# Statistical models of protein sequences

## Generative models & evolution-guided protein design

**Pierre Barrat-Charlaix**

Biozentrum, Univeristy of Basel

# Statistical modeling of protein sequences

## Protein family

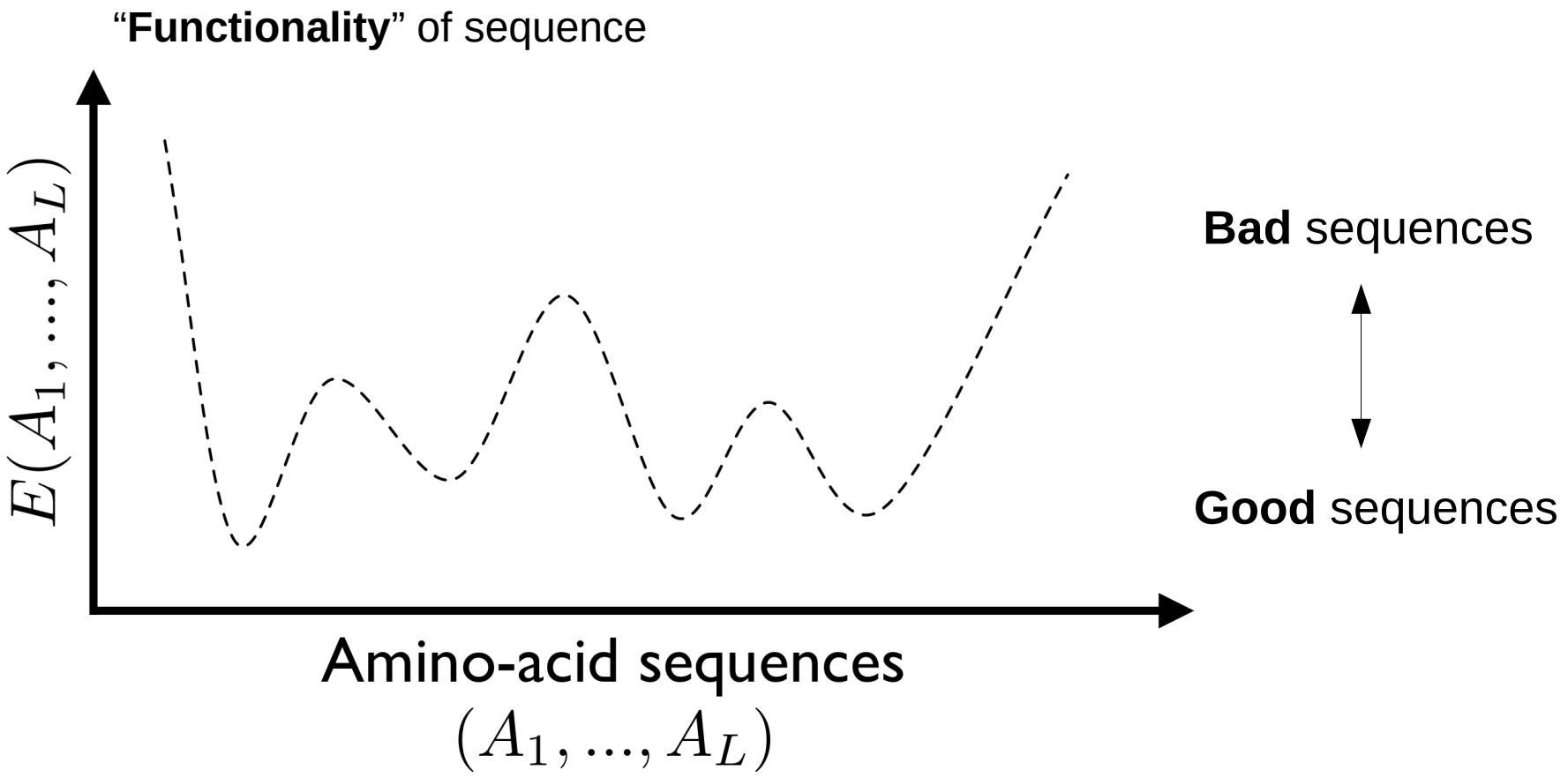
Evolutionary constraints

## Multiple Sequence Alignment

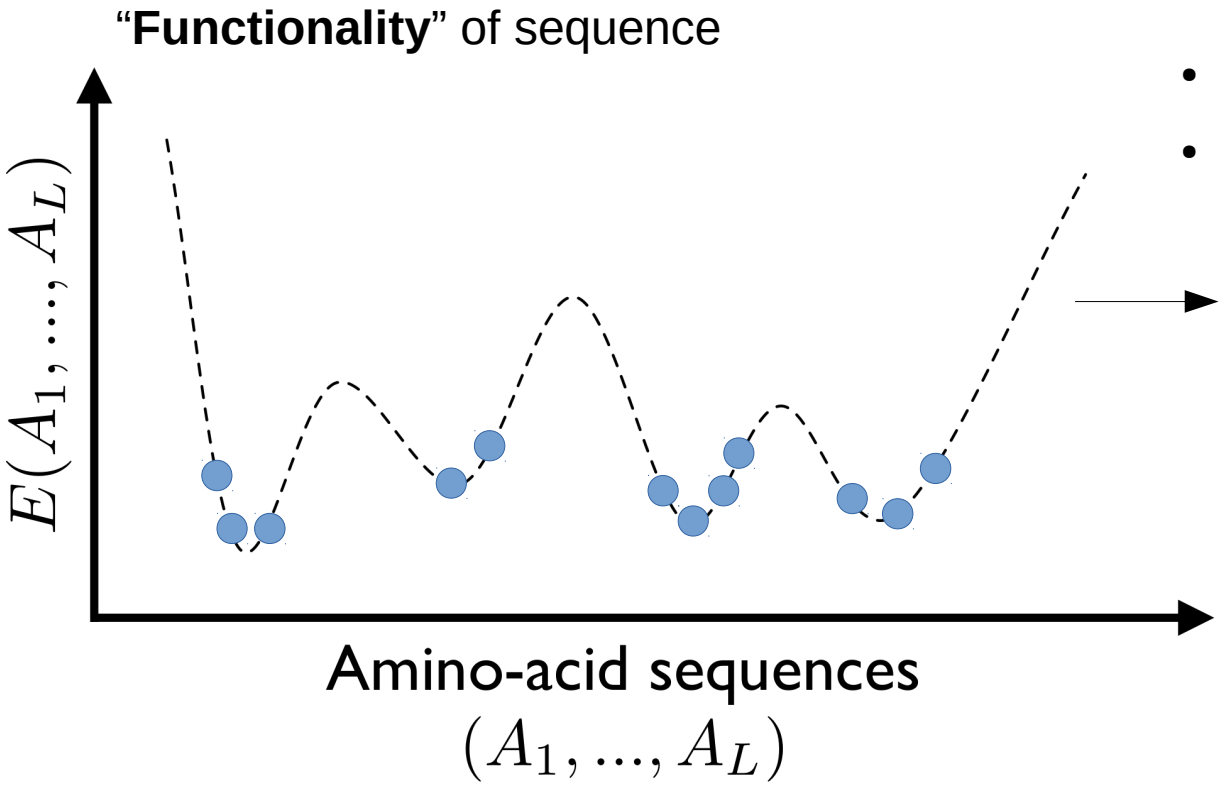
```
...
YHCDKCSMSFAAPSRLNKHMRTH
HKCSYCSKAFIKKTLLKAHERTH
-QCEECGKQFAYSHSLKTHMMTH
YVCNVCGNLFRQHSTLTIHMRTH
-TCEFCGKNFERNGNYVEHRRTH
FVCGVCNKGFNSRTYLLEHMNKH
YVCHFCGKAVTNRESLKTHVRLH
YSCNVCDKSFTQRSSLVVHQRTH
FECQICGKSFKRSVQLKYHMEIH
YKCATCQKSFKRSQELKSHGKLH
HACGICGKTFPNNSSLEKHKHIH
YVCDKCGRSFSQRSSLTIHQRYH
YTCNVCGKTVTTKKSYTNHVKIH
FKCGVCGKFYKNESSLKTHSKIH
-QCEECGEIFNHKSSLNKHLLKH
YACEYCDKRFGDKQYLTQHRRVH
FKCDECGQCFSQRSSLNRHKRYH
YECDICGICFNQRSTMTSHRRSH
```

# Information?

# Sequence functionality landscape



# Sequence functionality landscape



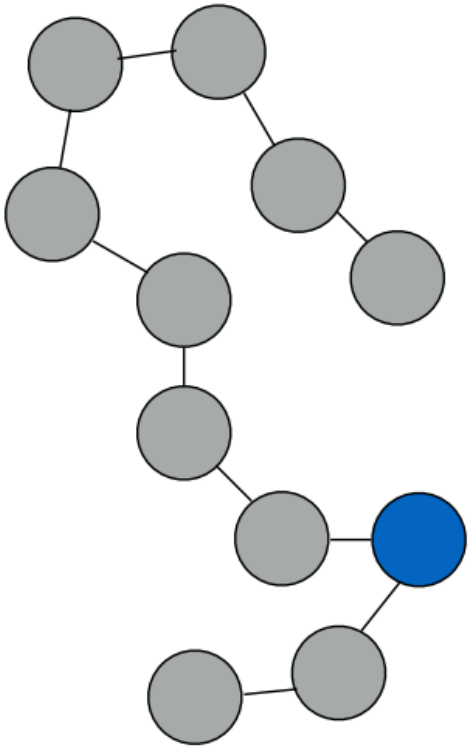

**Families of homologous sequences**

- Conserved structure and function
- Low sequence ID (20-30%)

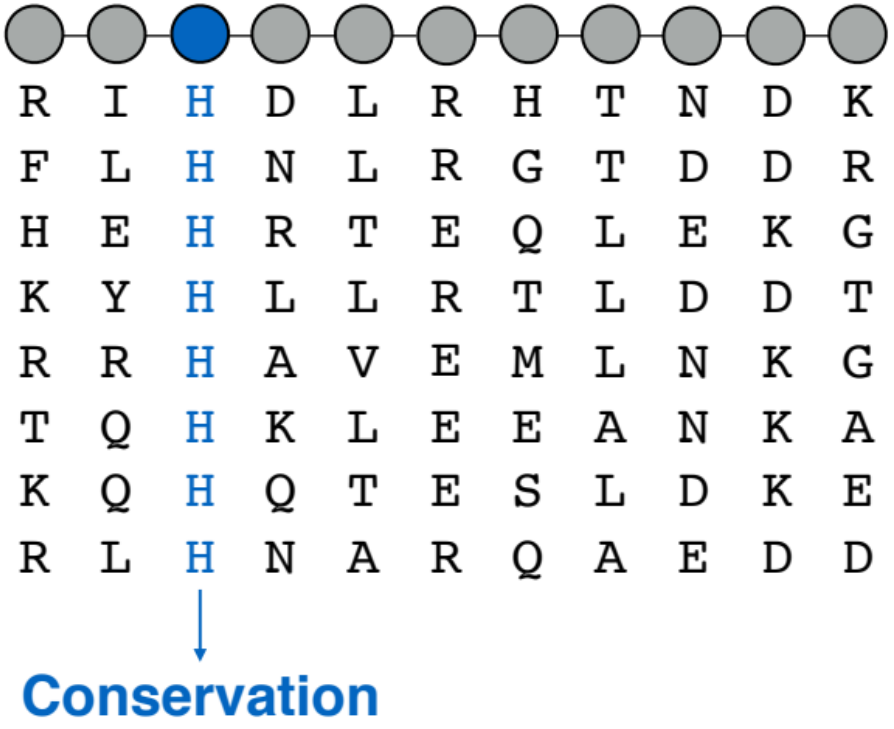
→ **Global sampling of sequence landscape**

**How can we model this ?**

# Profile models



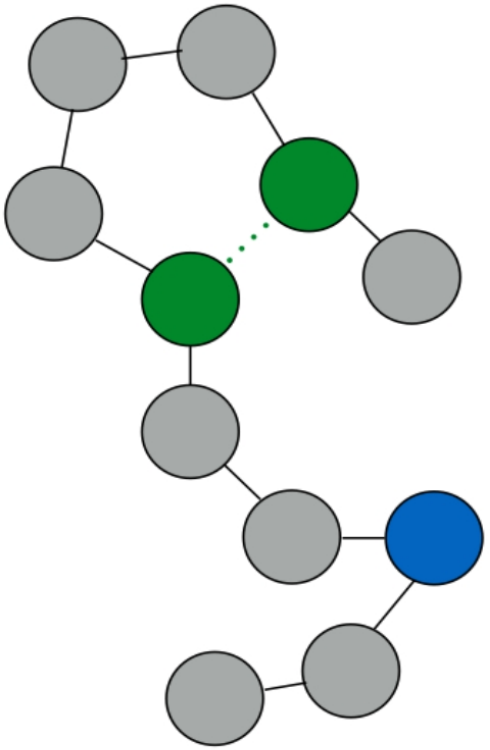
Evolutionary constraints



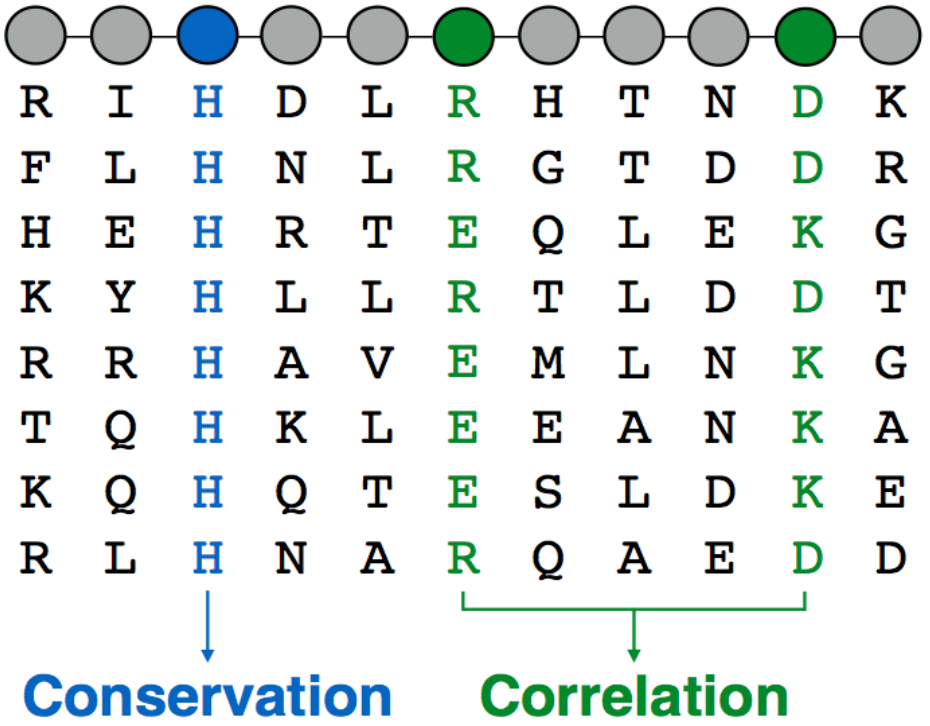

---

- Functionally important **positions**
- Homology detection (HMM)
- **Unable to capture relations between columns**

# Global statistical model



Evolutionary constraints



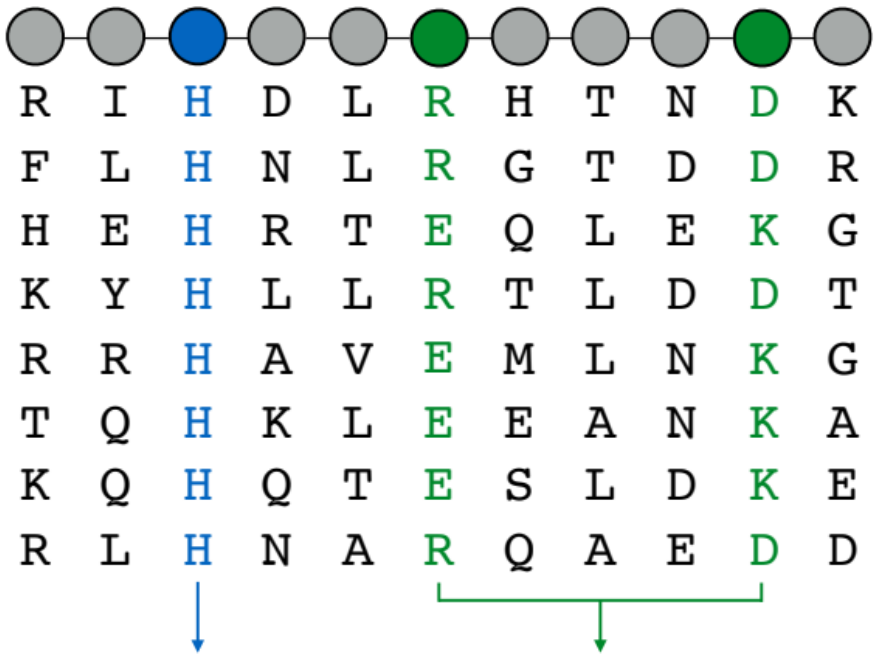

Couplings Fields

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

**Direct Coupling Analysis (DCA)**

# Maximum entropy formalism

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




$f_i(a)$ $f_{ij}(a, b)$

## Maximum entropy modeling

Find distribution $P(a_1 \ldots a_N)$

- With **maximal entropy** …

$$-\sum_{\{\vec{a}\}} P(\vec{a}) \log P(\vec{a}) \longrightarrow \text{Max}$$

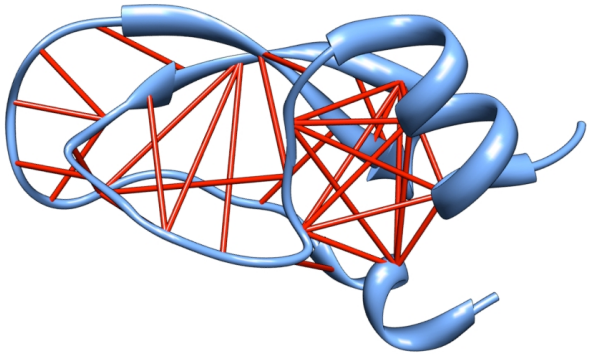
- While reproducing **pairwise statistics of data**

$$P_i(a) = f_i(a)$$
$$P_{ij}(a, b) = f_{ij}(a, b)$$

→ **Only information used is $f_{ij}(a, b)$ and $f_i(a)$**

# DCA: Successful model

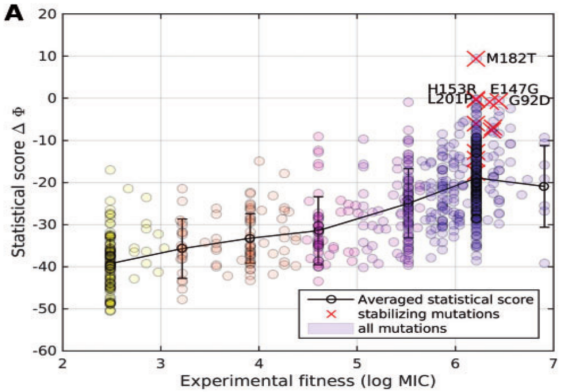
**Review**: Cocco *et al.,* 2018



- **Predicting 3D structure**

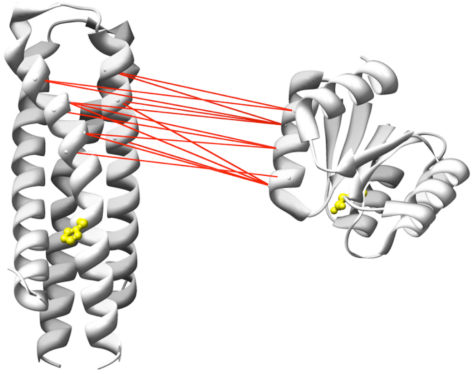
Morcos *et al.,* PNAS, 2011

Ovchinnikov *et al.,* Science, 2017




- **Predicting effect of mutations**

Figliuzzi *et al.,* MBE, 2015

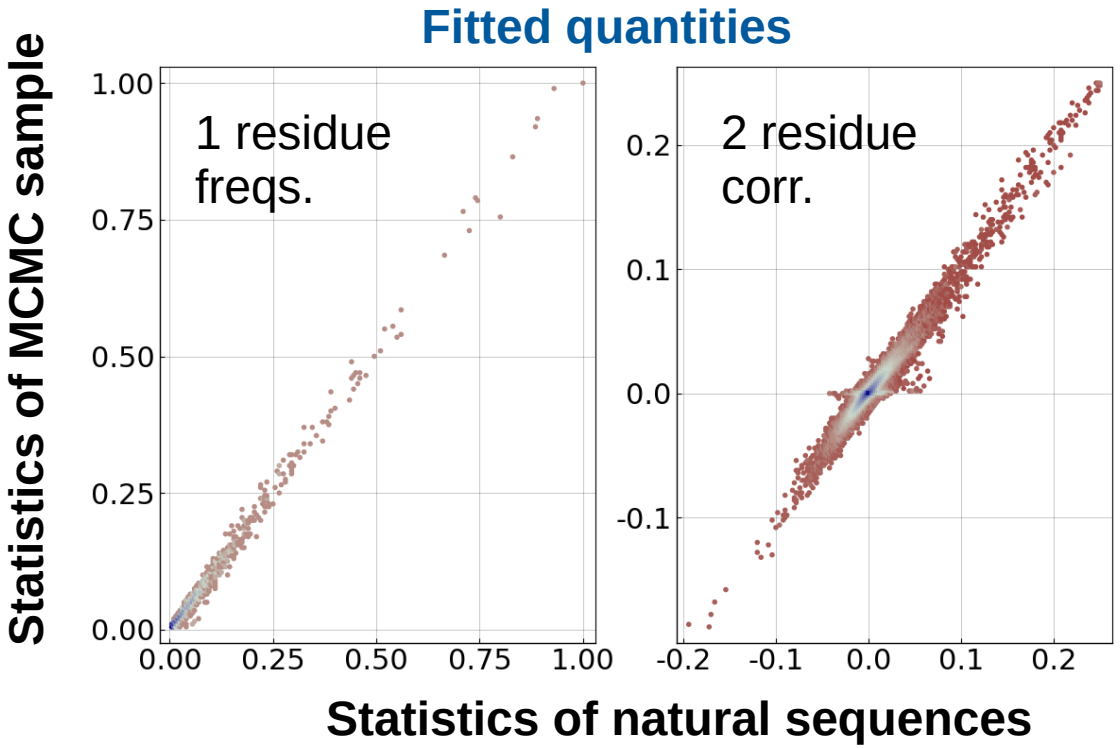


- **Predicting protein-protein interactions**

Gueudré *et al.,* PNAS, 2016

How good are DCA models at describing **functionality of a protein** ?

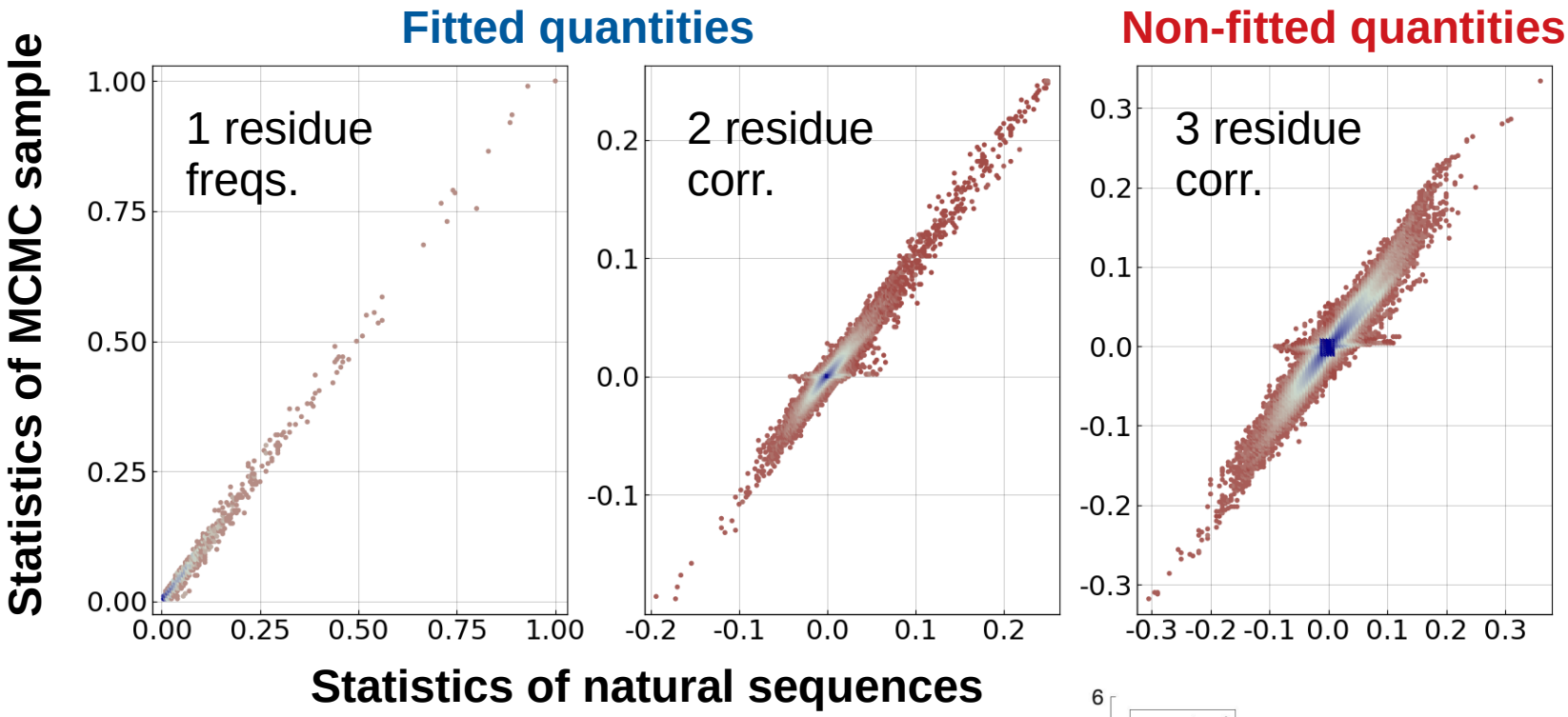
# Is the DCA model generative?



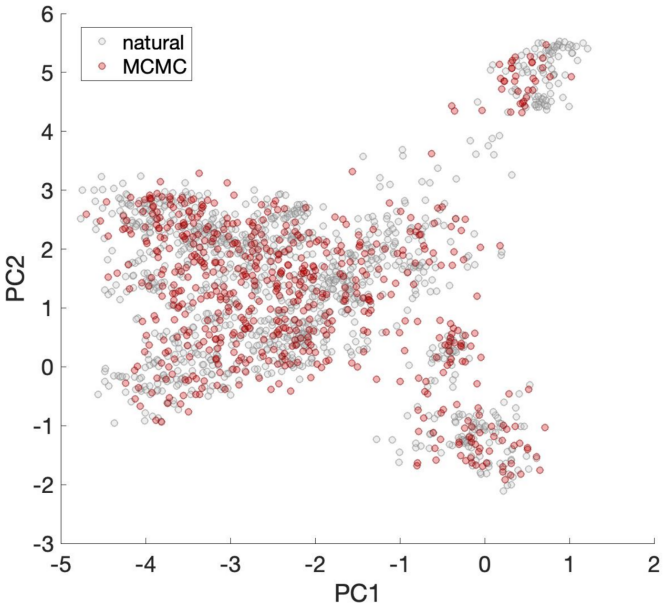

Sample from the **DCA sequence landscape**

$$P(a_1 \dots a_l) \propto \exp\{-E(a_1 \dots a_l)\}$$

# Is the DCA model generative?



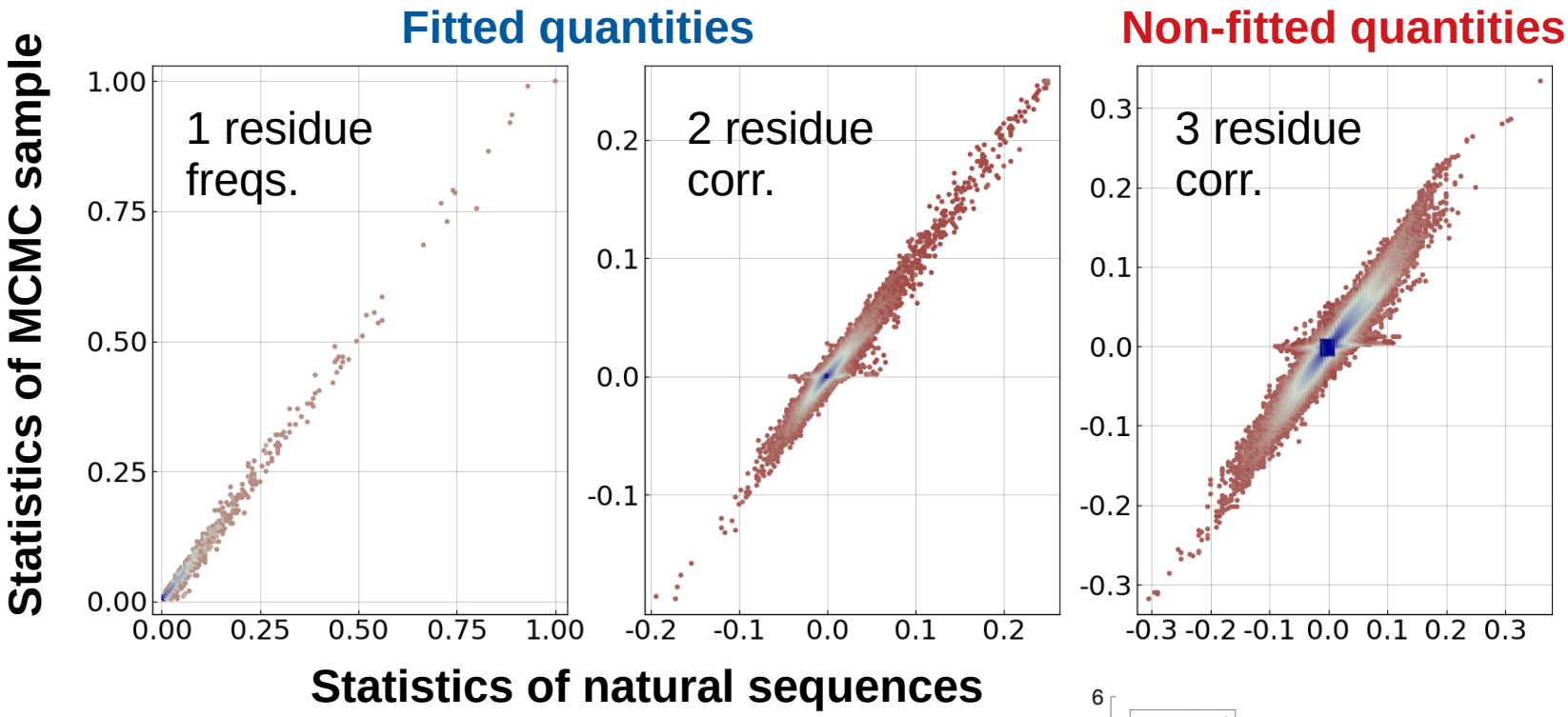

Projection onto the Principal Components of the MSA



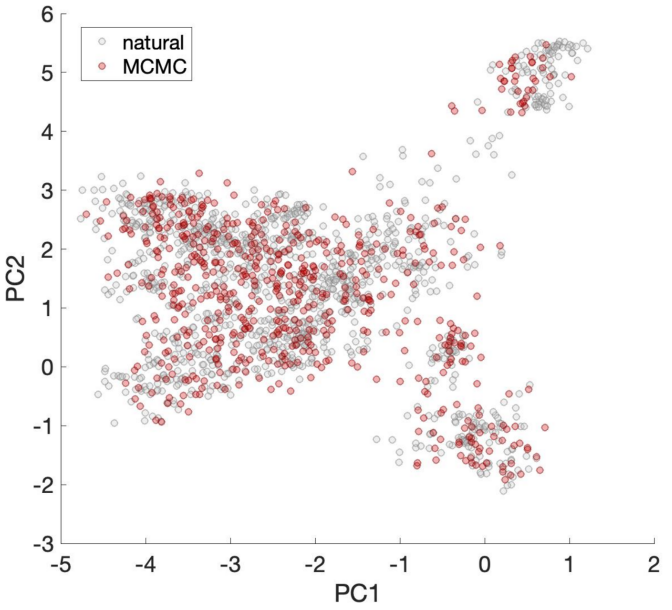

Sample from the **DCA sequence landscape**

$$P(a_1 \dots a_l) \propto \exp\{-E(a_1 \dots a_l)\}$$

# Is the DCA model generative?




Projection onto the Principal Components of the MSA




**DCA and natural sequences are statistically indistinguishable**

Figliuzzi *et al.,* MBE, 2018

→ **Are DCA sequences functional ?**

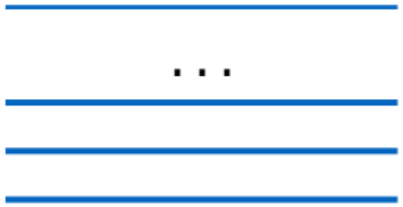
# Protein design

## Chorismate mutase

enzyme in the synthesis pathway of phenylalanine and tyrosine

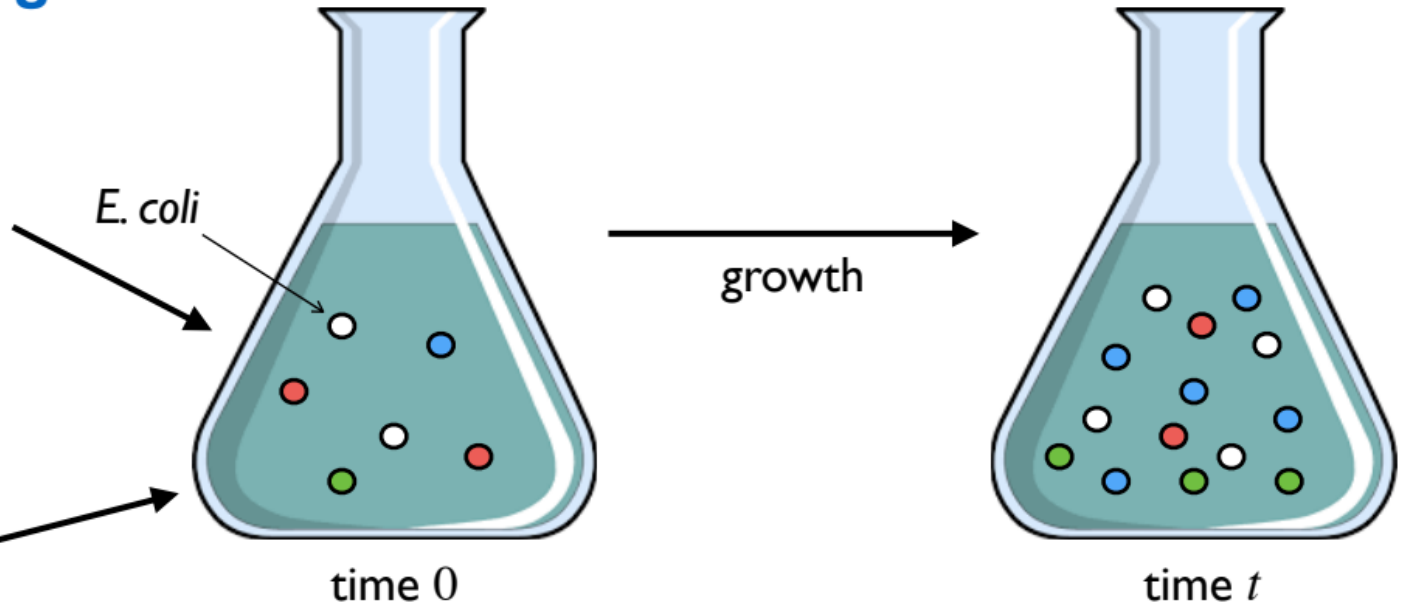
with Rama Ranganathan's group

1130 **natural homologs**



**DCA sequences**

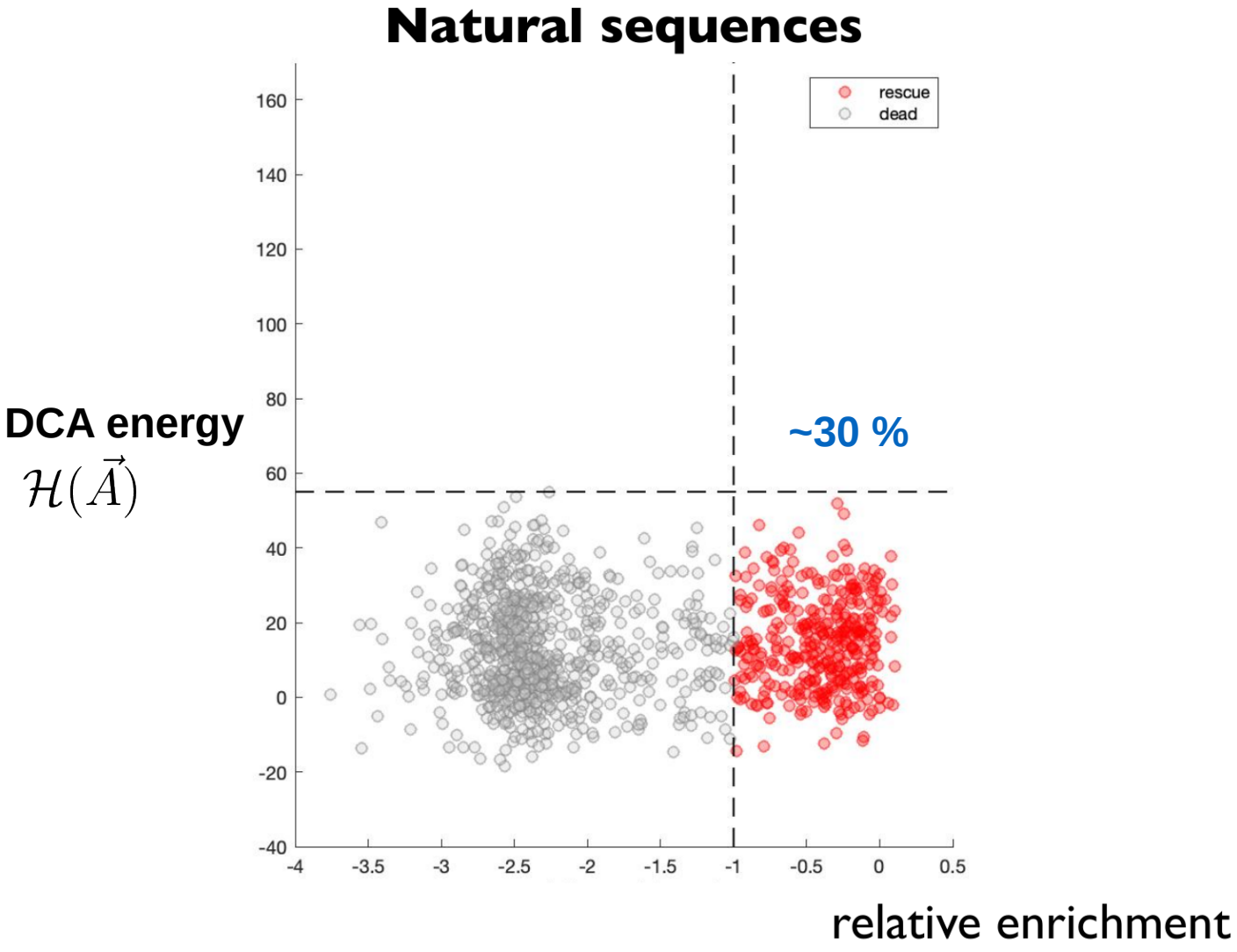
$$P(\vec{A}) \propto e^{-\mathcal{H}(\vec{A})}$$



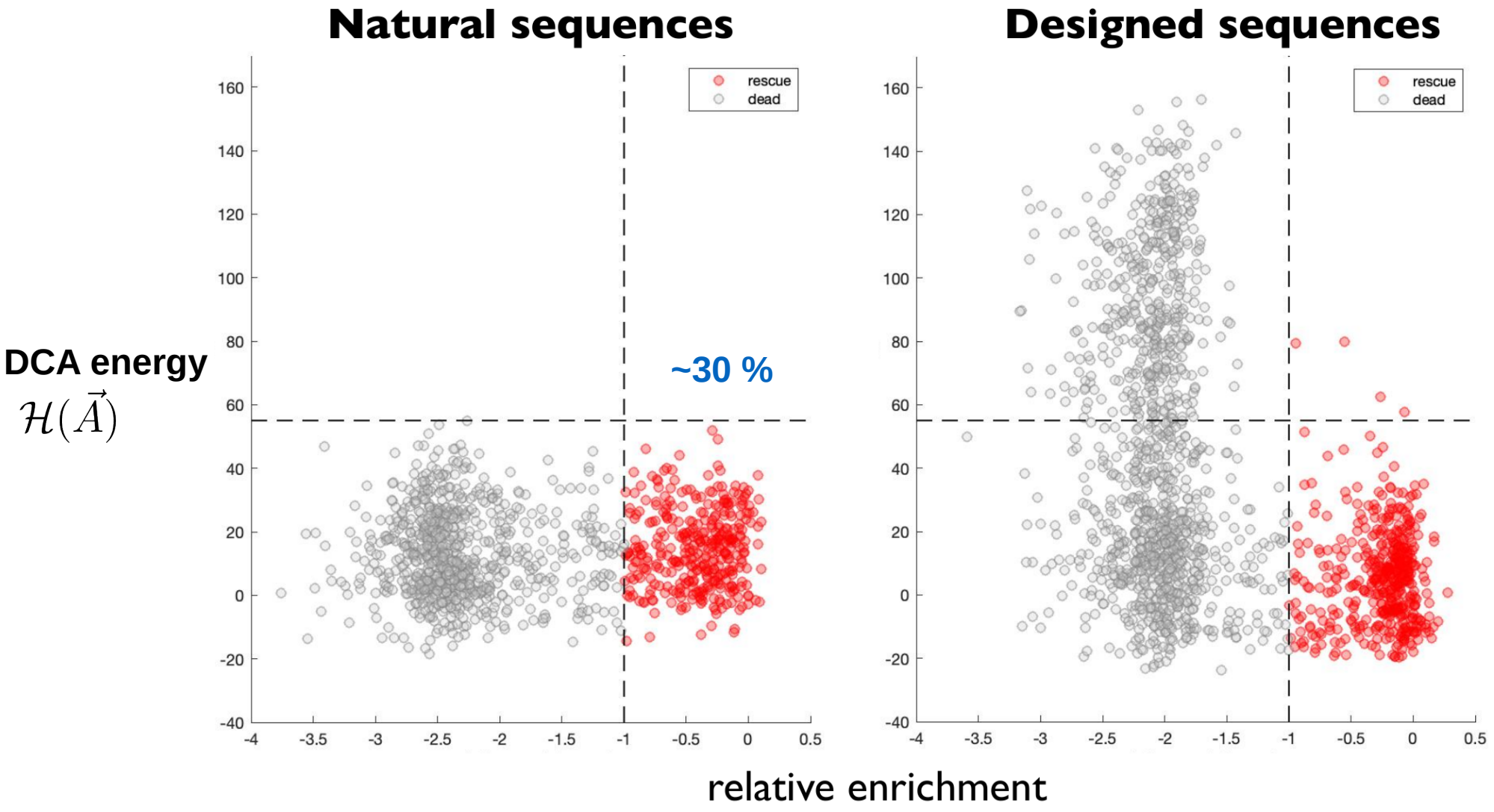

experiments by Bill Russ

Phenotype: $r.e. = \log \frac{f^t_{seq}}{f^0_{seq}} - \log \frac{f^t_{wt}}{f^0_{wt}}$ enrichment of designed sequence relative to wildtype (E. coli)

# Protein design: Chorismate mutase



# Protein design: Chorismate mutase




→ **Low energy DCA sequences are variable and functional**

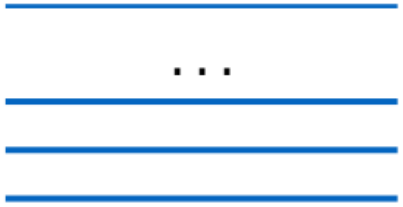
# Protein design

## Chorismate mutase

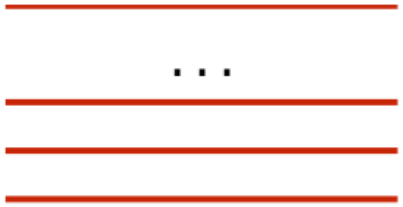
enzyme in the synthesis pathway of phenylalanine and tyrosine

with Rama Ranganathan's group

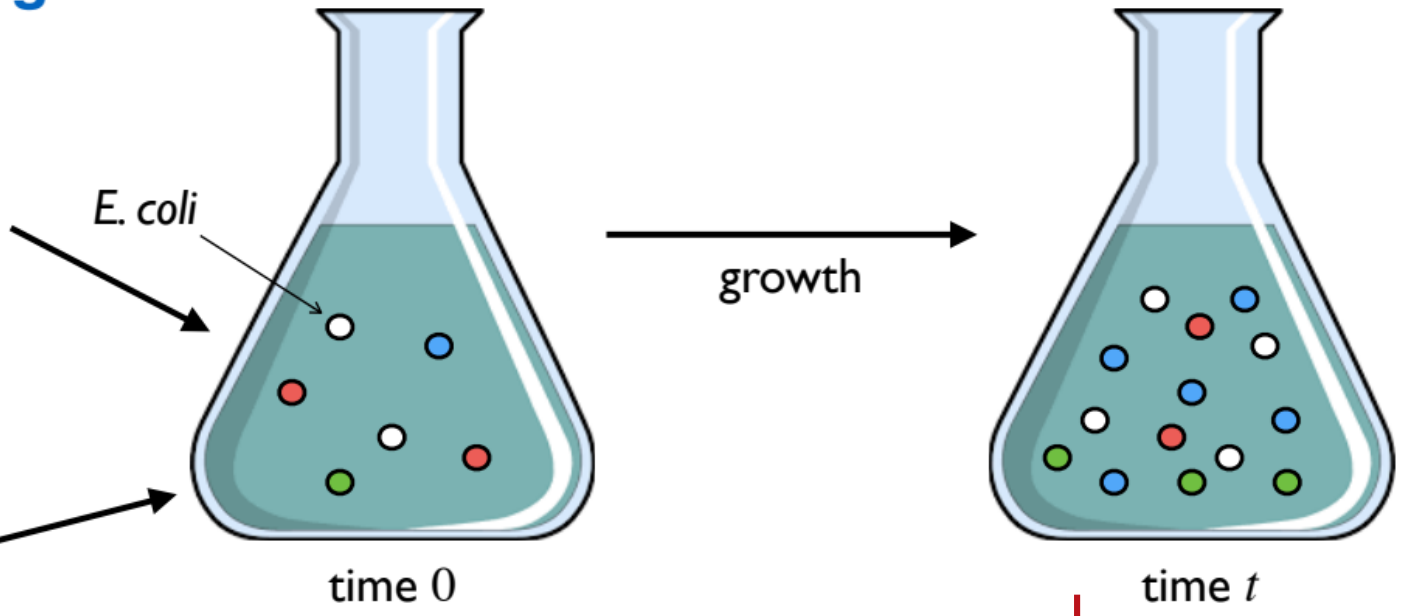
1130 **natural homologs**



**DCA sequences**



$$P(\vec{A}) \propto e^{-\mathcal{H}(\vec{A})}$$




experiments by Bill Russ

**Feedback!**

Phenotype: $r.e. = \log \frac{f^t_{seq}}{f^0_{seq}} - \log \frac{f^t_{wt}}{f^0_{wt}}$

enrichment of designed sequence relative to wildtype (E. coli)

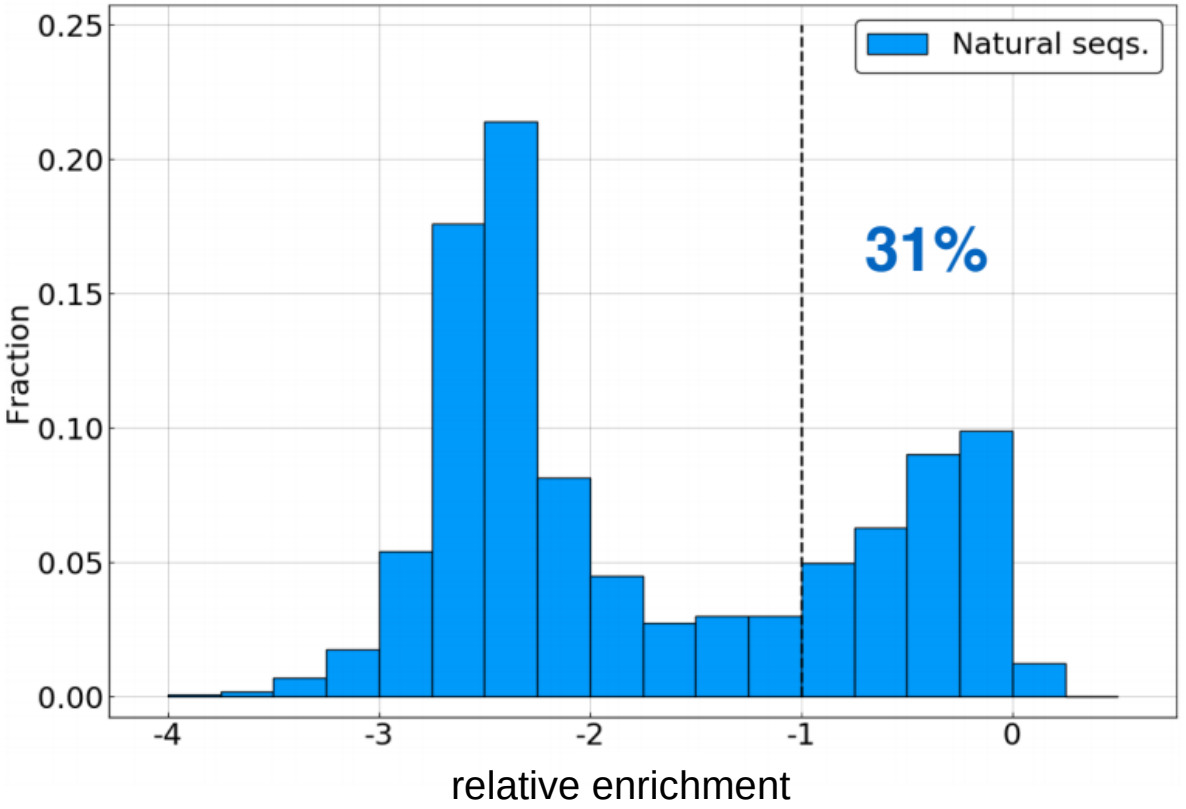
# Additional node

Not all natural seqs. are functional!

1130 **natural homologs**

| Sequence | x |
|---|---|
| ——— | 1 |
| ... | |
| ——— | 0 |
| ——— | 0 |
| ——— | 1 |

**Model with additional node**




$$\mathcal{H}(\vec{A}, x) = \mathcal{H}^{DCA}(\vec{A}) - \sum_{i=1}^{L} \xi_i(a_i, x) \longrightarrow P(x = 1|\vec{A}) \ ?$$

**Supervised** learning problem:
Infer parameters from **natural sequences** and **phenotypes**

→ **Test on designed sequences !**

# Additional node

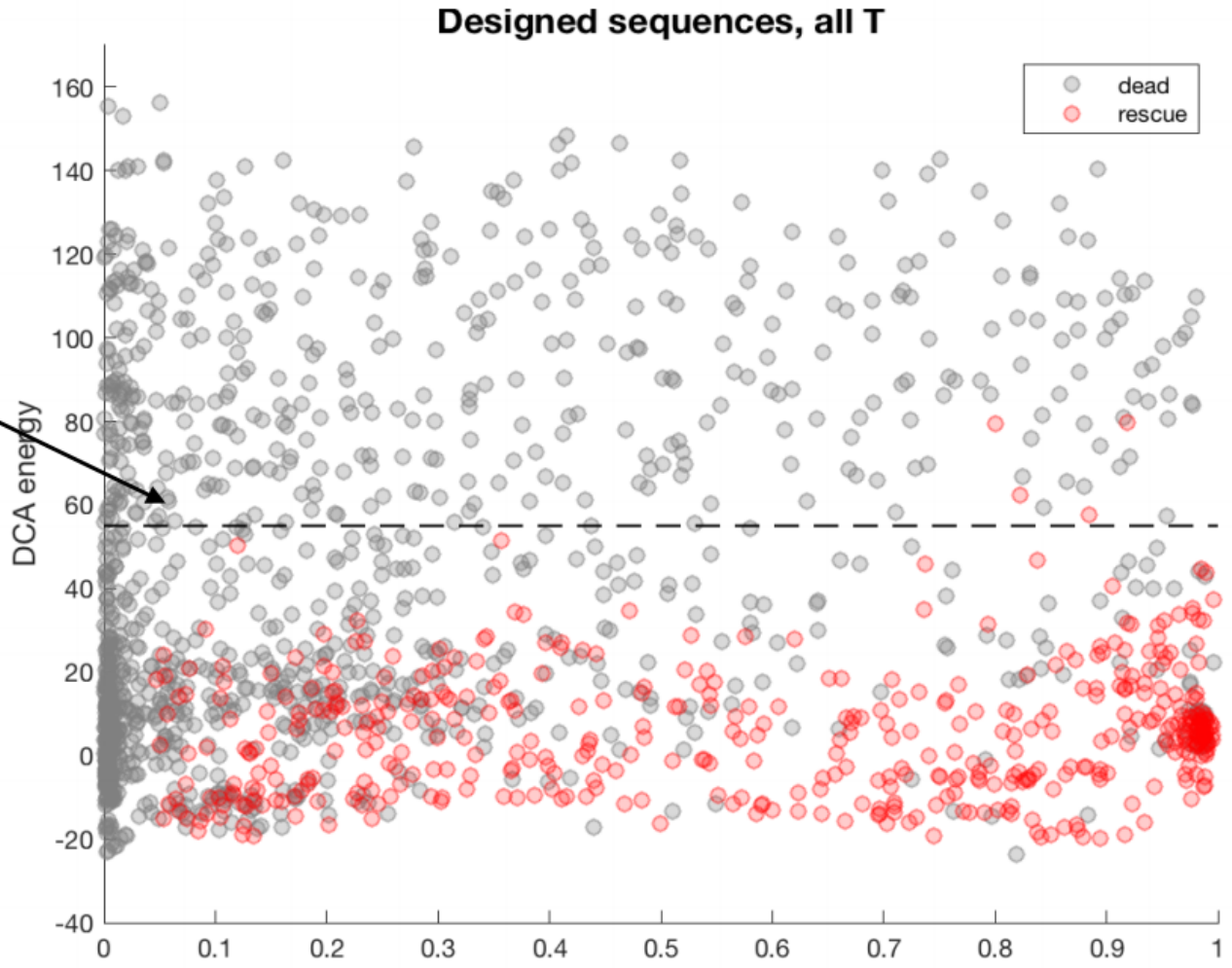
Highest natural seq.

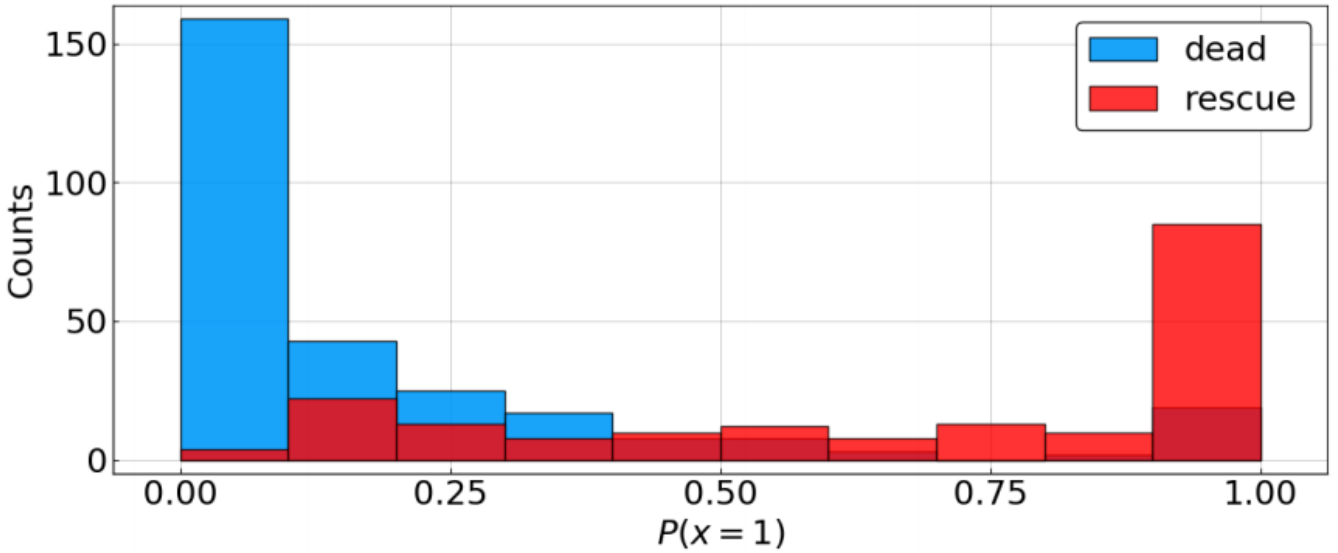
Low energy (<55)

+

High P(x=1) (>0.8)

82% functional!



# Conclusion

- Alignments of homologous proteins contain **sufficient information** for generating **non-natural functional sequences**

- This is done by modeling homologous sequences with a **pairwise exponential model**

  **Direct Coupling Analysis**
  - ➔ Fitted on **conservation** and **correlation** in the alignment
  - ➔ Reproduces **non-fitted quantities**
  - ➔ Can be improved using **experimental feedback**

# Acknowledgments

## Collaborators

Matteo Figliuzzi

Simona Cocco

Bill Russ

Martin Weigt

Rémi Monasson

Rama Ranganathan

# Acknowledgments

## Current group in Unibas

Richard Neher

Emma Hodcroft

Nicholas Noll

Eric Ulrich

# Thank you !