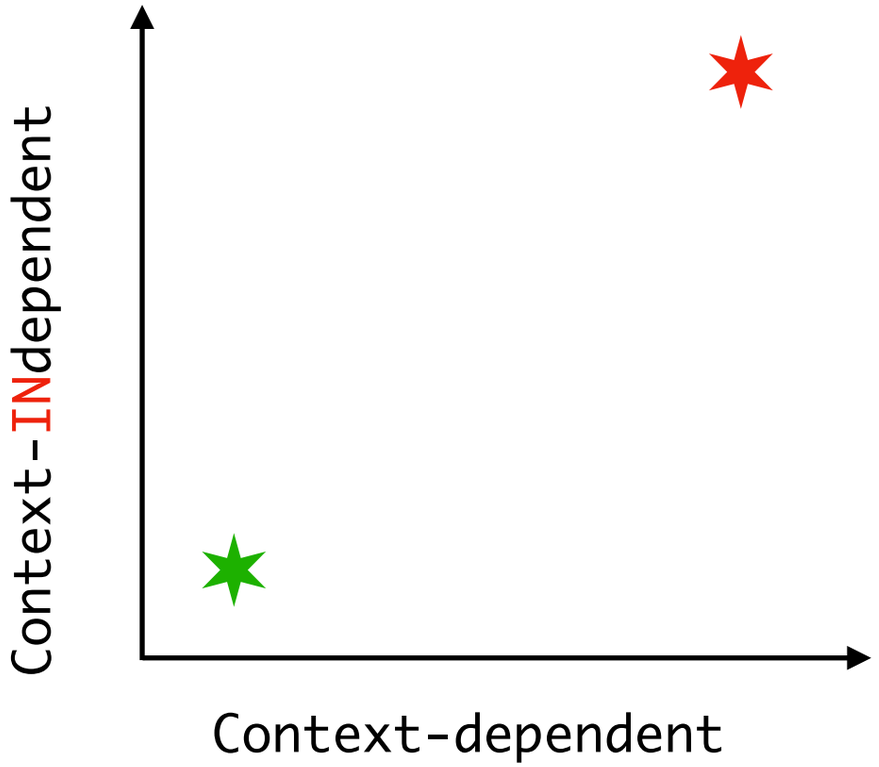
Quantifying epistasis

CIE

natural alignment



Mutability



CIE

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

CDE

```
CKP-KEYMPLRDC
```

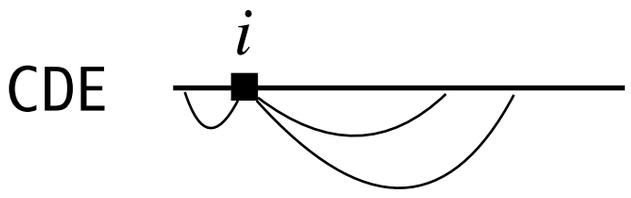
```
CKP-KEYMPLRDC
```

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

```
CKP-KEYMPLRDC
```

CDE

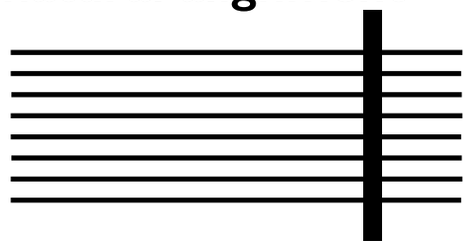
single sequence



Quantifying epistasis

CIE

natural alignment



Mutability

Context-INdependent

Context-dependent

CIE

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

CDE

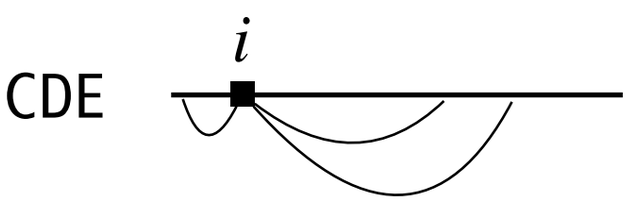
```
CKP-KEYMPLRDC
```

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

```
CKP-KEYMPLRDC
TRCAYV
```

CDE

single sequence

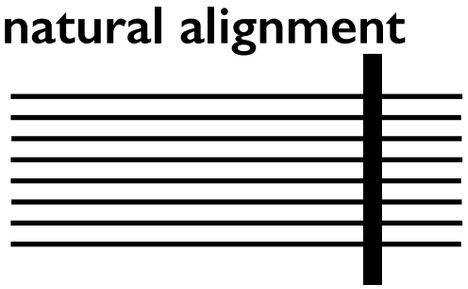


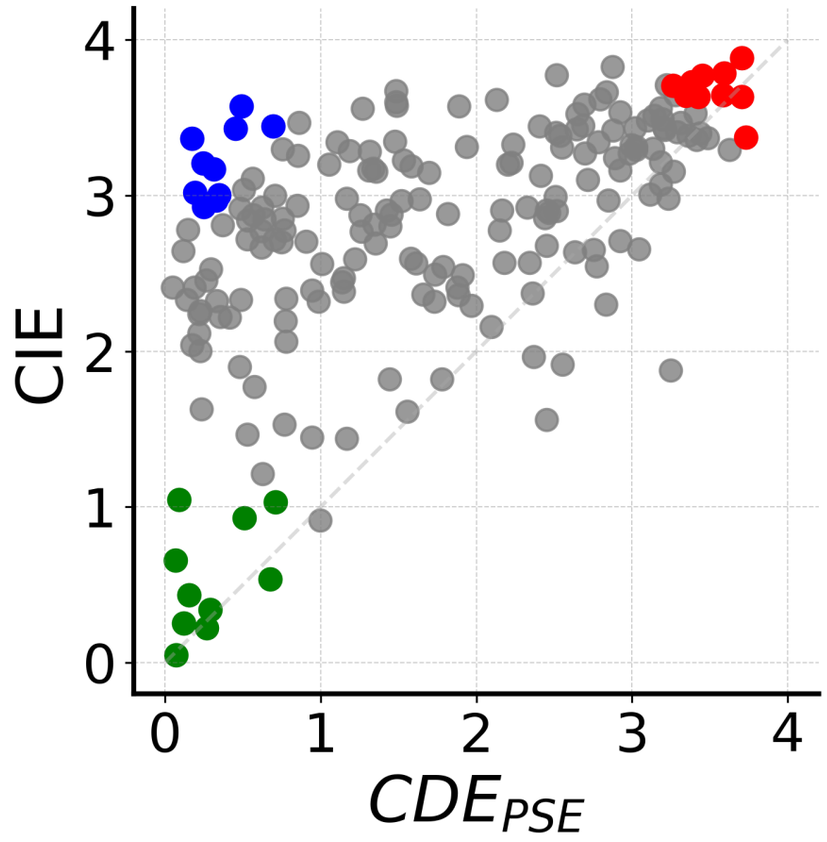

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

```
CKP-KEYMPLRDC
```

Quantifying epistasis

CIE





CIE

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

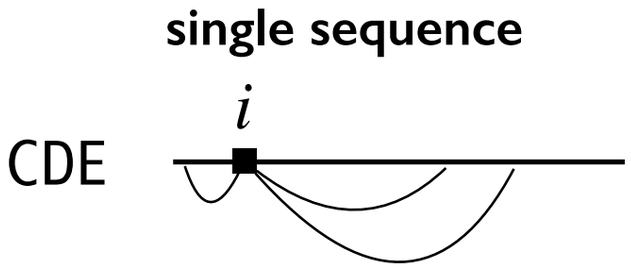
```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

CDE

```
CKP-KEYMPLRDC
```

```
CKP-KEYMPLRDC
```

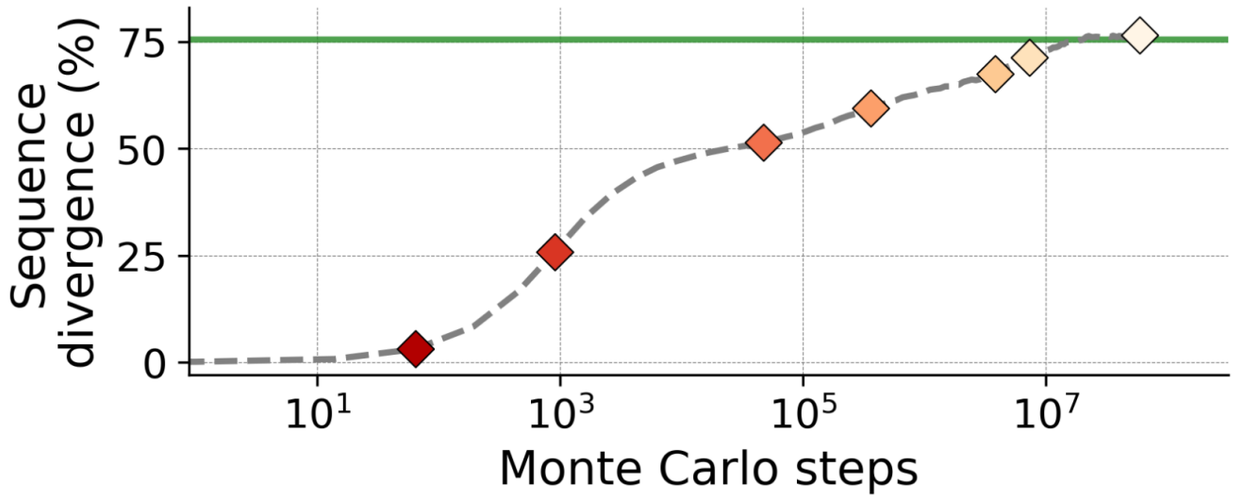
CDE

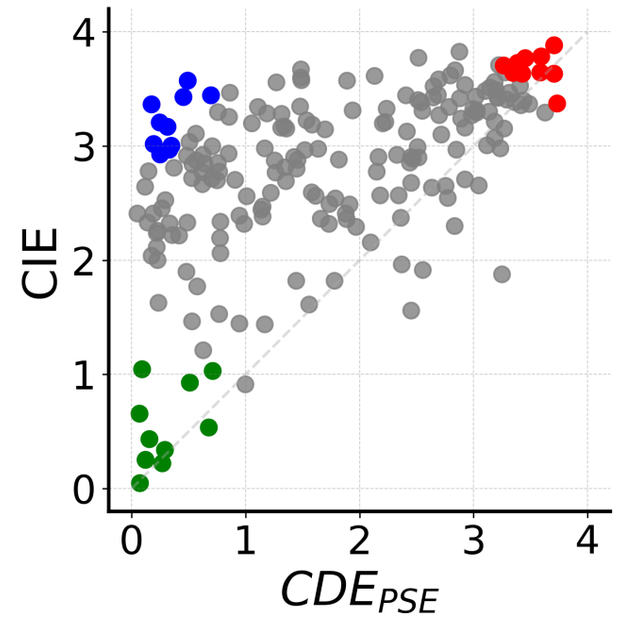


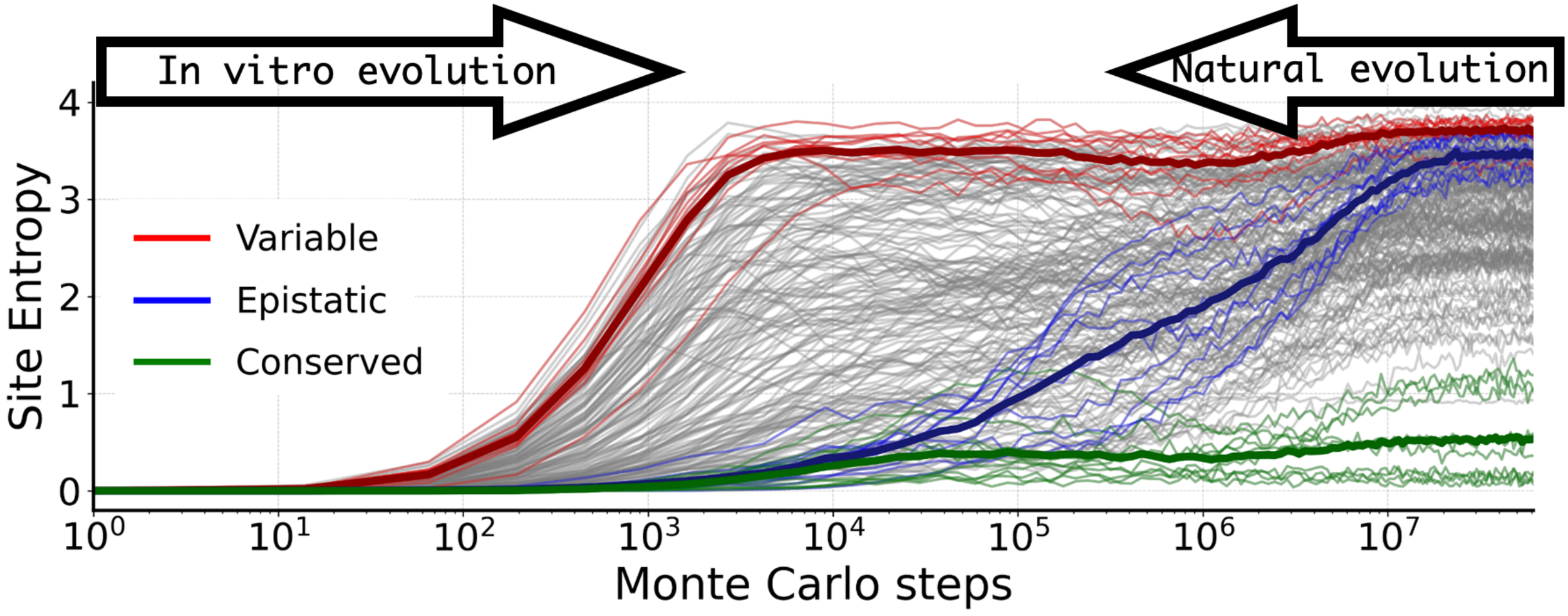

```
CKP-KEYMPLRDC
CREDKPYNPLRER
CWE-VPNNPLTEF
TQADVAWMPETGF
TCSQKA--PVRGT
ACADVAYSPVTGY
WCEDVANSPVTEV
```

```
CKP-KEYMPLRDC
```

Emergence of epistasis







Di Bari, Bisardi, Cotogno, Weigt, FZ, PNAS (2024)

Take home:

Epistasis is very heterogeneous across sites

It emerges on intermediate time scales due to collective effects

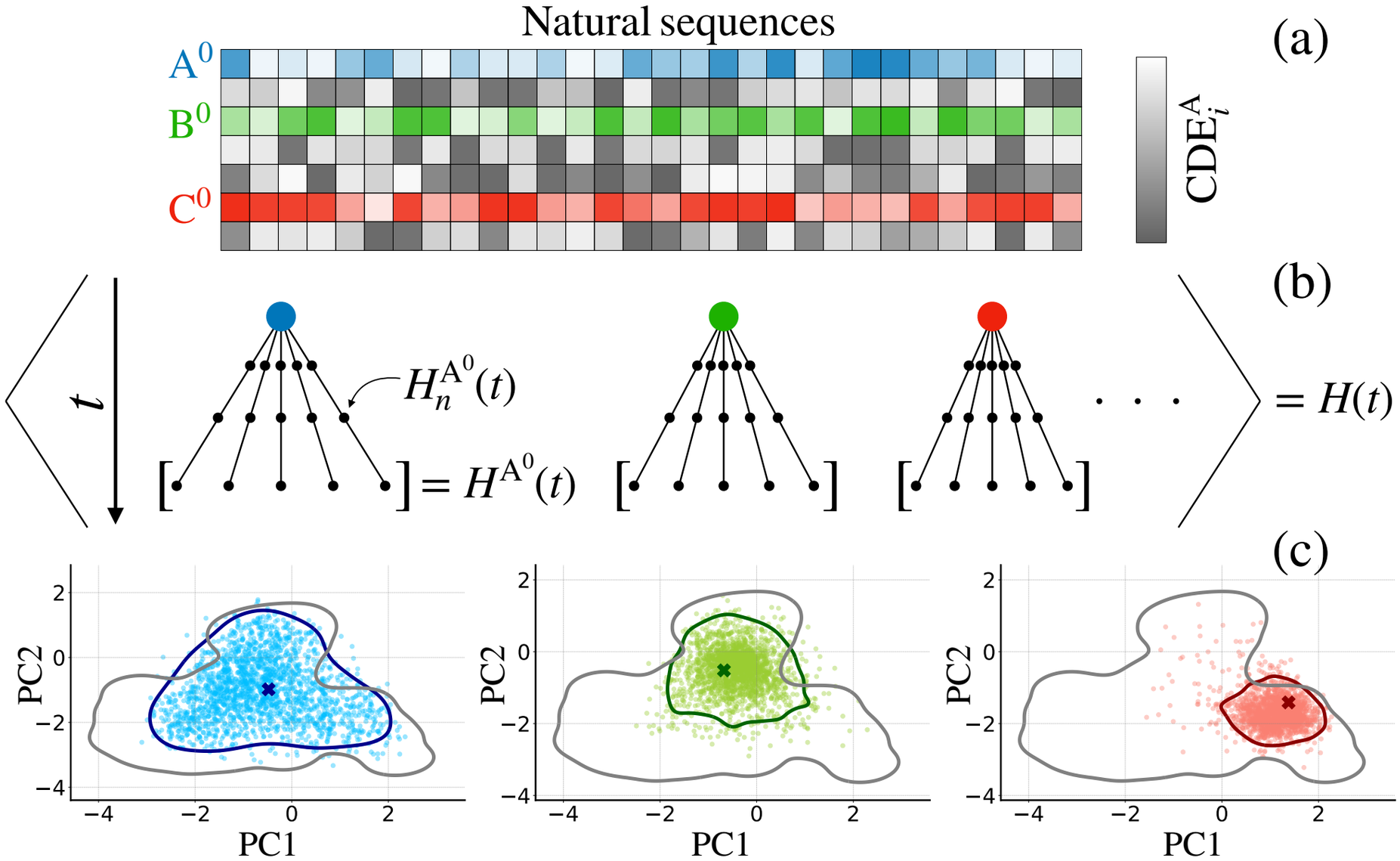
Measures of heterogeneity

Saverio
Rossi

Leonardo
Di Bari

Rossi, Di Bari, Weigt, FZ, arXiv:2412.01969 (2024)

Different sources of fluctuations



$H =$ number of mutations (Hamming dist.)

$[\,\bullet\,]$ over dynamical fluctuations

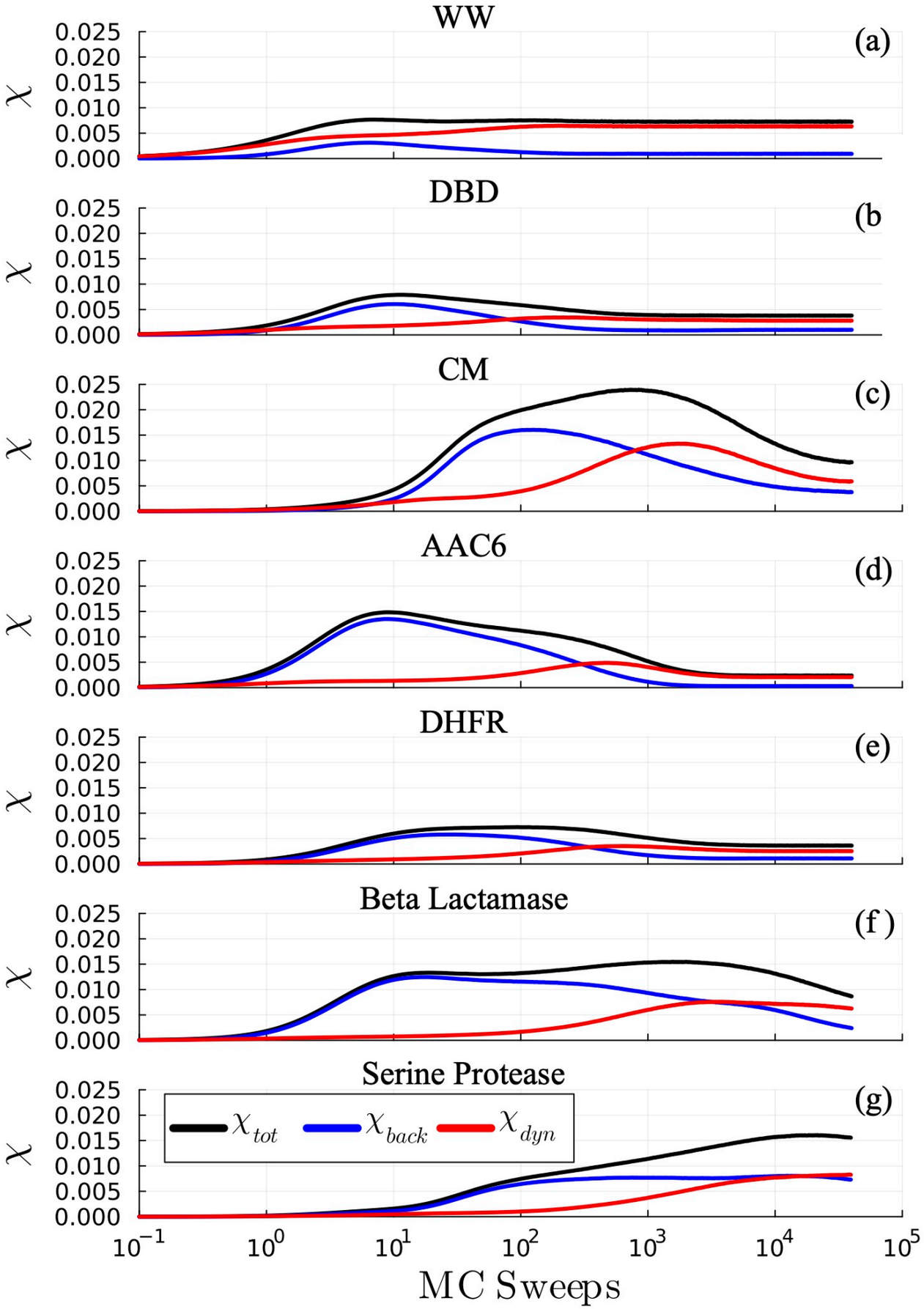
$\langle\,\bullet\,\rangle$ over initial sequence A

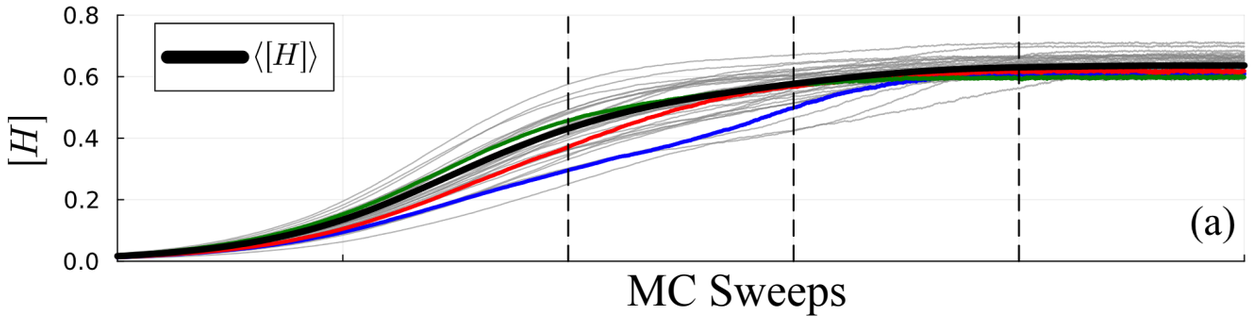
$$\chi_{dyn} = \langle \chi^A_{dyn} \rangle = \langle [H^2] - [H]^2 \rangle$$

$$\chi_{back} = \langle [H]^2 \rangle - \langle [H] \rangle^2$$

$$\chi_{tot} = \langle [H^2] \rangle - \langle [H] \rangle^2 = \chi_{back} + \chi_{dyn}$$

Different sources of fluctuations





Dynamical fluctuations subdominant at short/intermediate times

They only take over at very long times

Evolution is partially predictable from initial condition

Different sources of fluctuations

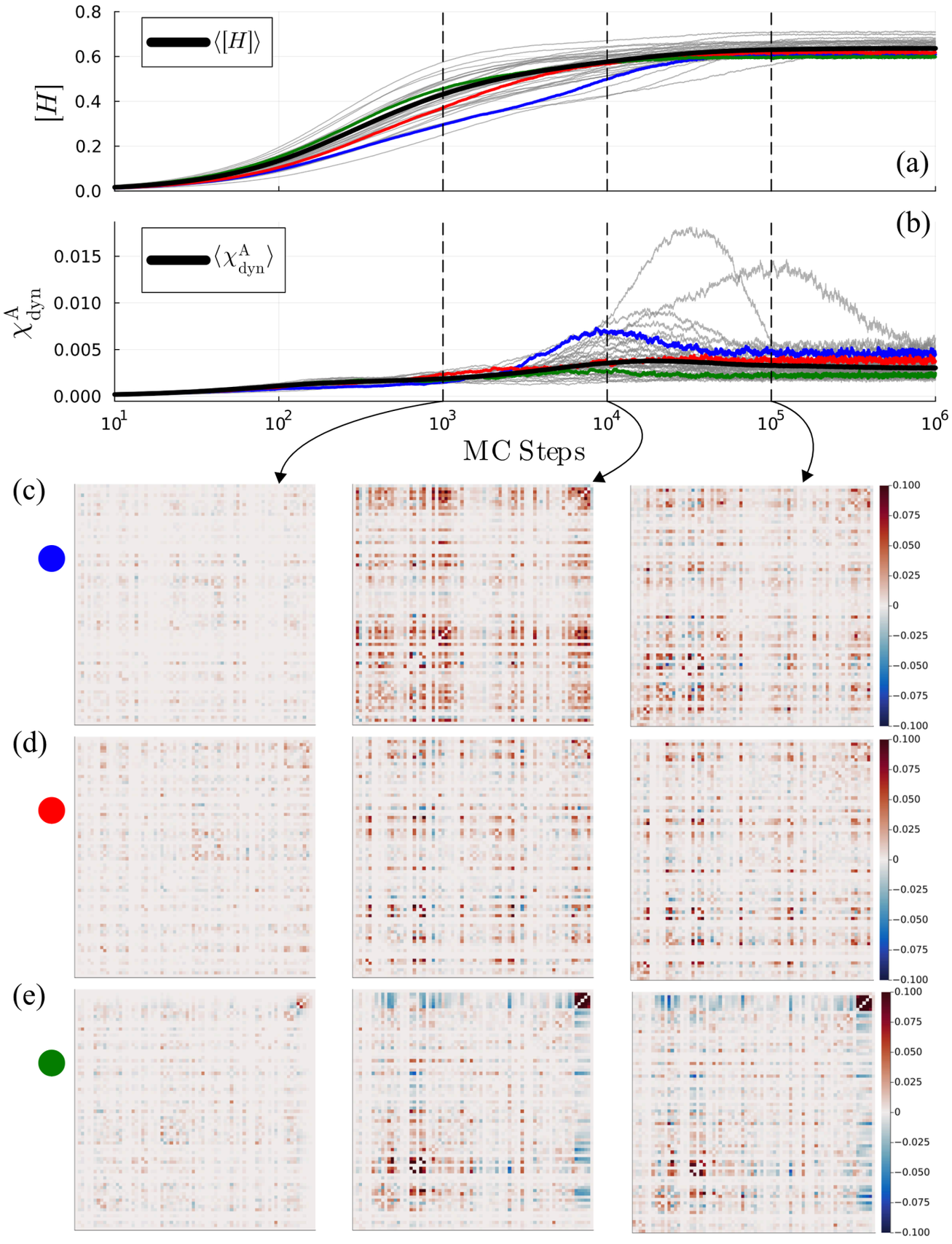
$$\chi^{A}_{dyn}(t,0) = \frac{1}{L^2}\sum_{ij} G^{\mathrm{A}}_{ij,\mathrm{dyn}}(t,0)$$

$$G^{\mathrm{A}}_{ij,\mathrm{dyn}}(t,0) = [\delta_{a_i^t,a_i^0}\delta_{a_j^t,a_j^0}] - [\delta_{a_i^t,a_i^0}][\delta_{a_j^t,a_j^0}]$$

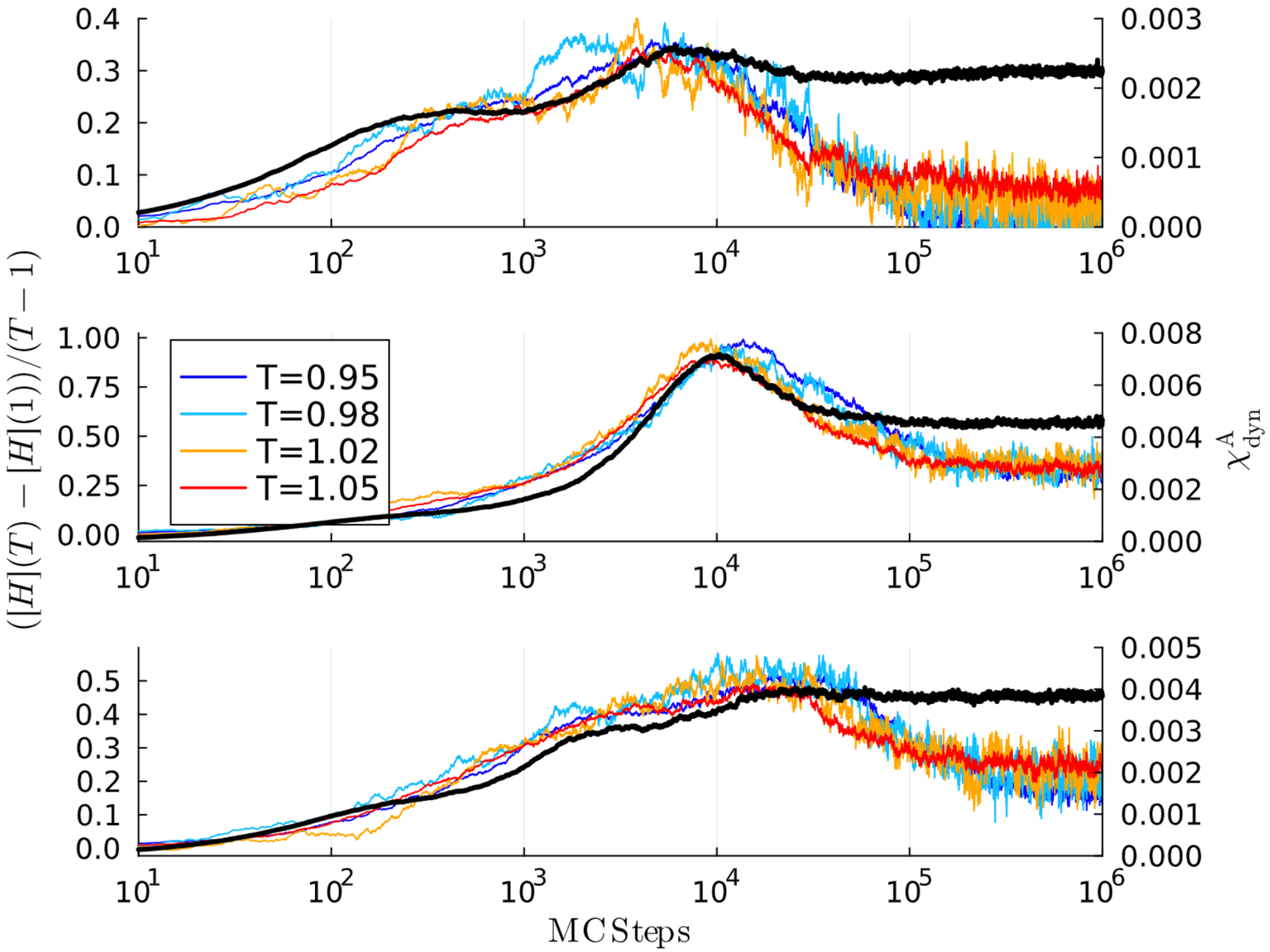
Clusters of strongly correlated sites

Depending on the initial sequence!

Sequence-specific dynamical ‘sectors’?



Fluctuation-dissipation relation



$$\chi^{A}_{dyn}(t,0) \approx \frac{d[H(t,0)]_A}{dT}$$

Berthier et al. Science (2005)

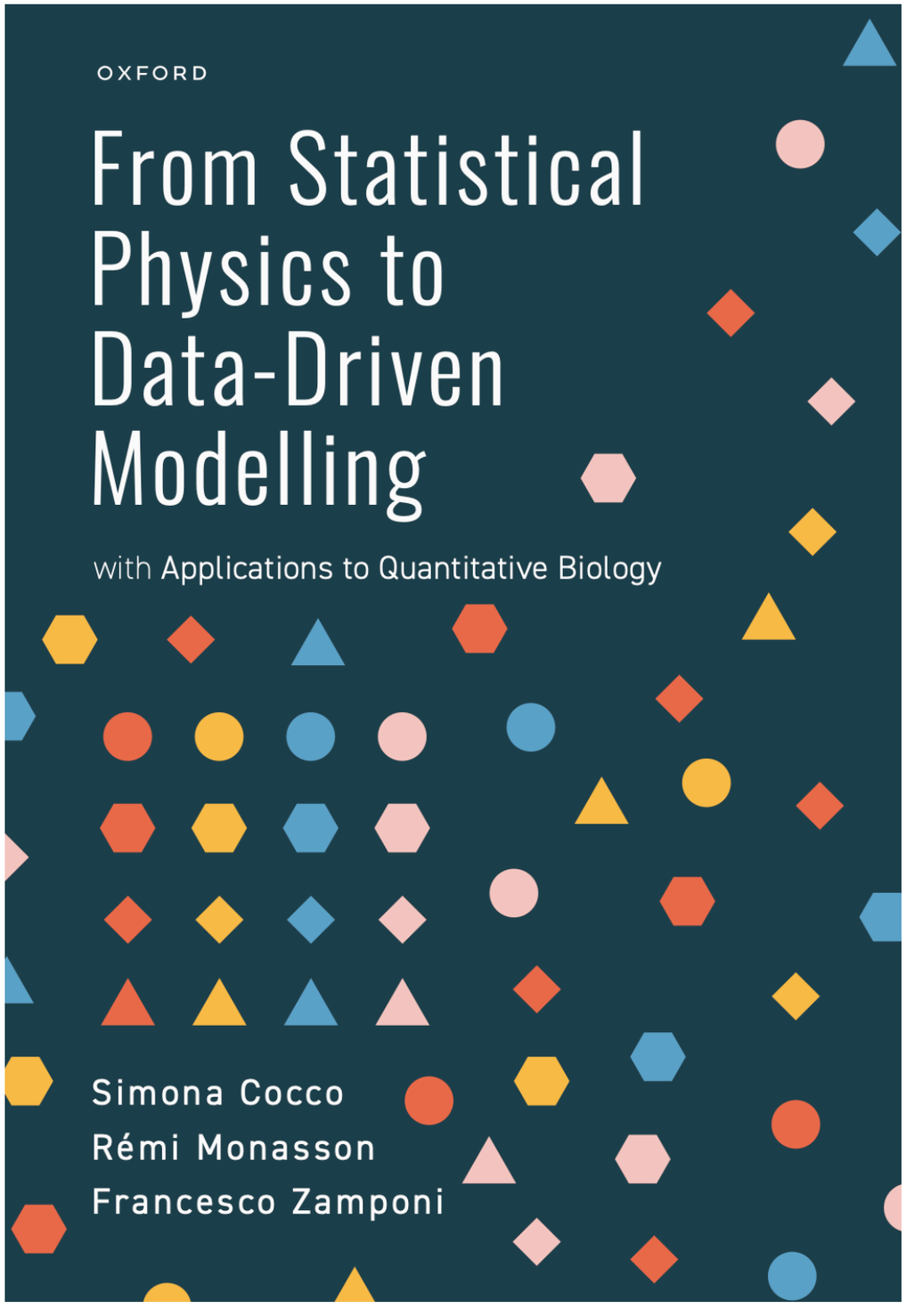
More heterogeneous sequences are more sensitive to perturbations
(remember: T = selection strength)

Take home:

Epistasis is very heterogeneous across sequences

We can quantify it using dynamical correlations and identify cooperatively rearranging sites

Advertisement



SAPIENZA
UNIVERSITÀ DI ROMA

Thank you for your attention!

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

Di Bari, Bisardi, Cotogno, Weigt, FZ, PNAS (2024)

Rossi, Di Bari, Weigt, FZ, arXiv:2412.01969 (2024)