# A machine-learning project:

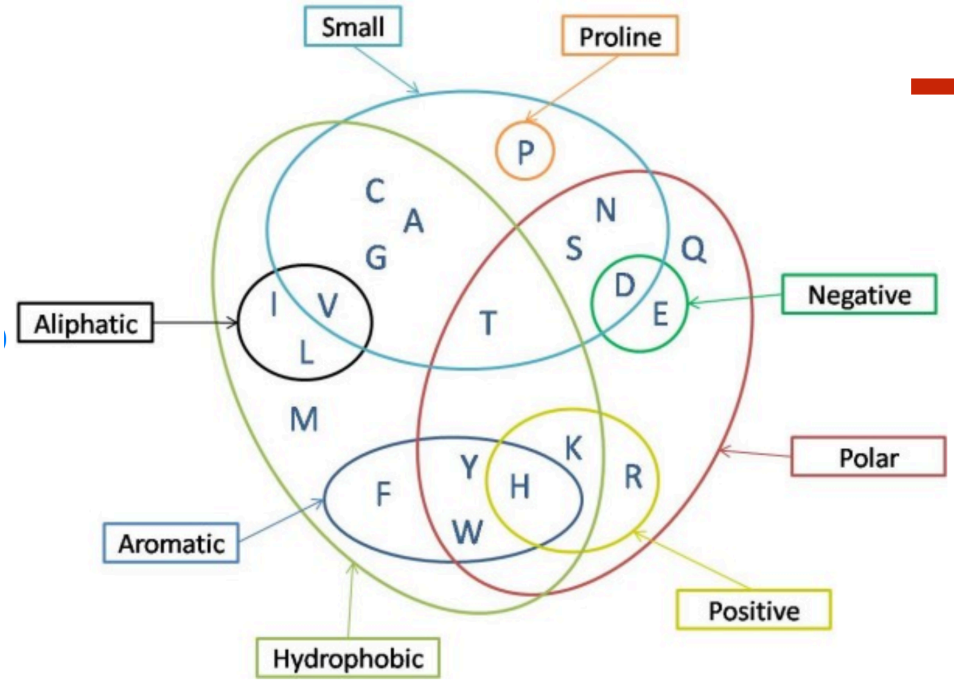
## unsupervised and supervised characterization of protein sequence data

Martin Weigt

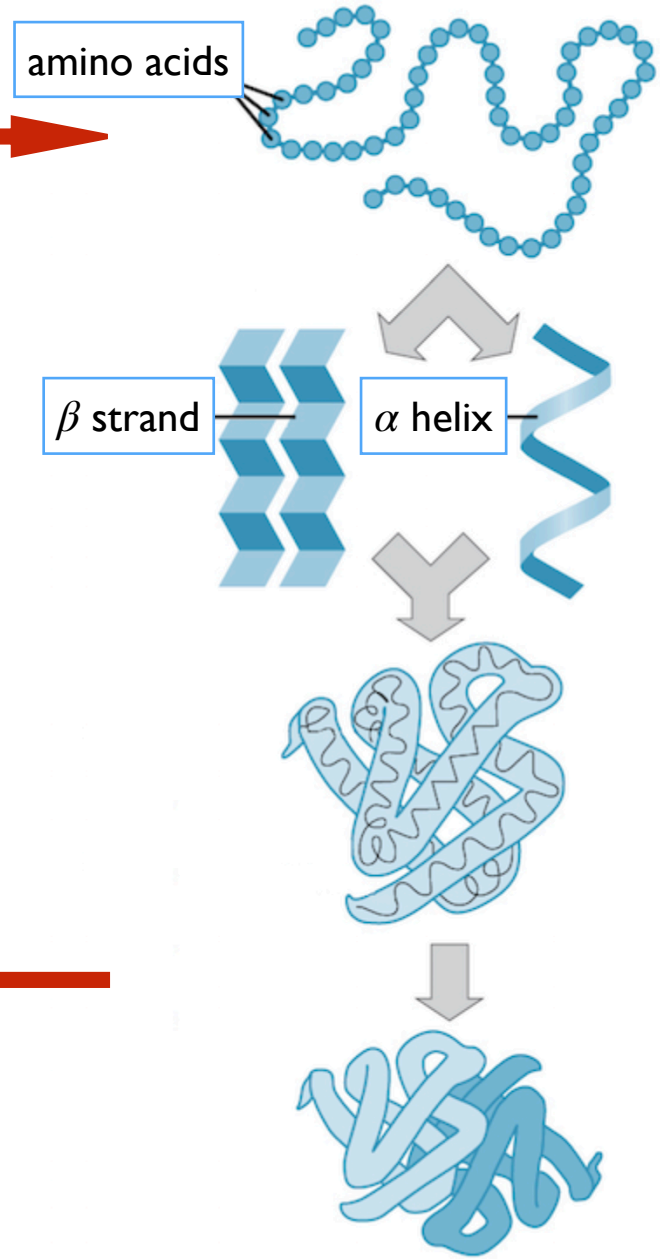
Laboratoire de Biologie Computationnelle et Quantitative
Sorbonne Université

# Proteins - workhorses of the cell

## Built from 20 amino acids (residues):




## Orders of protein structure:

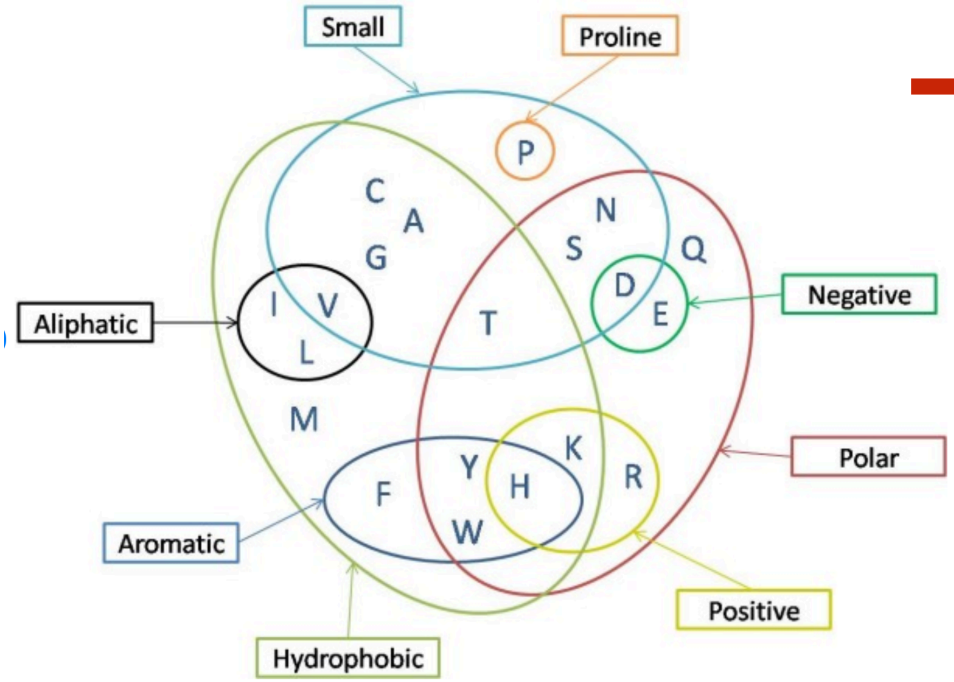



## Multitude of functions:

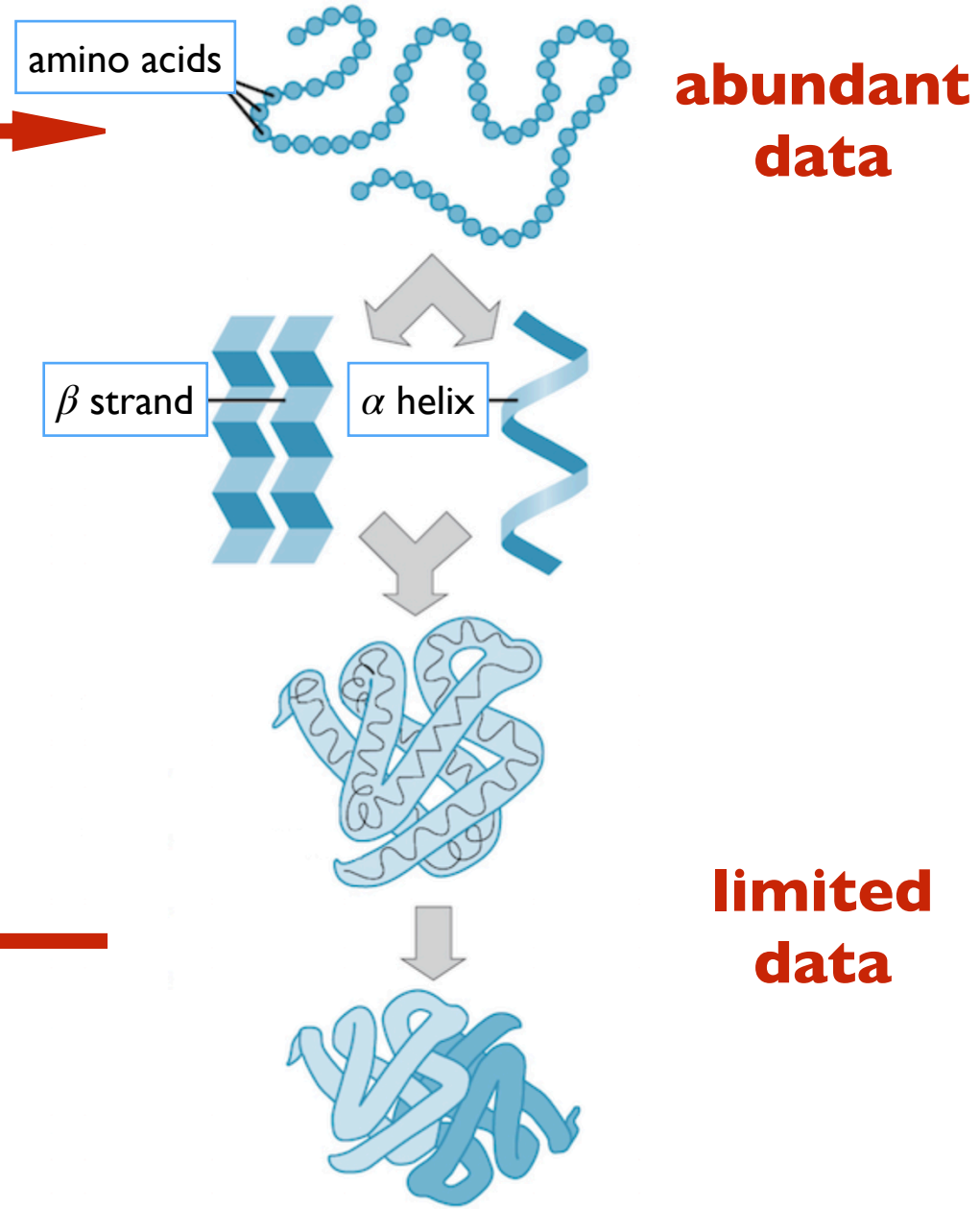
- enzymes & catalysis
- transport
- cellular structure
- signal transduction
- protein synthesis…

# Proteins - workhorses of the cell

## Built from 20 amino acids (residues):




## Orders of protein structure:




**abundant data**

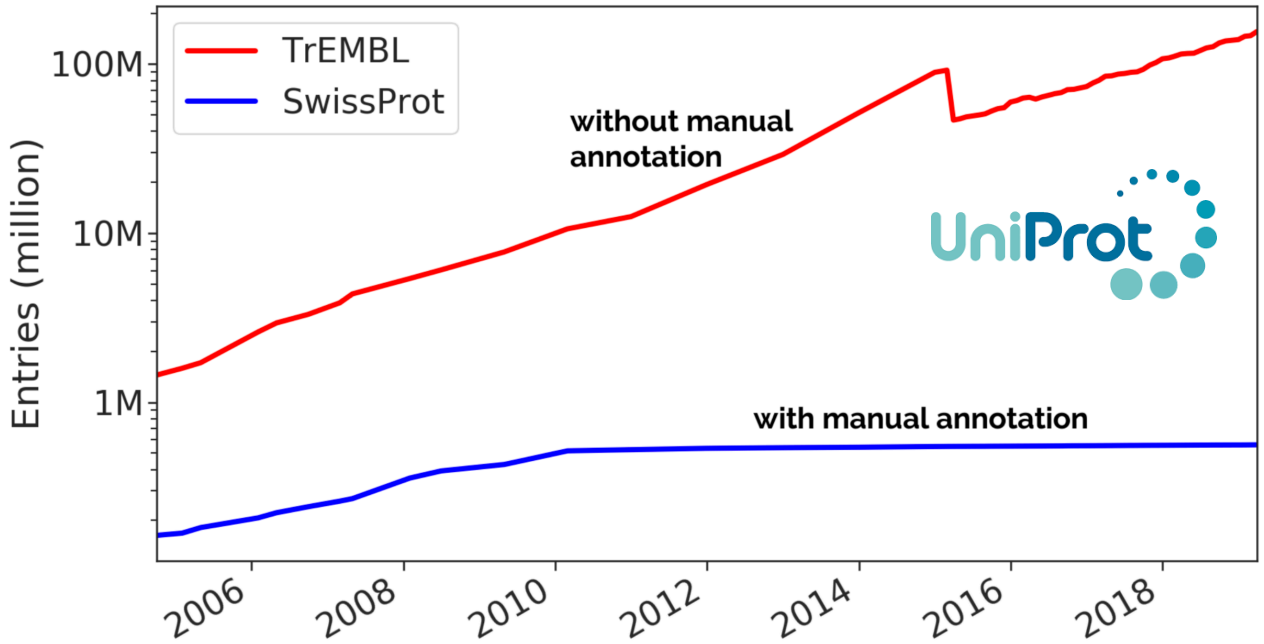
**limited data**

## Multitude of functions:

- enzymes & catalysis
- transport
- cellular structure
- signal transduction
- protein synthesis…

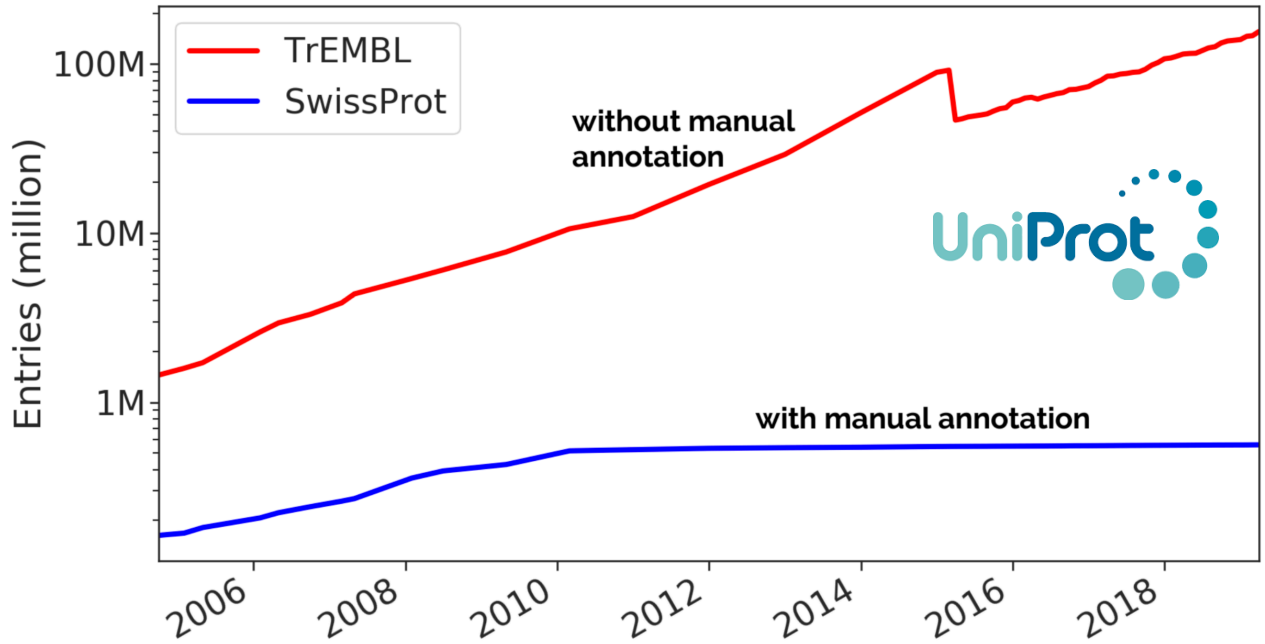
# Sequence data

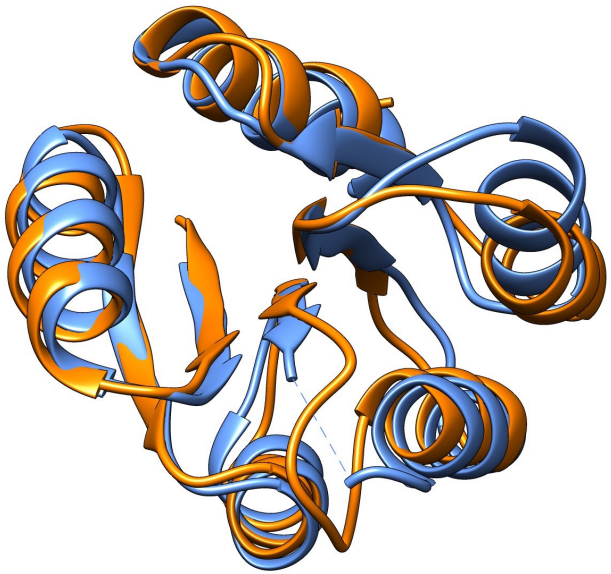
**Sequence data** are rapidly accumulating…



# Sequence data

**Sequence data** are rapidly accumulating…

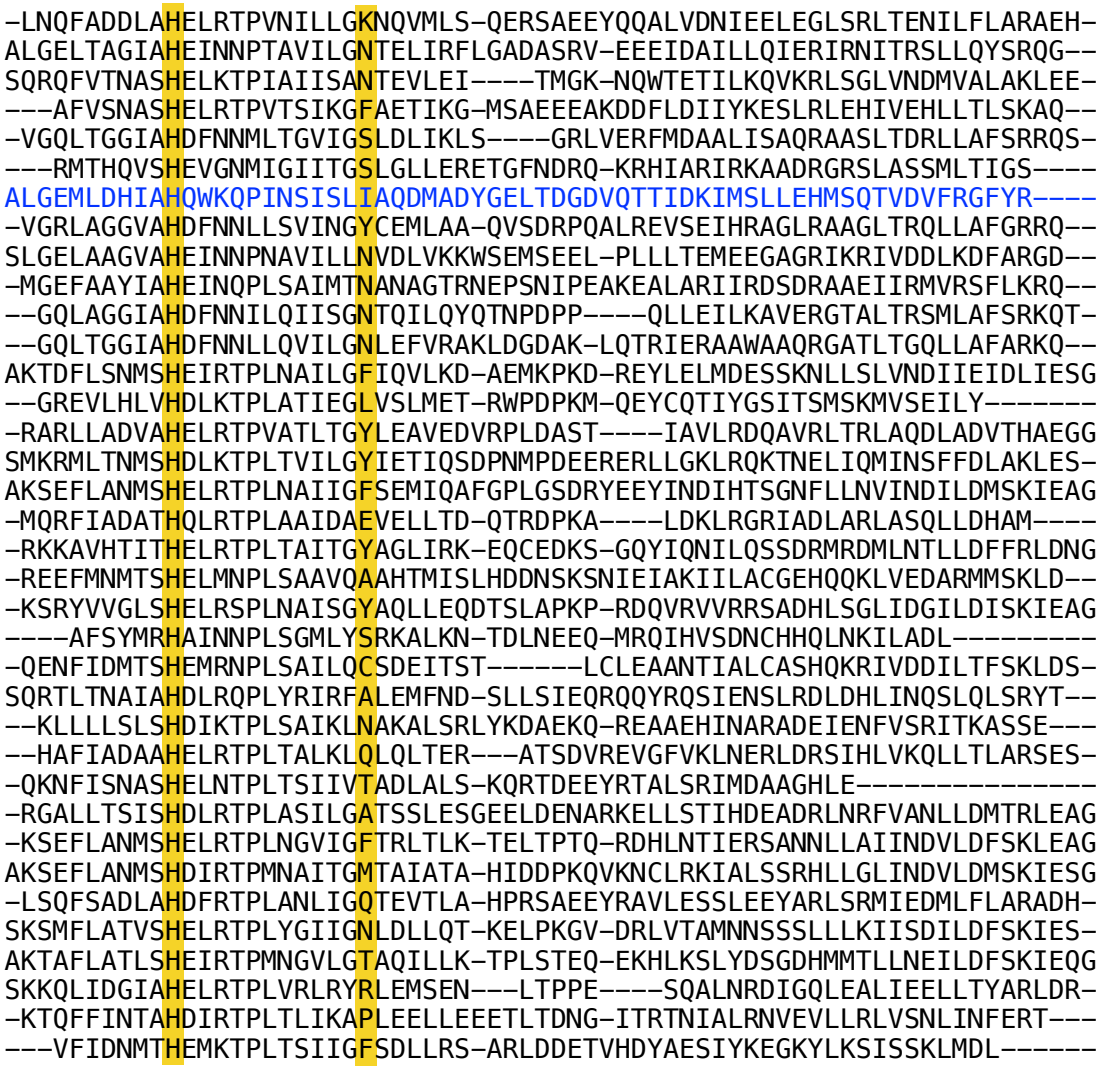


…and organized into **protein families**:

- common evolutionary origin (homologs)
- **conserved biological structure / function**
- **diverged sequences** (20-30% seq ID)
- available as multiple-sequence alignments in public **Pfam** database

# Looking for information in multiple-sequence alignments



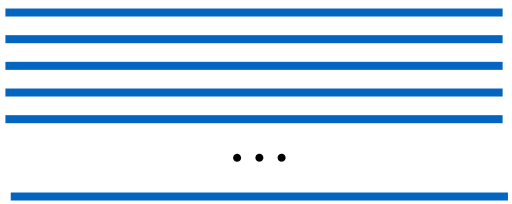

...

- 50-500 positions sequence length
- $10^3$-$10^5$ homologous sequences

# Statistical sequence models

**Data** – MSA of homologous sequences

$$\{(a_i^\mu, ..., a_L^\mu)\}_{\mu=1,...,M}$$



→

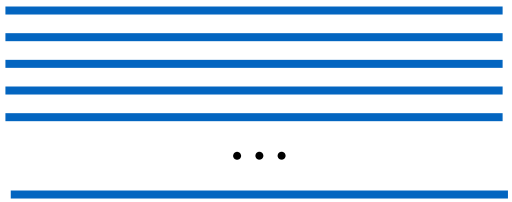
**Statistical sequence model**

$$P(a_1, ..., a_L)$$

# Statistical sequence models

**Data** – MSA of homologous sequences

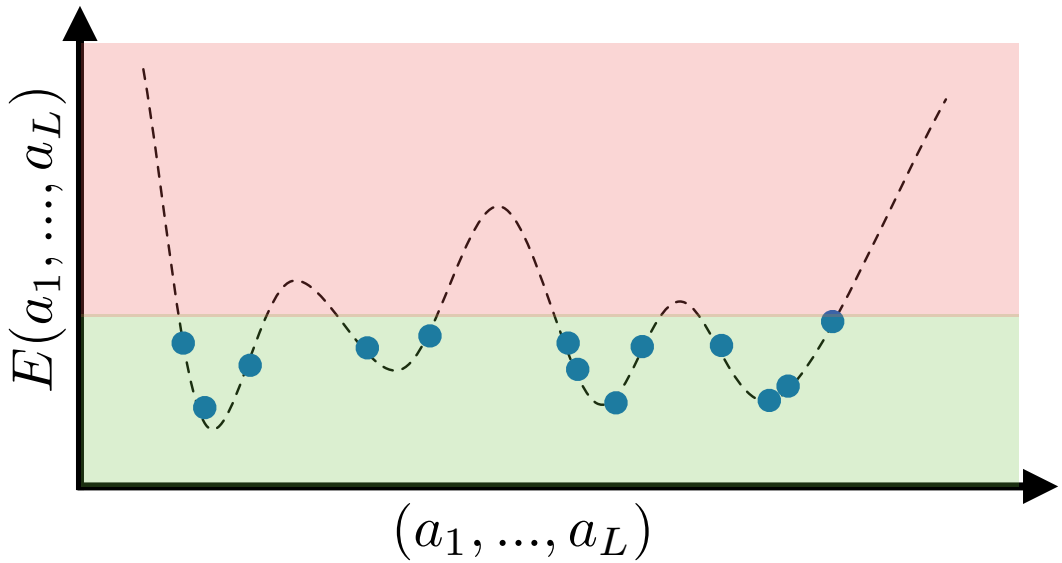
$$\{(a_i^\mu, \ldots, a_L^\mu)\}_{\mu=1,\ldots,M}$$



**Statistical sequence model**

$$P(a_1, \ldots, a_L) \sim \exp\{-E(a_1, \ldots, a_L)\}$$

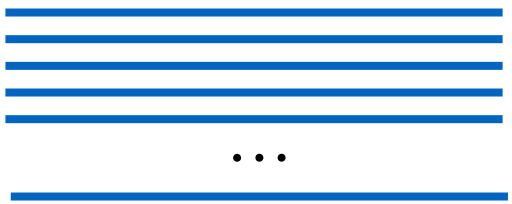
"statistical energy"
"sequence landscape"



# Statistical sequence models

**Data** – MSA of homologous sequences

$$\{(a_i^\mu, ..., a_L^\mu)\}_{\mu=1,...,M}$$



**Statistical sequence model**

$$P(a_1, ..., a_L)$$
$$\sim \exp\{-E(a_1, ..., a_L)\}$$

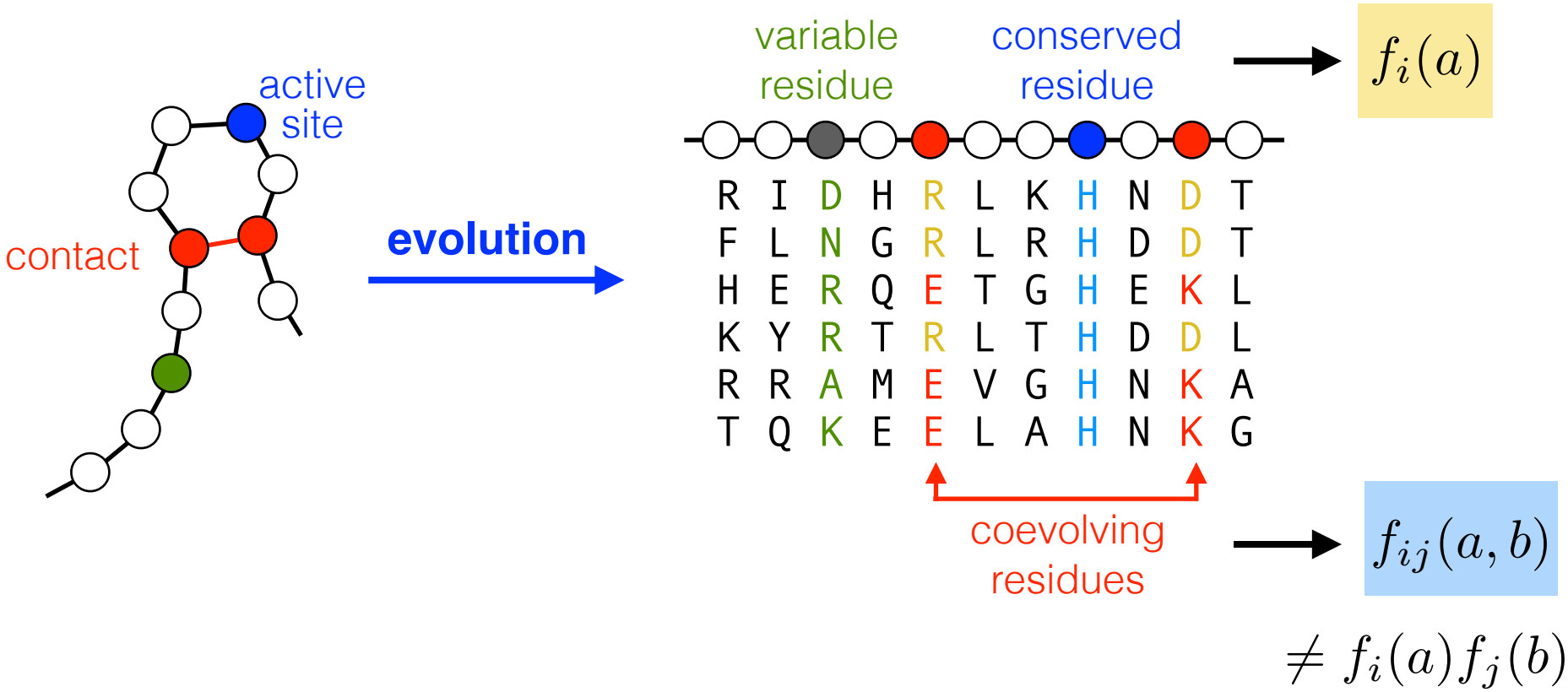
"statistical energy"
"sequence landscape"

**Attention:**
**Data insufficient to do this without modeling**

$$L \in (50, 500)$$
$$M \in (10^3, 10^5)$$

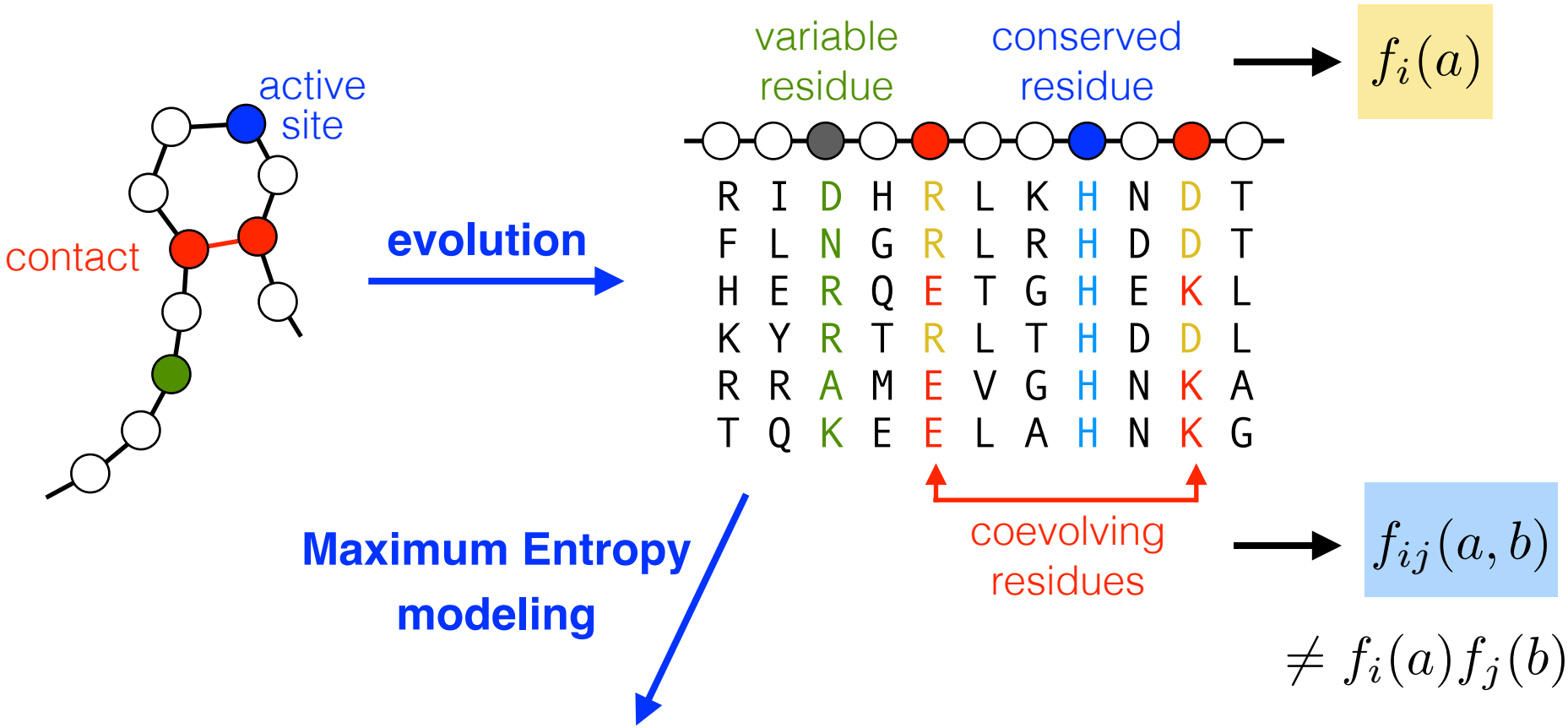
$20^L \in (10^{65}, 10^{650})$ parameters

# Conservation and coevolution in proteins




review: [Cocco, Feinauer, Figliuzzi, Monasson, MW '18]

# Conservation and coevolution in proteins




**Direct Coupling Analysis (DCA)** – Boltzmann machine / Potts model / Markov Random Field

$$P(a_1, ..., a_L) \sim \exp\left\{ \sum_{i<j} J_{ij}(a_i, a_j) + \sum_i h_i(a_i) \right\}$$

review: [Cocco, Feinauer, Figliuzzi, Monasson, MW '18]

# How to learn the parameters?

- **Boltzmann-machine learning:**
  - start with initialised parameters (fields/couplings)
  - calculate marginals

$$P_{ij}(a,b) = \sum_{a_1,\ldots,a_L} P(a_1,\ldots,a_L)\delta_{a_i,a}\delta_{a_j,b}$$

  - update parameters to fit marginals (gradient ascent of log-likelihood)

$$\Delta J_{ij}(a,b) = \varepsilon[f_{ij}(a,b) - P_{ij}(a,b)]$$

  - iterate until sufficiently precise fitting

➡ exact calculation requires exponential time ~$21^L$

➡ approximations (MCMC, PCD) needed for computational efficiency

# All models are wrong, but some are useful

[George E.P. Box, 1976]

1. Are pairwise couplings **useful** ?

2. Are pairwise couplings **sufficient** ?

3. Are pairwise couplings **necessary** ?

# All models are wrong, but some are useful

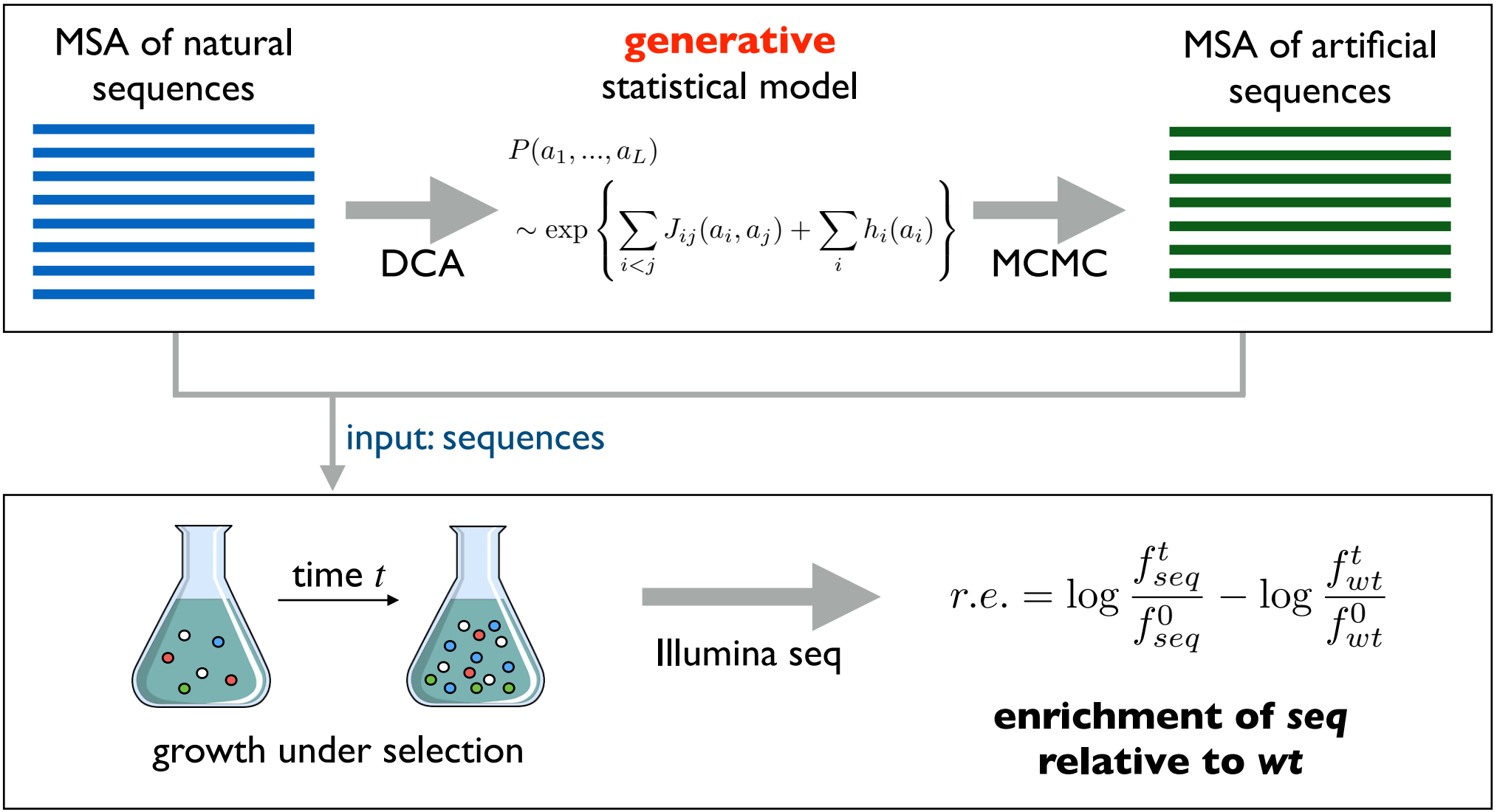
[George E.P. Box, 1976]

1. Are pairwise couplings **useful** ? ▸ **predictive, interpretable**

2. Are pairwise couplings **sufficient** ? ▸ **generative**

3. Are pairwise couplings **necessary** ? ▸ **minimal, parsimonious**

# From data to sequence design

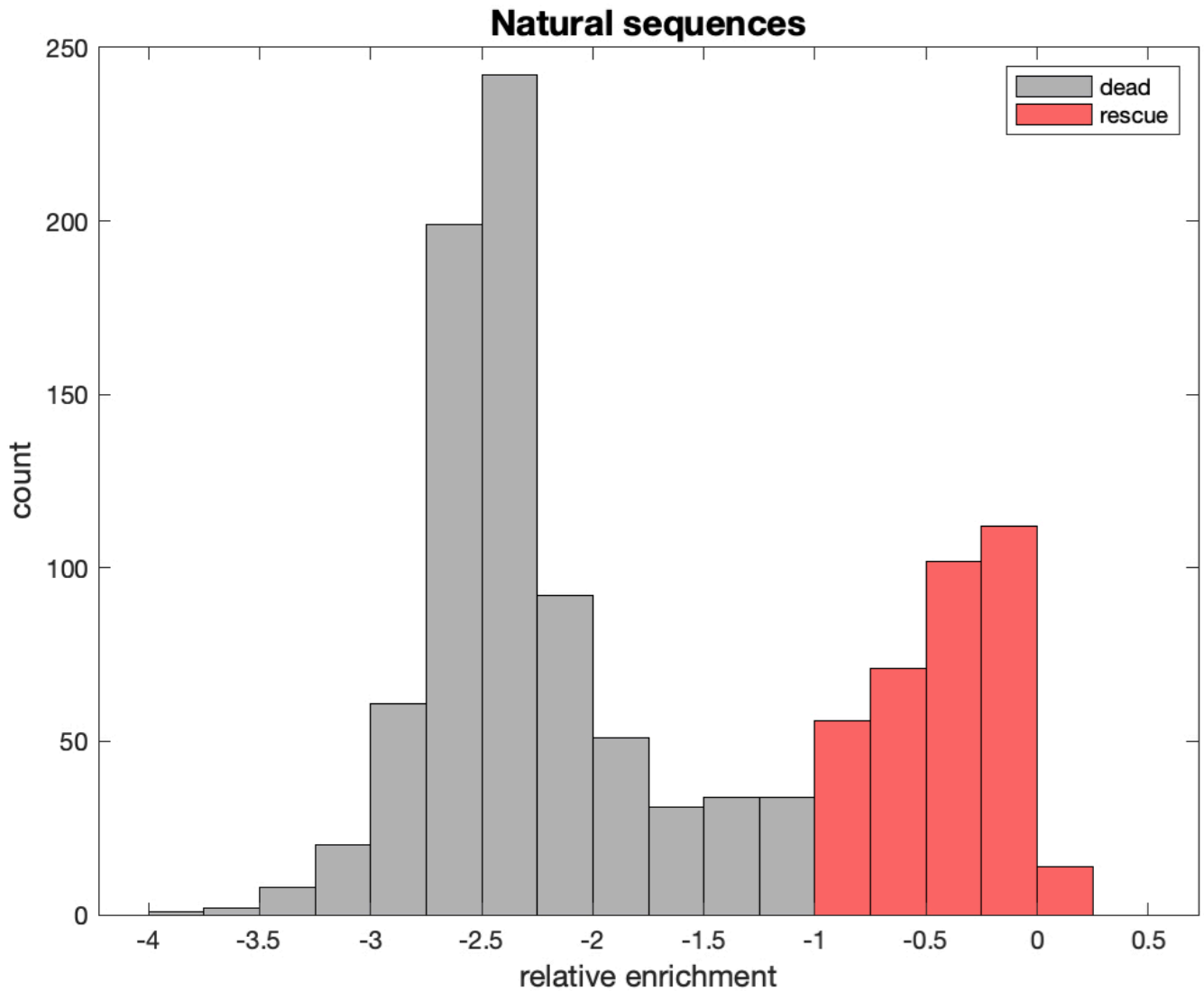



***in vivo* high-throughput assay**

Russ, Figliuzzi, Stocker, Barrat-Charlaix, Socolich, Kast, Hilvert, Monasson, Cocco, Weigt and Ranganathan, *Science* (2020)

# Chorismate mutase design

- enzyme in the synthesis pathway of phenylalanine and tyrosine
  - conditionally essential gene in *E. coli*
- natural alignment of 1259 sequences of length $L = 96$



# Sequence design : what to expect?

## Sequence spaces

- sequences of 20 amino acids, length 96 : $\simeq 10^{125}$
- sequences coherent with conservation : $\simeq 10^{86}$
- sequences coherent with coevolution : $\simeq 10^{54}$

# Sequence design : what to expect?

## Sequence spaces

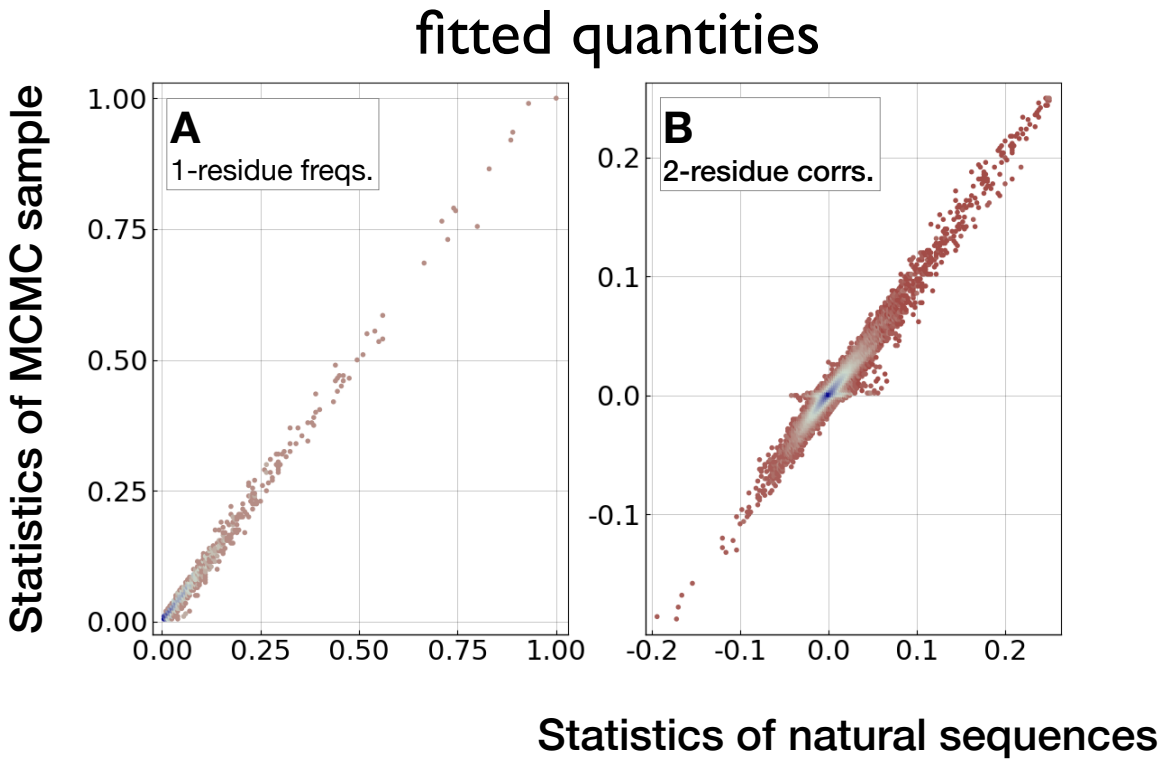
- sequences of 20 amino acids, length 96 : $\simeq 10^{125}$
- sequences coherent with conservation : $\simeq 10^{86}$
- sequences coherent with coevolution : $\simeq 10^{54}$

## Atoms

- atoms in the universe : $\simeq 10^{78} - 10^{82}$
- atoms on earth : $\simeq 10^{50}$
- atoms in human cell : $\simeq 10^{14}$

➡ very good generative models are needed

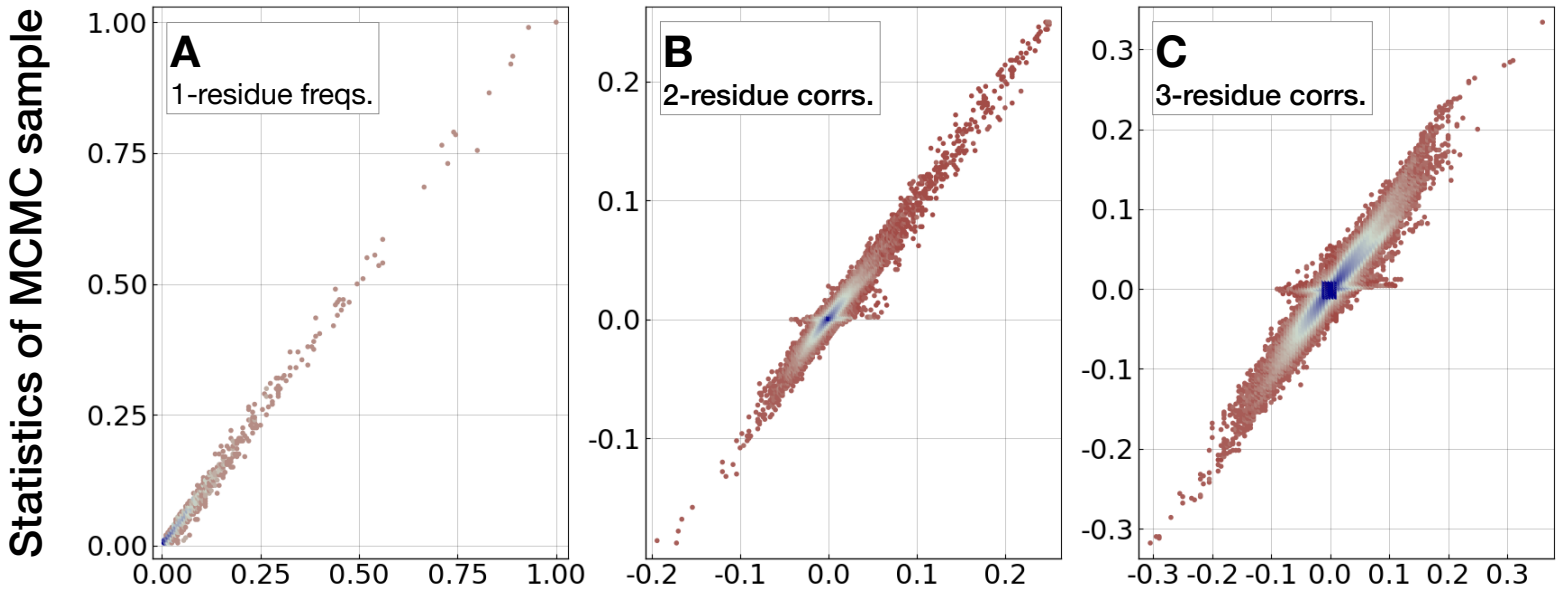
# Is the DCA model generative ?



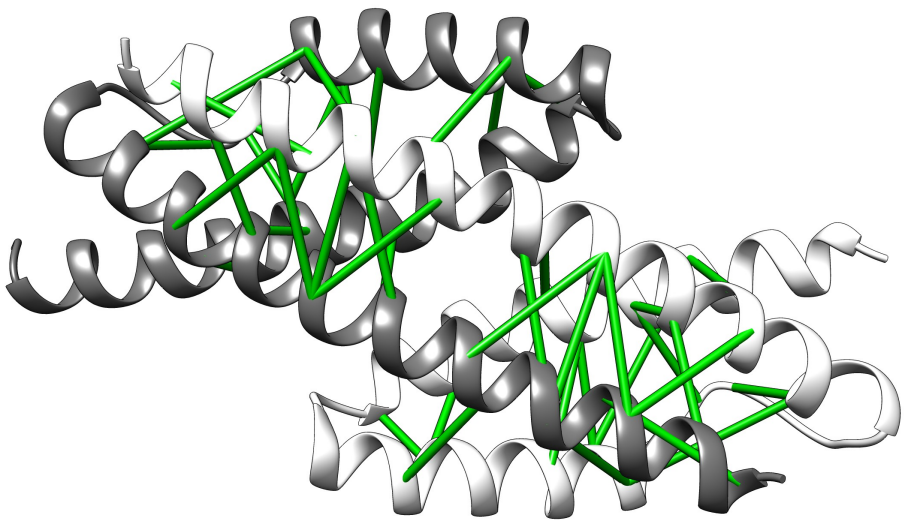
# Is the DCA model generative ?

fitted quantities

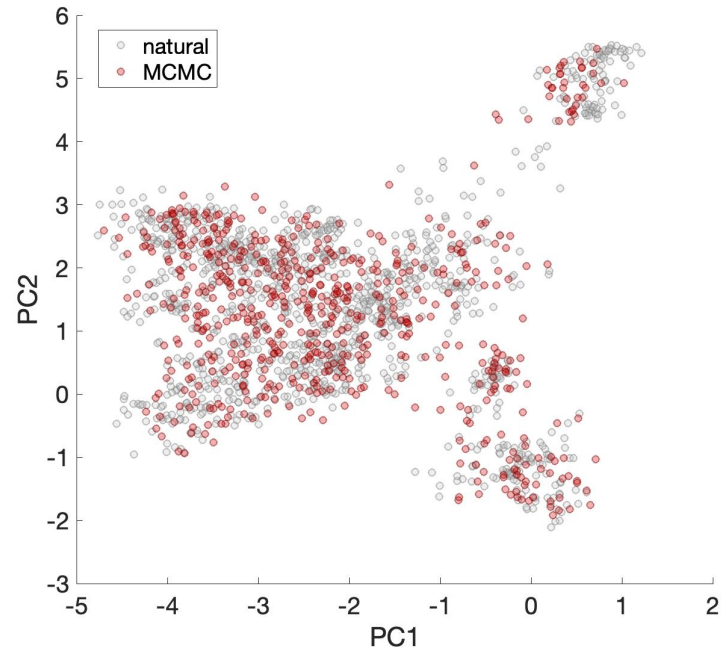
non-fitted quantities




Statistics of natural sequences

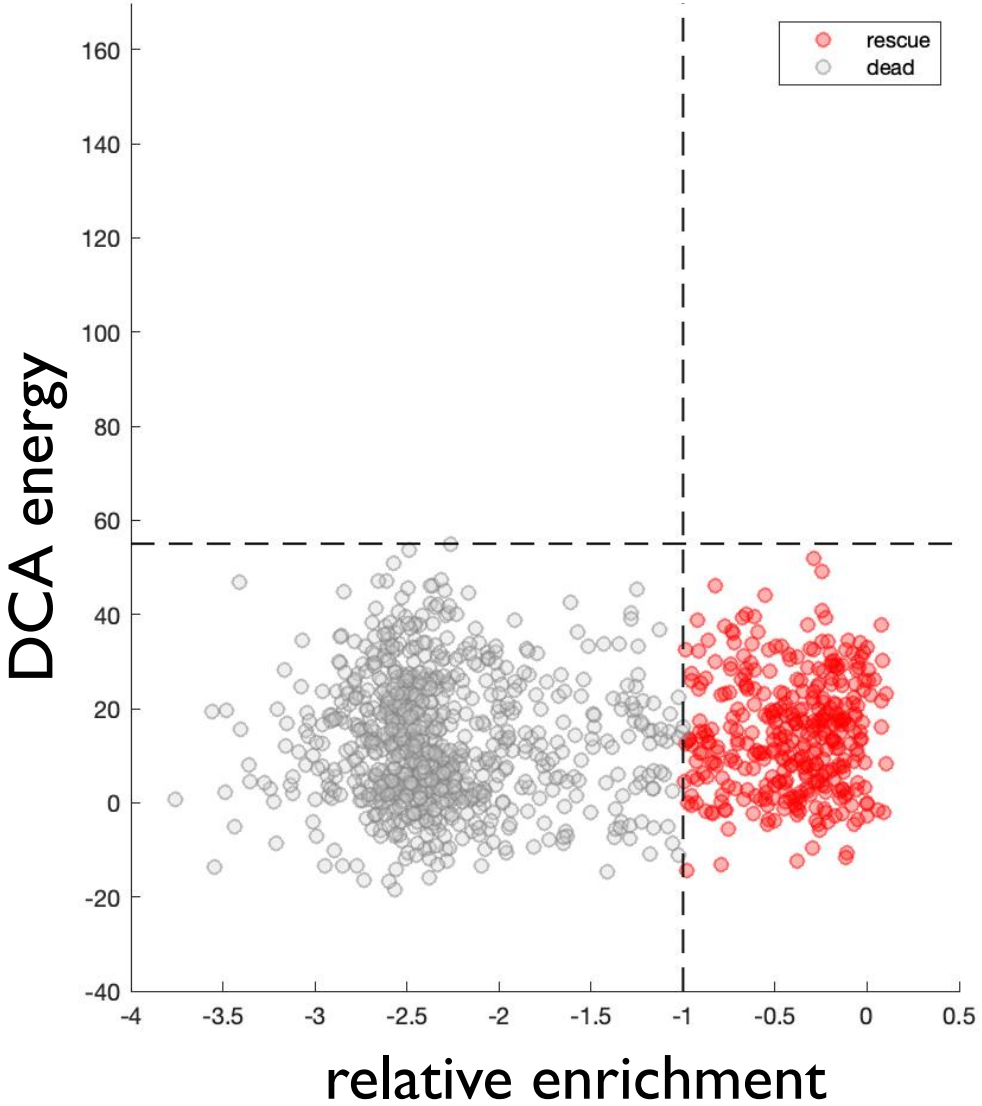


**Strongest couplings = 3D contacts**

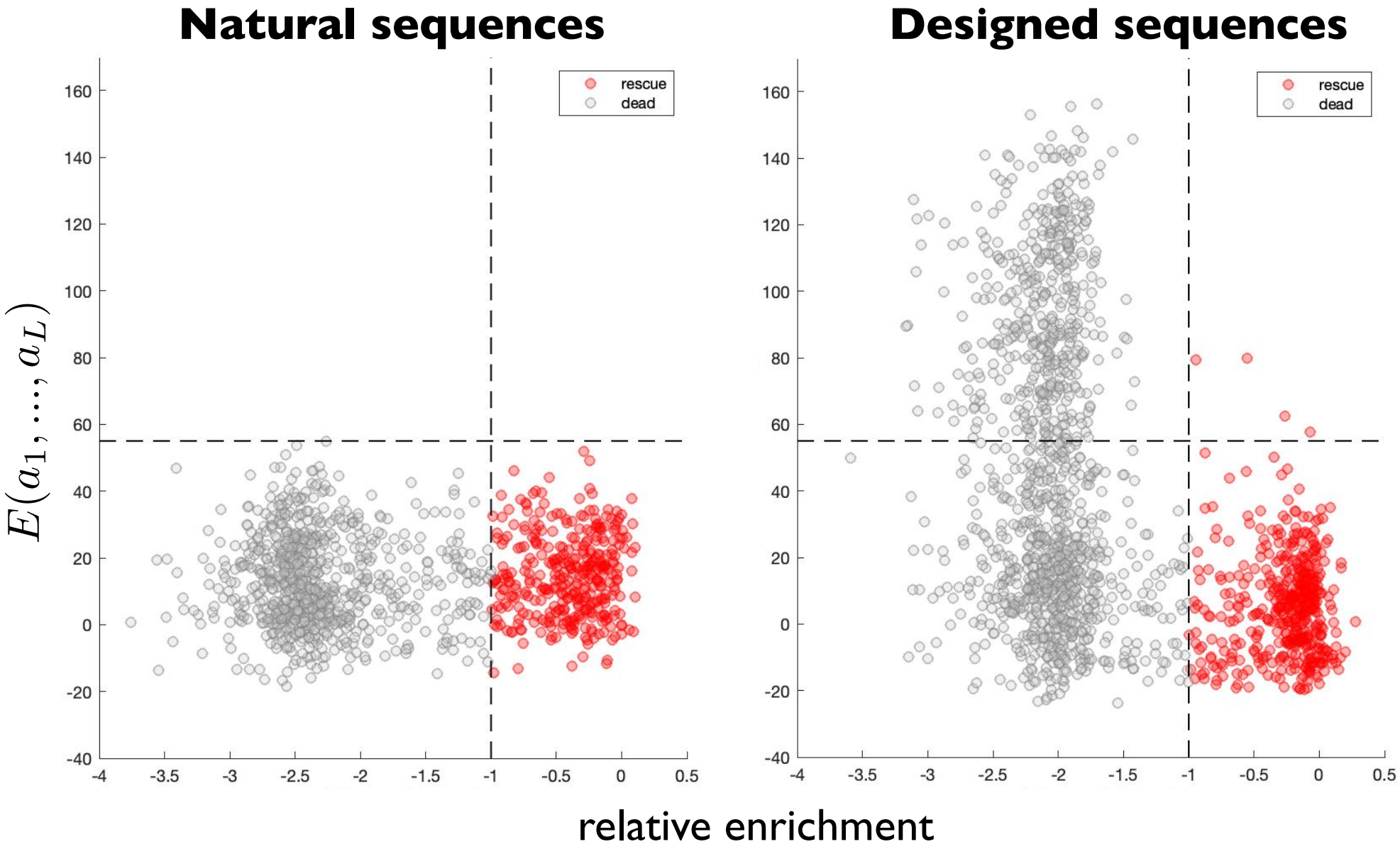


# Chorismate mutase design

## Natural sequences

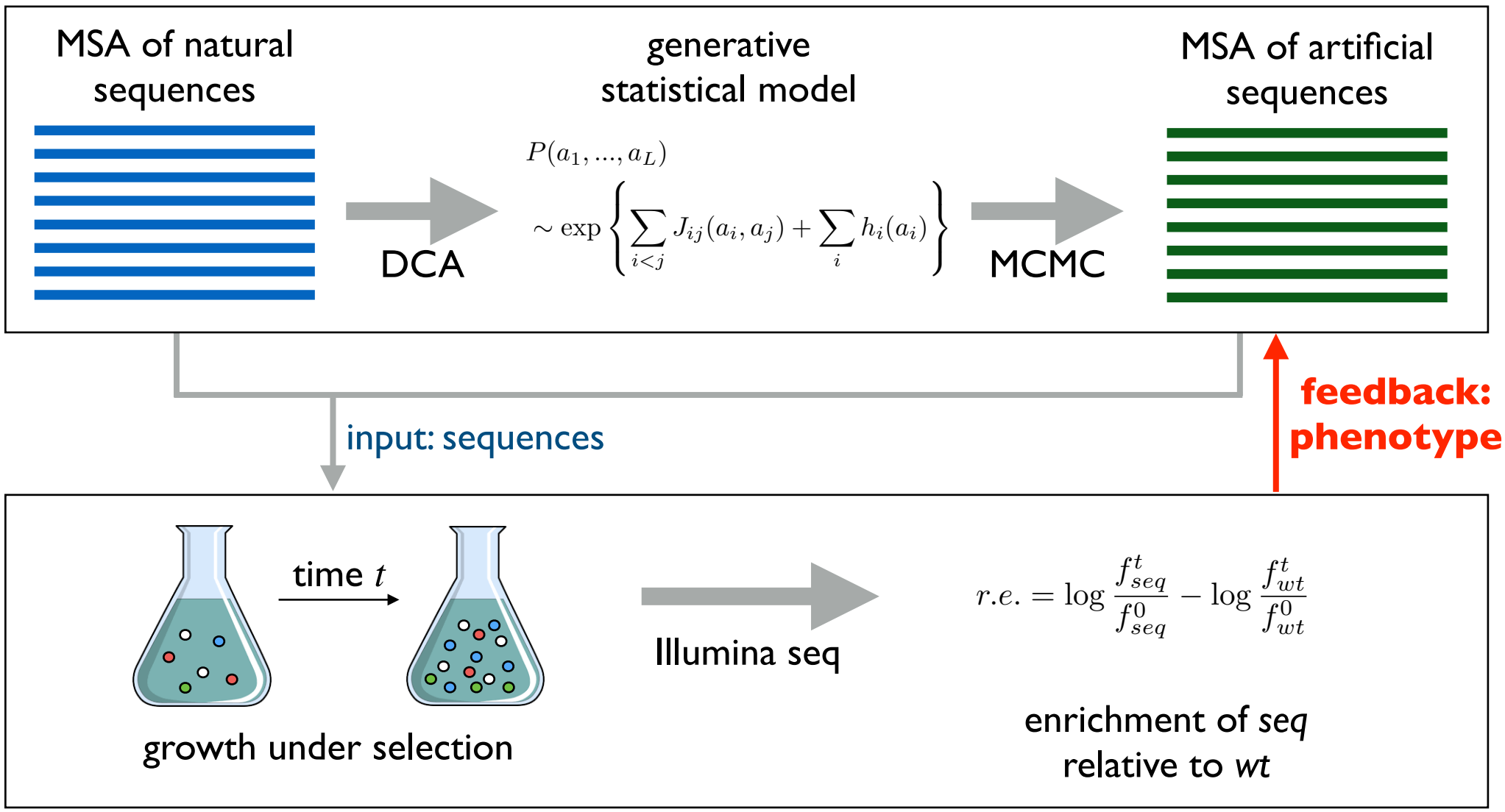


# Chorismate mutase design




- **functional & variable sequences** found at low DCA energies
- profile models (conservation, no coevolution) - no functional seqs

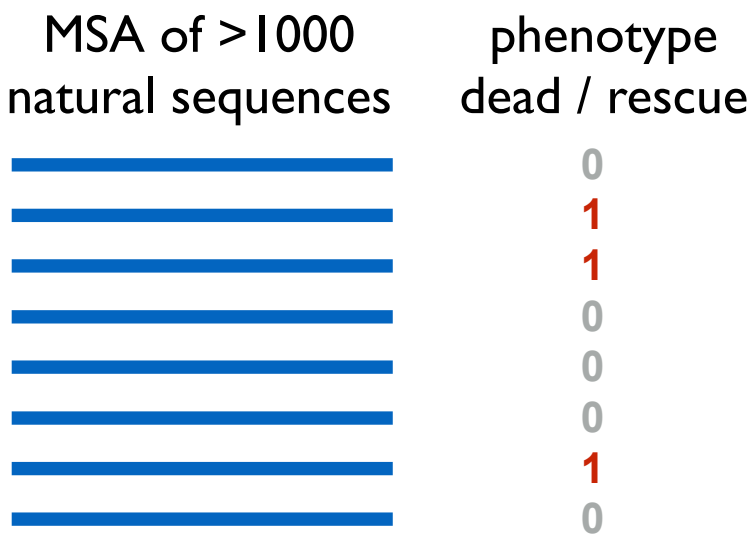
# Experimental feedback for better design




***in vivo* high-throughput assay**

# Integrating phenotypic information

augmented data:




**supervised** problem:

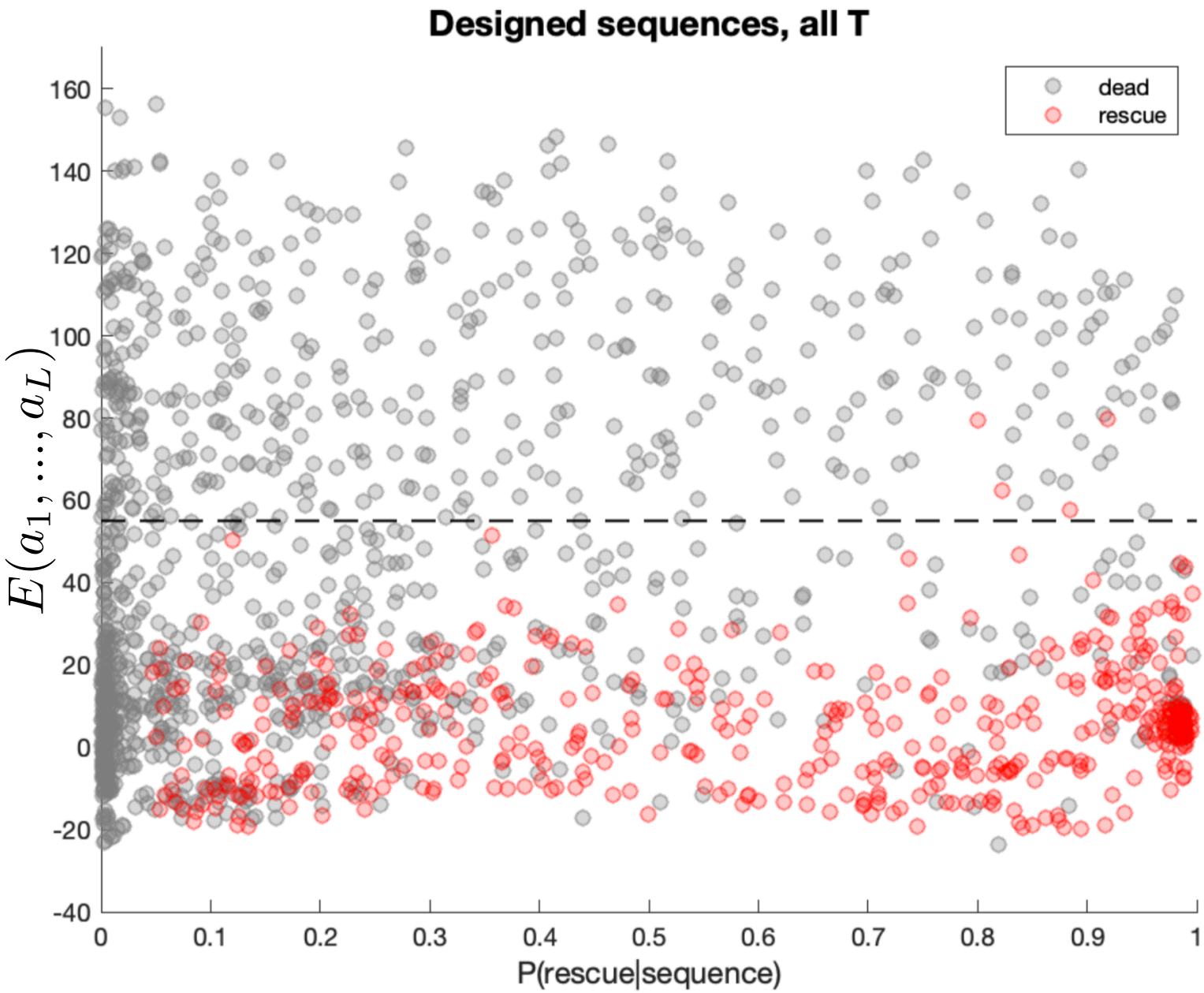
- predict phenotype $x$ from sequence $(a_1, ..., a_L)$

$$P(x = 1|a_1, ..., a_L) = \frac{\exp\{\sum_i K_i(a_i)\}}{1 + \exp\{\sum_i K_i(a_i)\}}$$

- infer parameters from natural sequences & phenotypes

# Integrating phenotypic information

- validate on designed sequences

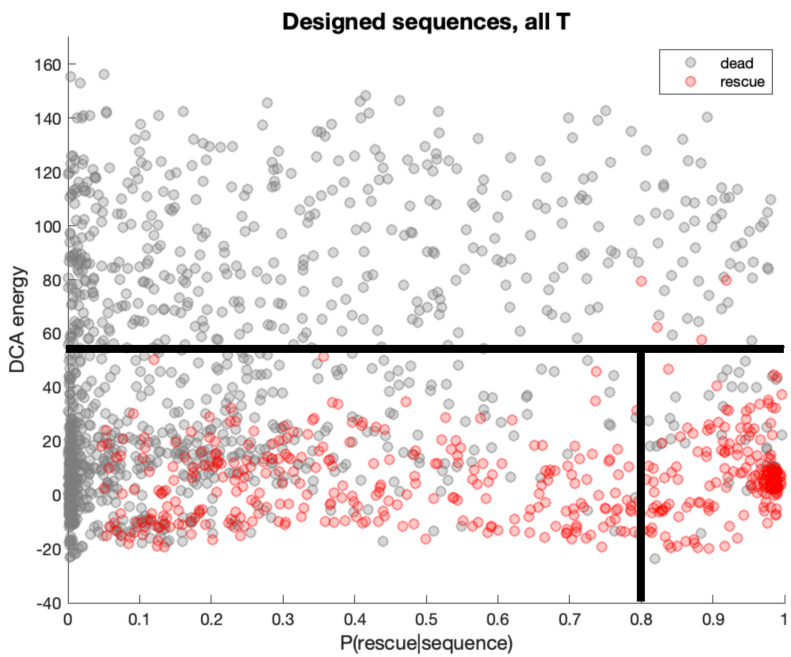


# Should we be astonished ?

## Sequence spaces:

- sequences of 20 amino acids, length 96 : $\simeq 10^{125}$
- sequences coherent with conservation : $\simeq 10^{86}$
- sequences coherent with coevolution : $\simeq 10^{54}$

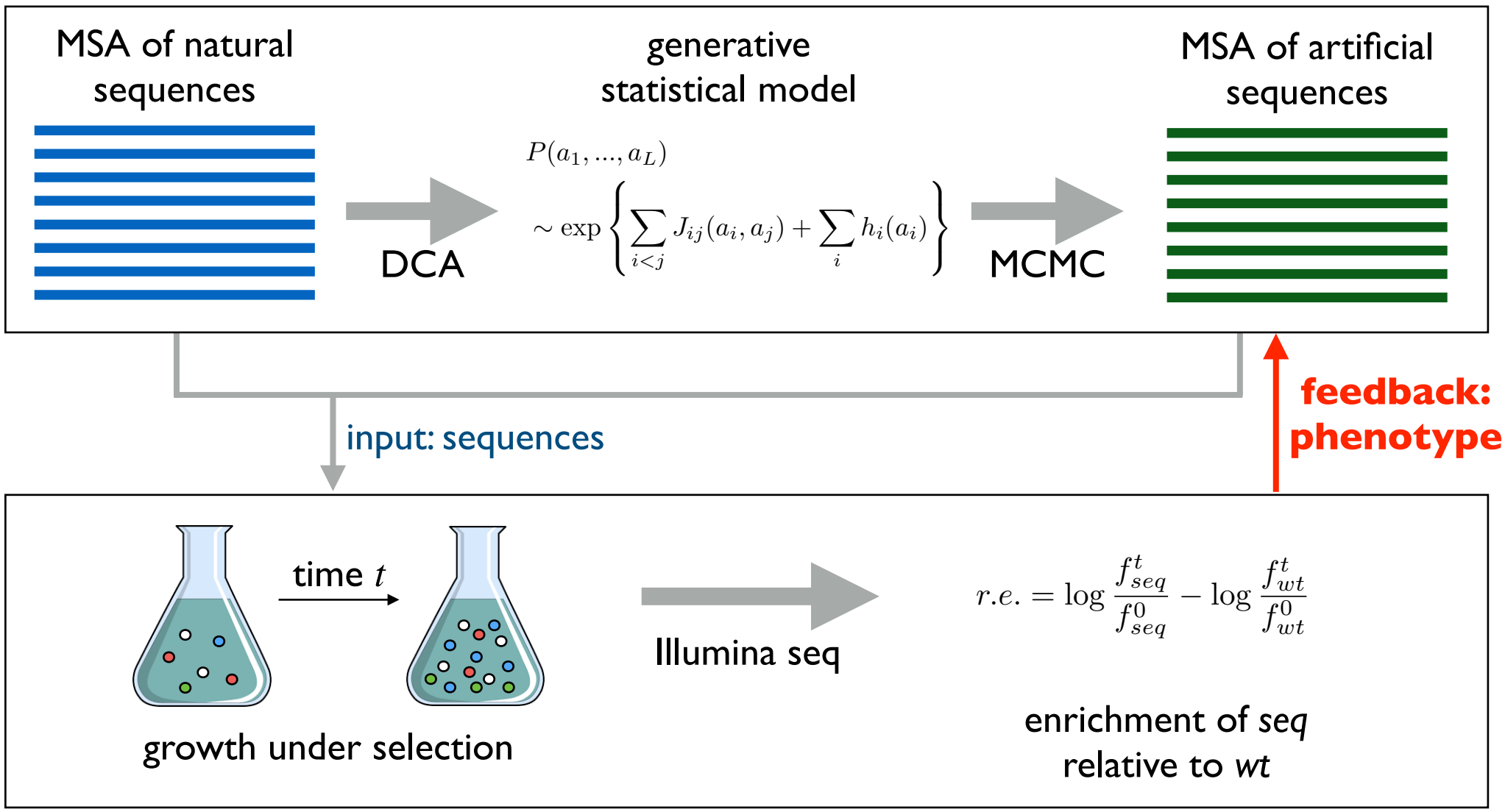
## Success probabilities:




- all designed sequences : 29%
- low DCA energy : 39%
- low DCA energy and high P(rescue) : 79%

Russ, Figliuzzi, Stocker, Barrat-Charlaix, Socolich, Kast, Hilvert, Monasson, Cocco, Weigt and Ranganathan, *Science* (2020 in press)

# Experimental feedback for better design




***in vivo* high-throughput assay**