

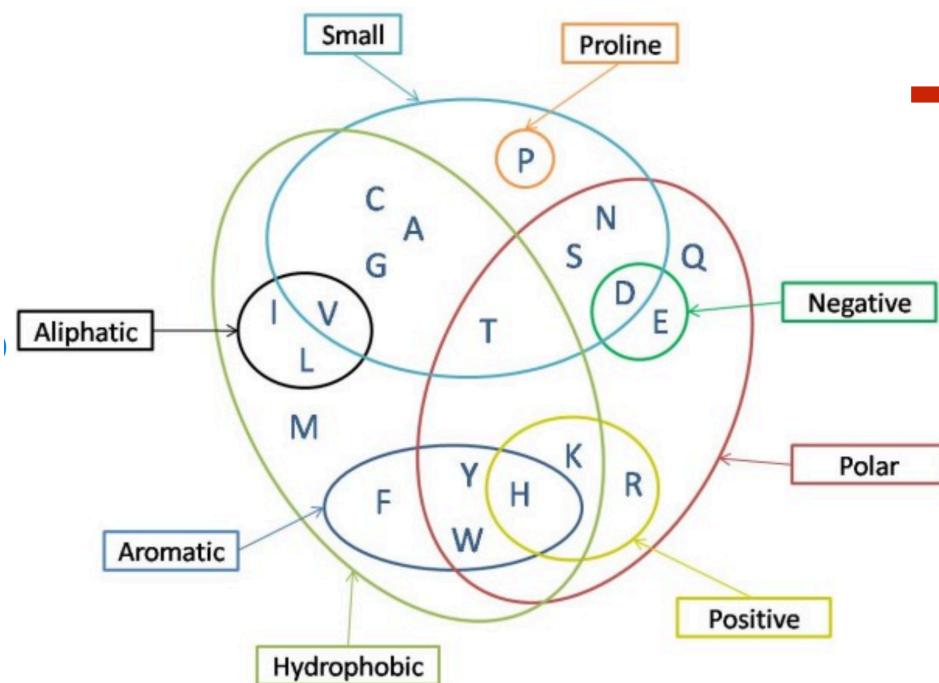
A machine-learning project: unsupervised and supervised characterization of protein sequence data

Martin Weigt

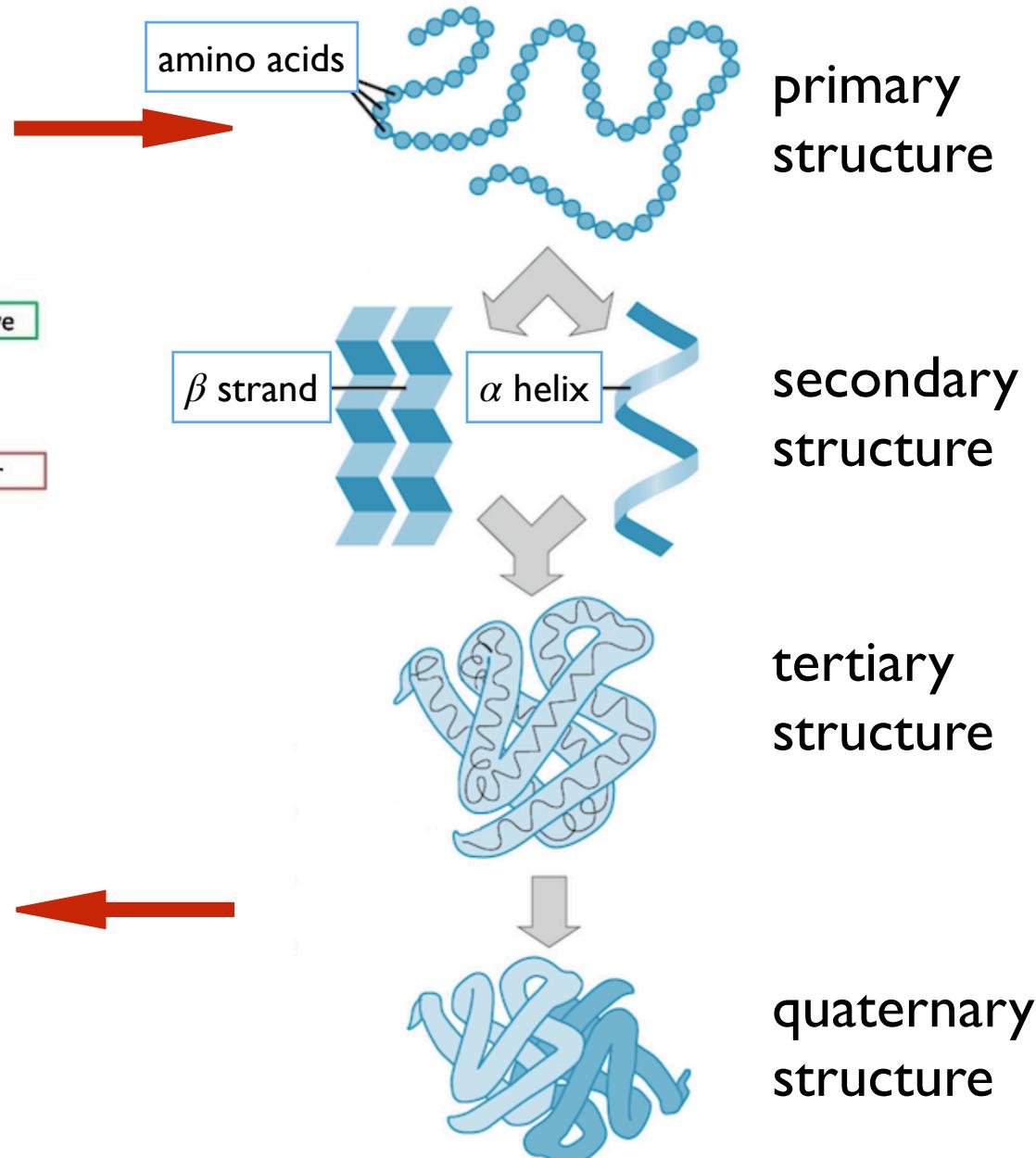
Laboratoire de Biologie Computationnelle et Quantitative
Sorbonne Université

Proteins - workhorses of the cell

Built from 20 amino acids
(residues):



Orders of protein structure:

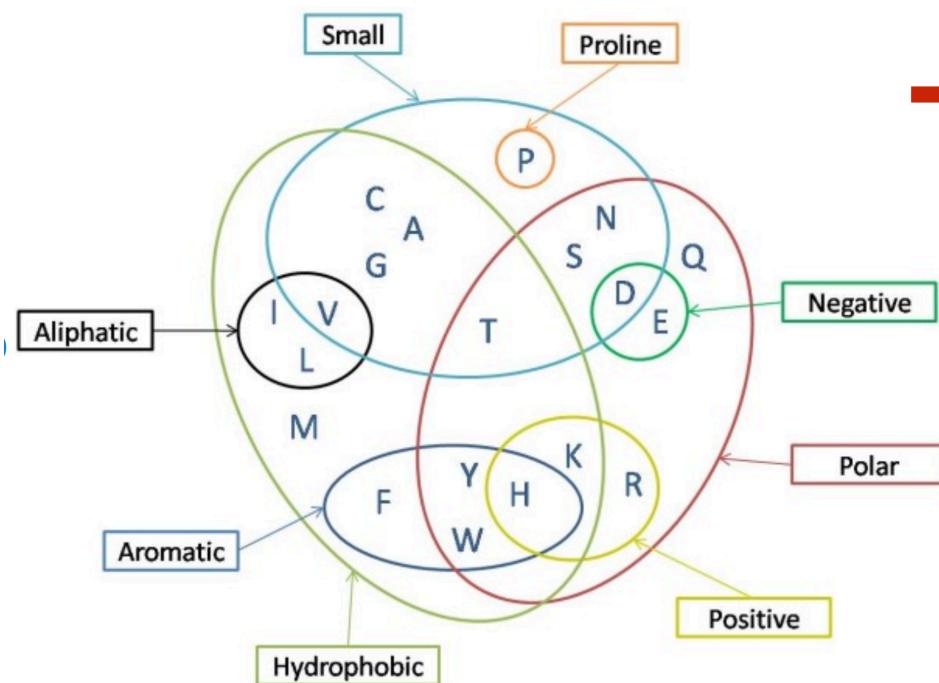


Multitude of functions:

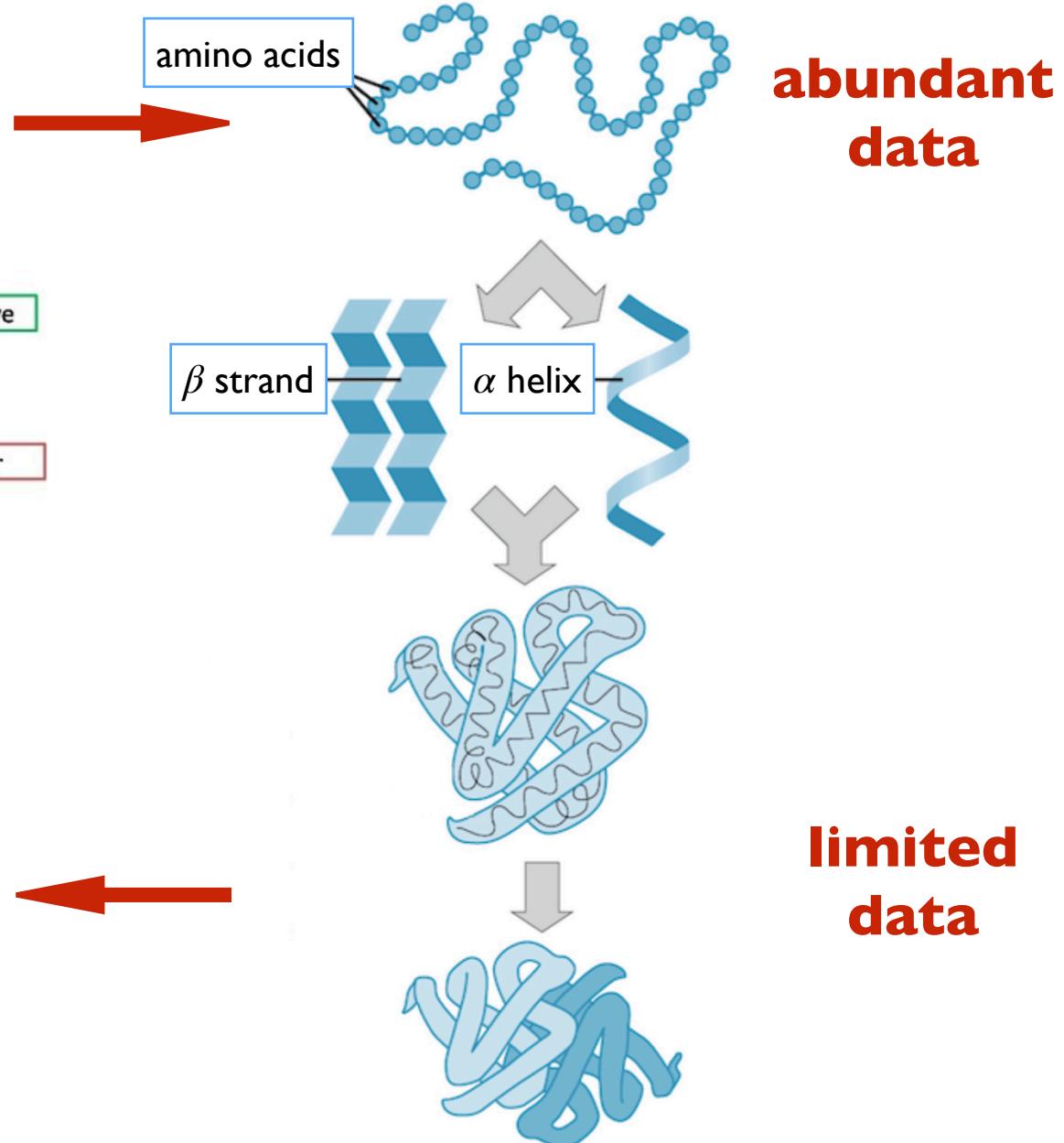
- enzymes & catalysis
- transport
- cellular structure
- signal transduction
- protein synthesis...

Proteins - workhorses of the cell

Built from 20 amino acids
(residues):



Orders of protein structure:



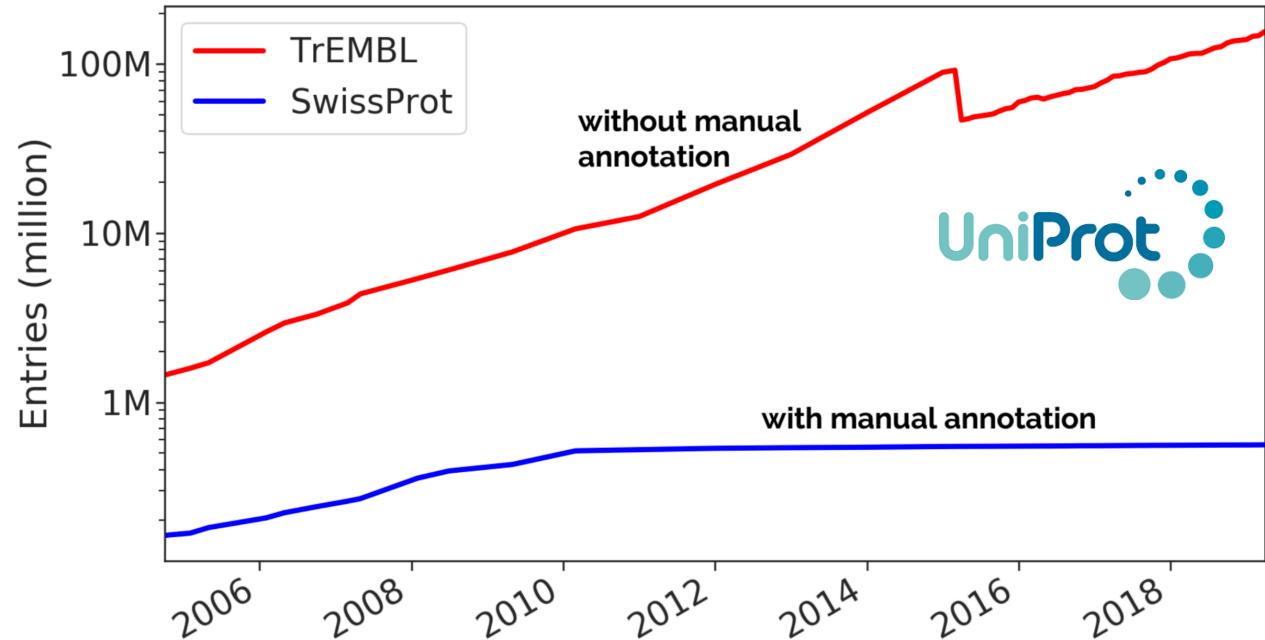
Multitude of functions:

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

limited
data

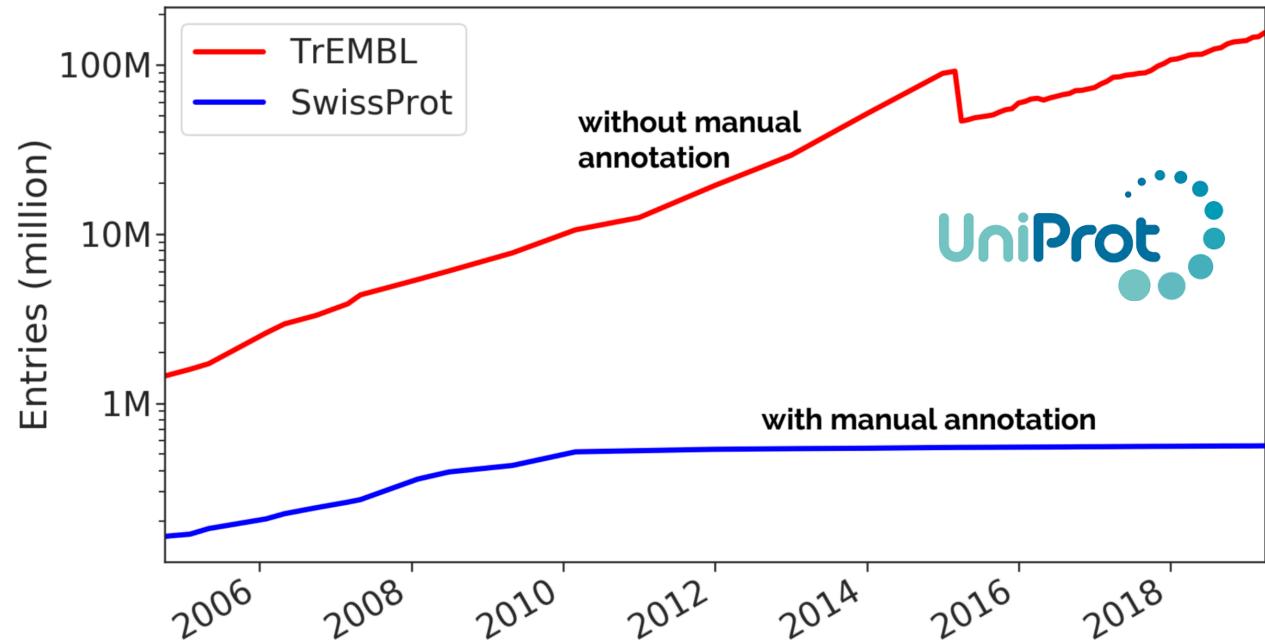
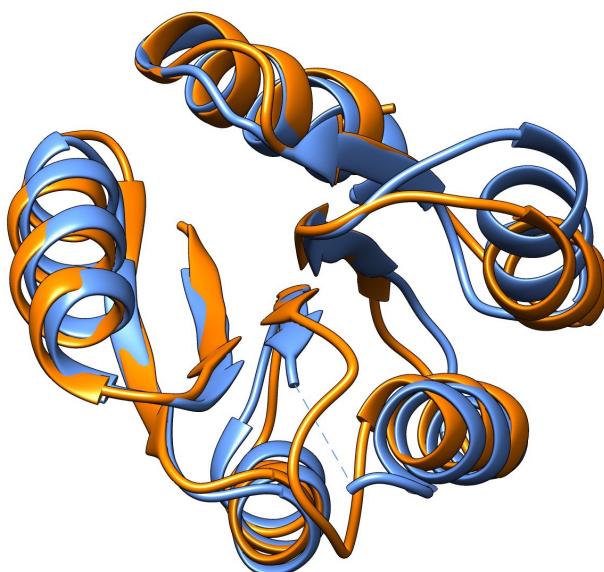
Sequence data

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

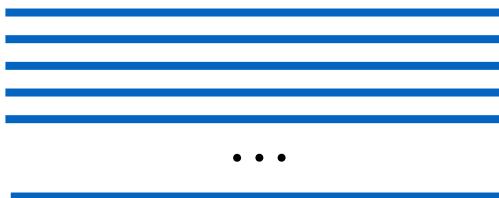
-LNQFADDLAHEL RTPVNILLGKNQVMLS-QERSAEEYQQALVDNIEELEGLSRLTENILFLARAEH-
ALGE LTAGIAHEINNPTAVILGNTELIRFLGADASRV-EEEIDAILLQIERIRNITRSLLQYSRQG--
SQRQFVTNASHELKTPIAIIISANTEVLEI----TMGK-NQWTETILKQVKRLSGLVNDMVALAKLEE--
---AFVSNA SHELRTPVTSIKGFAETIKG-MSAEEEAKDDFLDIYKESLRLEHIVEHLLTLSKAQ--
-VGQLTGGIAHDFNNMLTVIGS LDLIKLS----GRLVERFMDAALISAQRAASLTDRLLAFSRRQS--
---RMTHQVSHEVGNMIGIITGSLLERETGFNDRQ-KRHIARIRKAADRGRSLASSMLTIGS----
ALGEMLDHIAHQWKQPINSISLIAQDMADYGE LTDGDVQTTIDKIMSLLEHMSQTVDFRGFYR----
-VGRLAGGV AHD FNNL LSVINGYCEMLAA-QVSDRPQALREVSEIHRAGLRAAGLTRQLLAFGRRQ--
SLGELAAGVAHEINNPNAVILLNVDLVKKWSEMSEEL-PLLLTE MEEGAGRIKRI VDDLKD FARGD--
-MGEFAAYIAHEINQPLSAIMTNANAGRNEPSNIPEAKEALARIIRDSDRAAEIIRMVR SFLKRQ--
--GQLAGGIAHDFNNILQIISGNTQILQYQTNPDP-----QLEILKAVERGTALTRSM LAFSRKQT--
--GQLTGGIAHDFNNL LQVILGNLEFVRAKLDGDAK-LQTRIERAAWAAQRGATLTGQLLAFARKQ--
AKTDFLSNMSHEIRTPNAILGFIQVLKD-AEMKPKD-REYLELMDESSKNLLSLVNDIIEIDLIESG
--GREVLHLVHDLKTPLATIEGLVSLMET-RWPDPKM-QEYCQTIYGSITSMSKMVSEILY-----
-RARLLADVAHEL RTPVATLTGYLEAVEDVRPLDAST----IAVLRDQAVRLTRLAQDLADVTHAEGG
SMKRMLTNMSHDLKTP TVLGYIETIQSDPNMPDEERERLLGKL RQKTNEI QMINSFFDLAKLES-
AKSEFLANMSHEIRTPNAIIGFSEMIQAFGPLGSDRYEEYINDIHTSGNFLNVINDILDMSKIEAG
-MQRFIADATHQLRTPLAAIDA EVELLTD-QTRDPKA----LDKLRGRIADLARLASQLLDHAM----
-RKKAVHTT IHEL RTPLTAITGYAGLIRK-EQCEDKS-GQYIQN ILQSSDRMRDMLN TLLDFFRLDNG
-REEFMNMTSHELMNPLSAAVQAHTMISLHDDNSKS NIEIAKIILACGEHQQLKVEDARMM SKLD--
-KSRYVVGLSHELRSPNLAISGYAQLLEQDTSLAPKP-RDQVRVVRSSADHLSGLIDGILDISKIEAG
----AFSYMRHAINNPLSGMLYSRKALKN-TDLNEEQ-MRQIHVSDNCHHQLNKILADL-----
-QENFIDMTSHEMRNPLS AILQCSDEITST-----LCLEANTIALCASHQKRI VDDILTFSKLDS-
SQRTLTNAIAHDLRQPLYRIRFALEMFD-SLLSIEQRQYRQSIENSRLDLDH LINQSLQLSRYT--
--KLLLLLSHDIKTPLSAIKLNAKALSR LYKDAEKQ-REAAEHINARADEIE N FVS RITKASSE--
--HAFIADA AHEL RTPL TALKLQLQ LTER---ATSDVREVG FVKLNERLDRS IHLVKQ LLTL ARSES-
-QKNFISNASHELNTPLTSIIVTADLALS-KQRTDEEYRTALS RIMDAAGHLE-----
-RGALLTSISHDLRTPLASILGATSSLESGEELDENAR KELLSTI HDEADRLNRFVANLLDMTRLEAG
-KSEFLANMSHEIRTPNMAITGMTAIATA-HIDDPKQVKNCLRKIALSSR HLLGLINDVLDMSKIESG
-LSQFSADLAHD FRTPLANLIGOTEVTLA-HPRSAEEYRAVLESSLEEYARLSR MIEDMLFLARADH-
-SKSMFLATVSH ELRTPL YGIIGNLDLLQT-KELPKGV-DRLVTAMNNSSLLLKIISDILDFSKIES-
-AKTAFLATLSHEIRTPMNGVLGT A QILLK-TPLSTEQ-EKHLKSLYD SGDHM MTLLNEI LDFSKIEQG
SKKQLIDGIAHELRPLVRLRYRLEMSEN---LTPPE----SQALNRD IQLEALI EELLTYARLDR-
-KTQFFINTAHD IRTPLT LIKAPLEELLEEE LT DNG-ITRTNIALRNVEVLLRLVSNLINFERT---
---VFIDNMTH EMKPLTSIIGFS DLLRS-ARLDDETVHDYAESIYKEGKYLKSISSKLM DL-----

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

Statistical sequence models

Data – MSA of homologous sequences

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



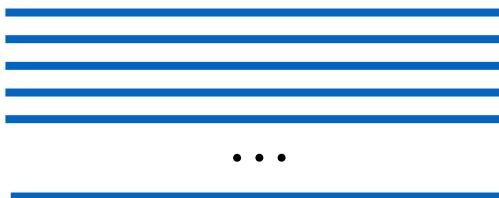
Statistical sequence model

$$P(a_1, \dots, a_L)$$

Statistical sequence models

Data – MSA of homologous sequences

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



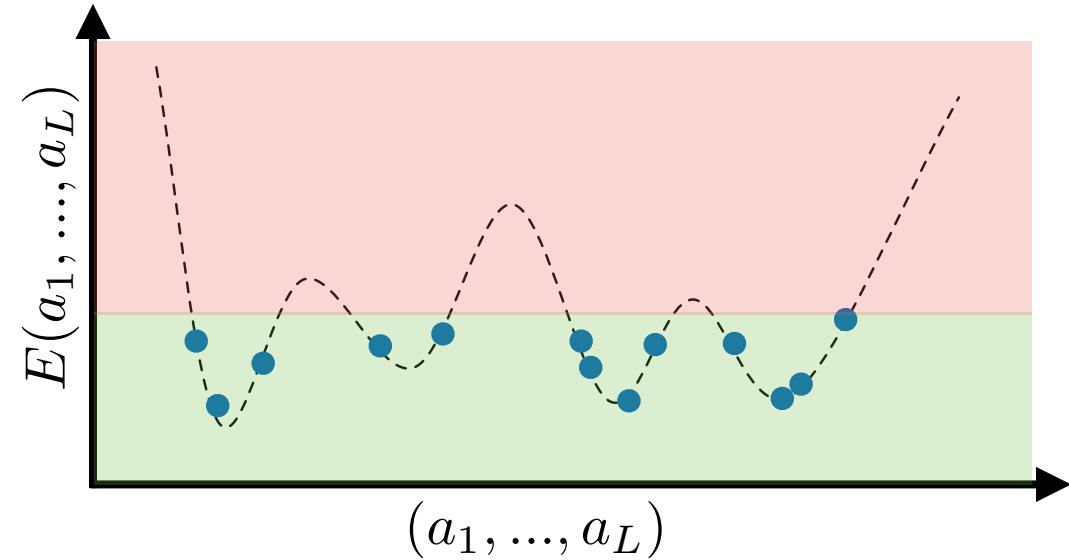
Statistical sequence model

$$P(a_1, \dots, a_L)$$

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

“statistical energy”

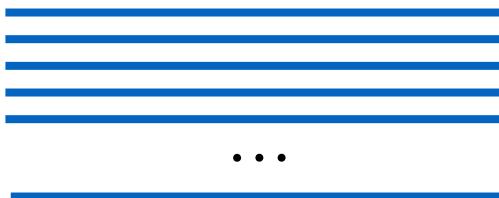
“sequence landscape”



Statistical sequence models

Data – MSA of homologous sequences

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



Statistical sequence model

$$P(a_1, \dots, a_L)$$

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

“statistical energy”

“sequence landscape”

Attention:

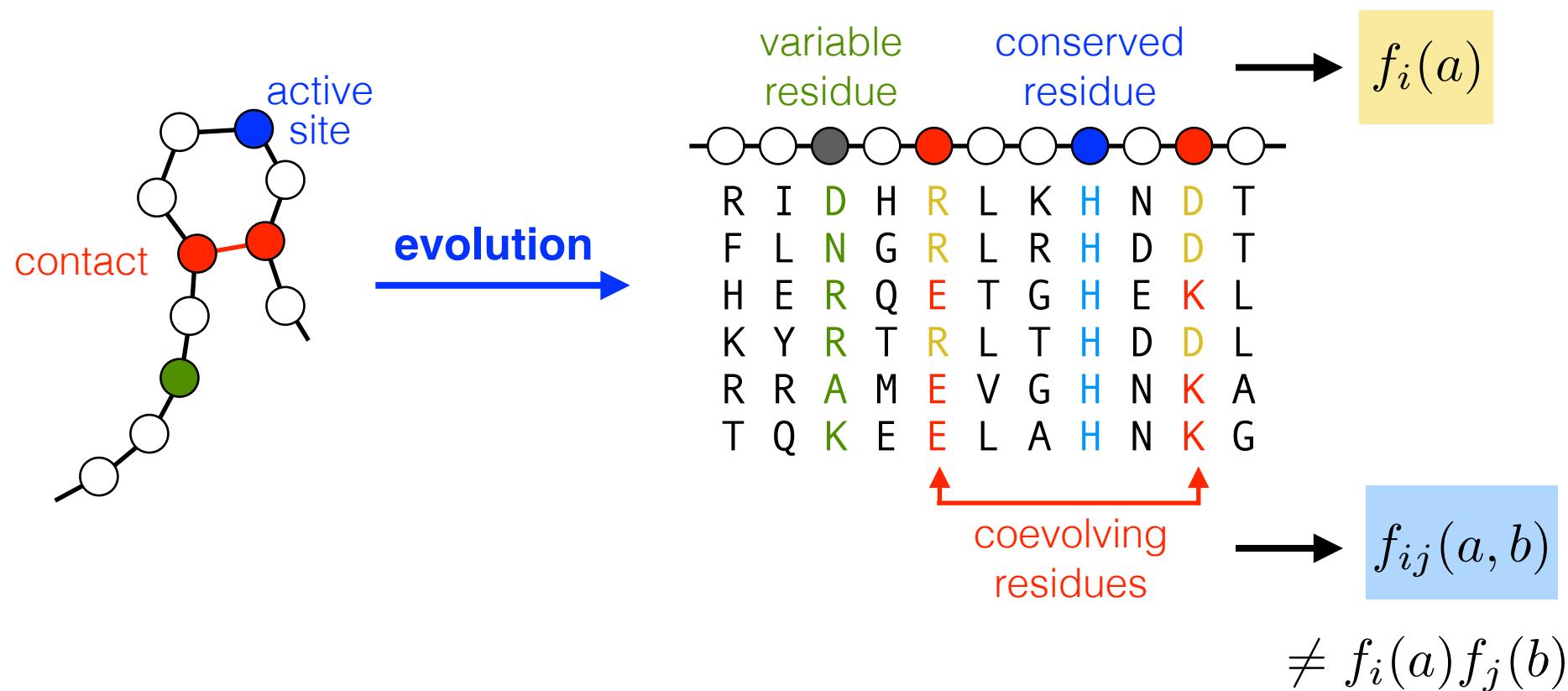
Data insufficient to do this without modeling

$$L \in (50, 500)$$

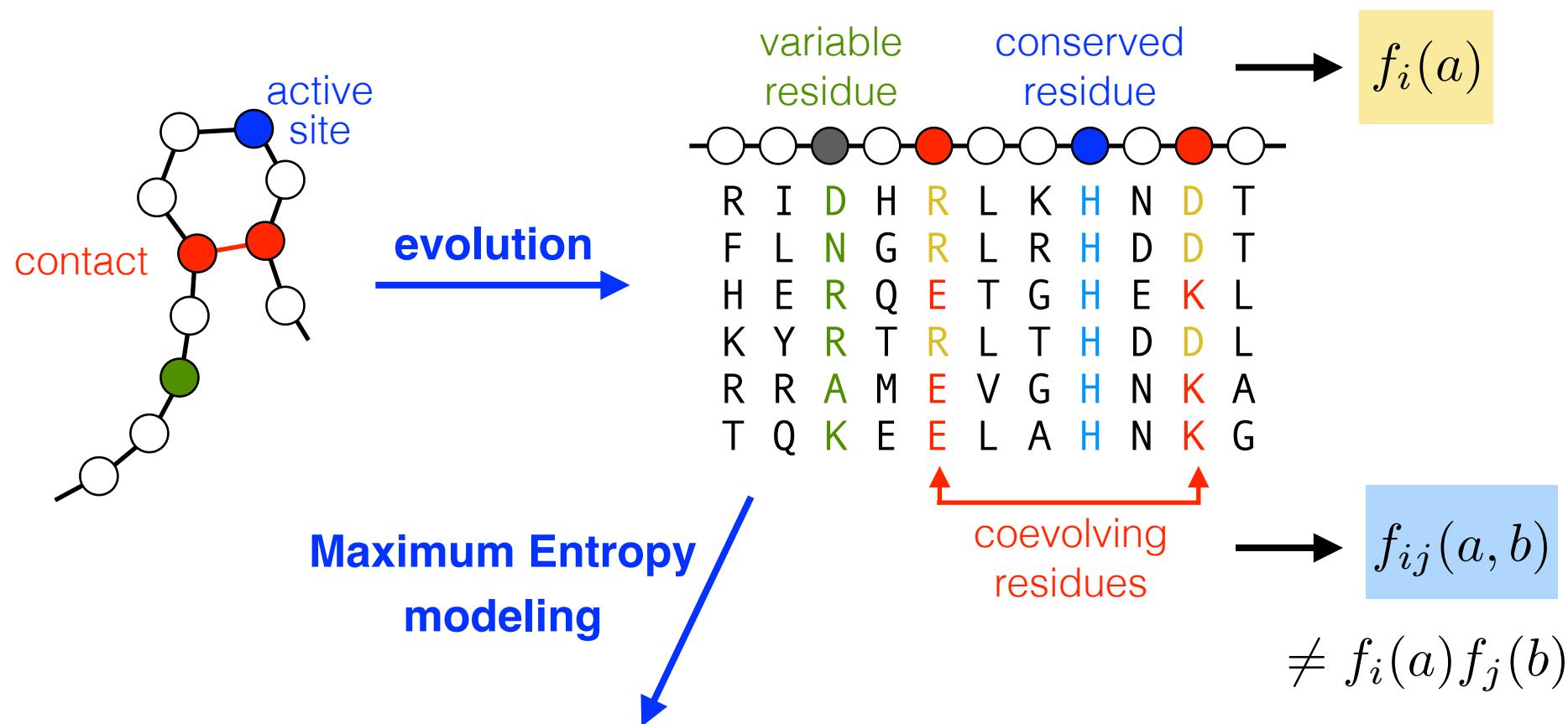
$$M \in (10^3, 10^5)$$

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

Conservation and coevolution in proteins



Conservation and coevolution in proteins



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

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

How to learn the parameters?

- **Boltzmann-machine learning:**

- start with initialised parameters (fields/couplings)
- calculate marginals

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

- 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 $\sim 2^L$
- approximations (MCMC, PCD) needed for computational efficiency

All models are wrong, but some are useful

[George E.P. Box, 1976]

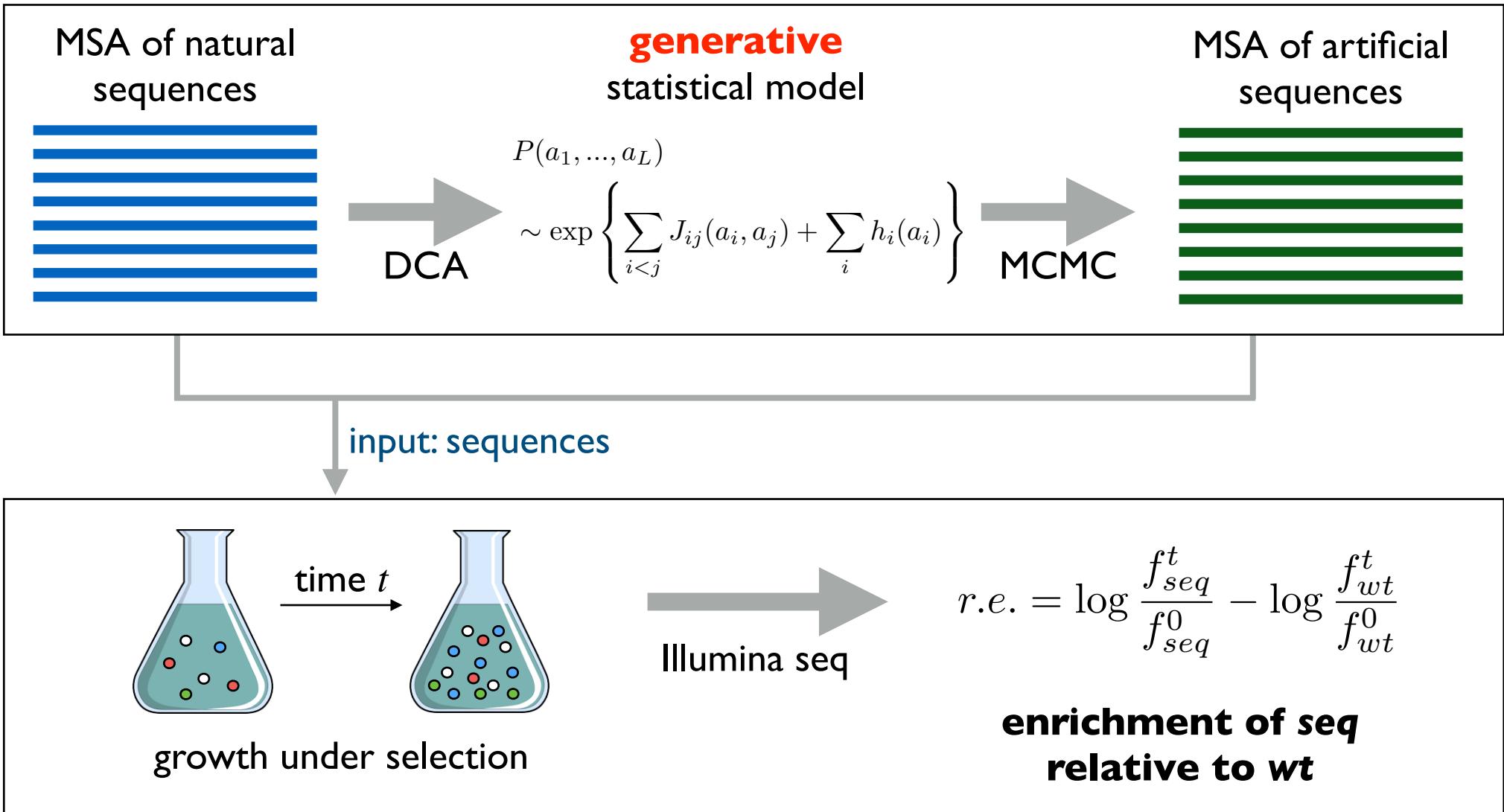
1. Are pairwise couplings **useful** ?
2. Are pairwise couplings **sufficient** ?
3. Are pairwise couplings **necessary** ?

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**

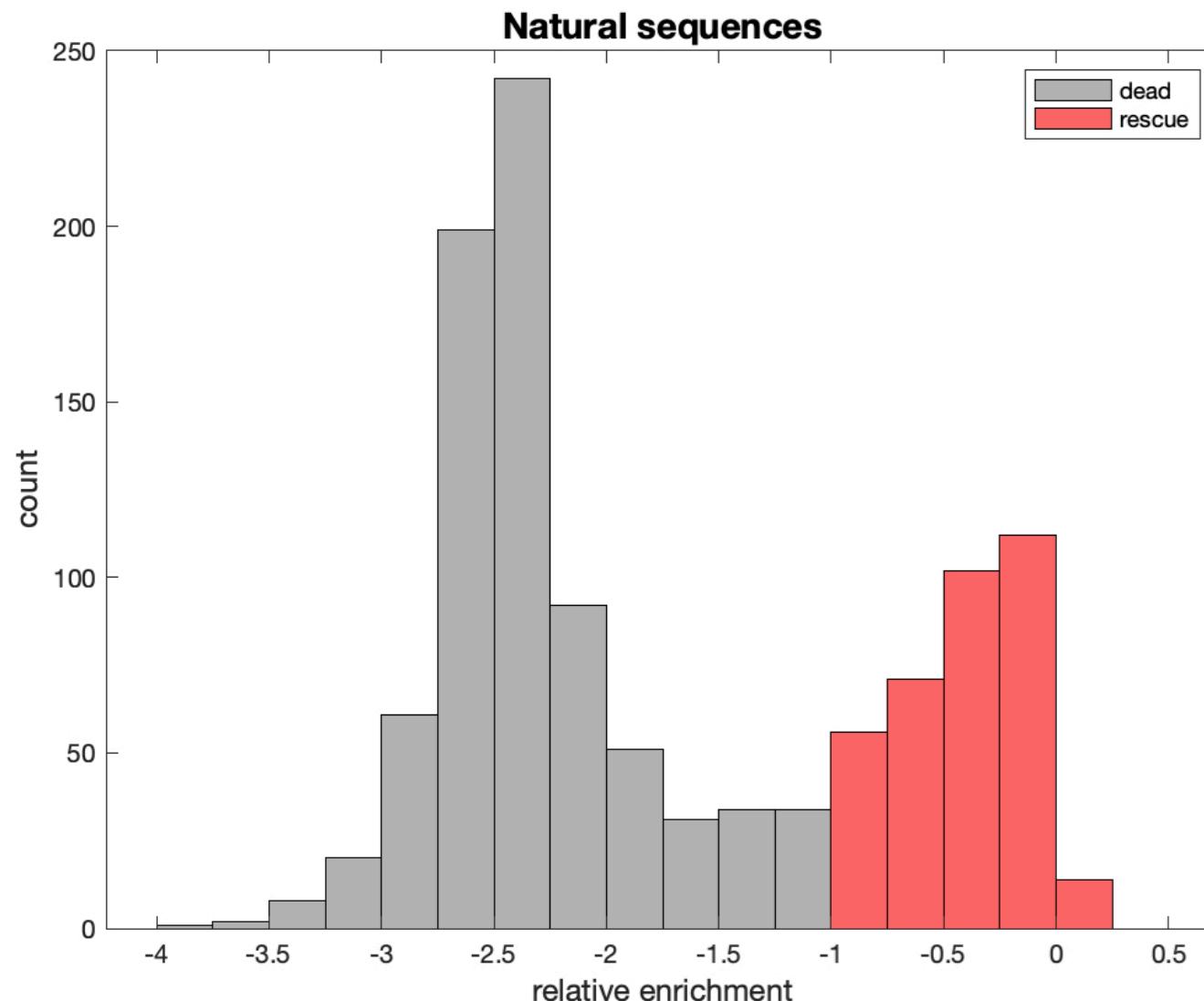
From data to sequence design



in vivo high-throughput assay

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 : $\simeq 10^{54}$

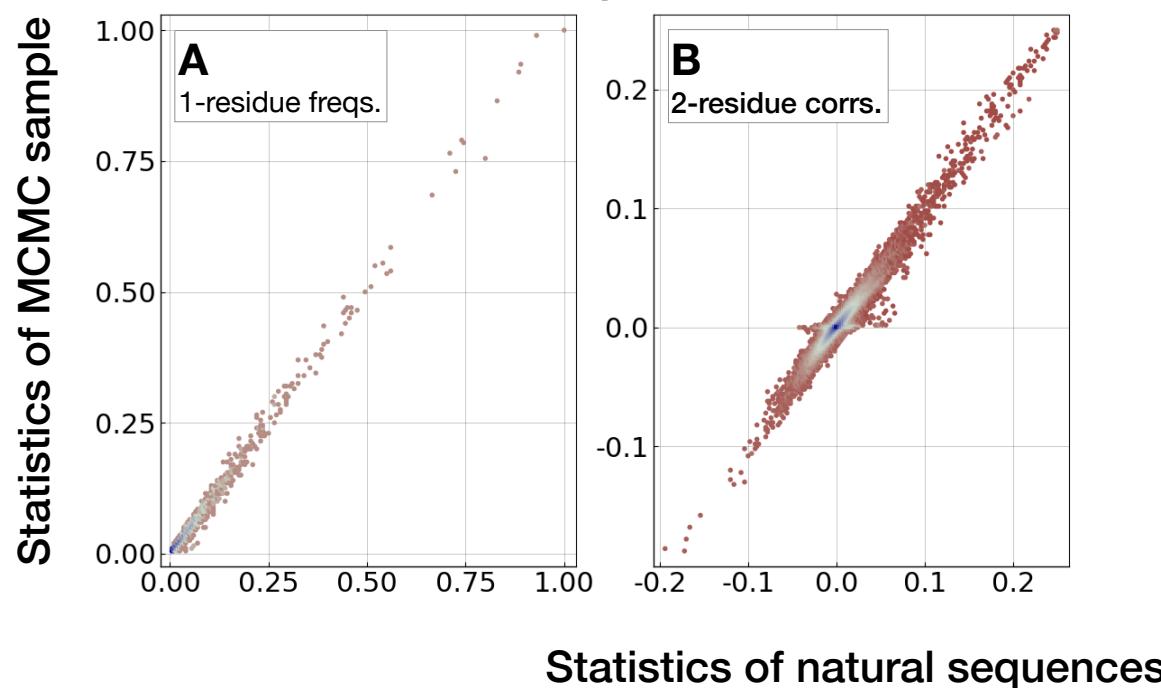
Atoms

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

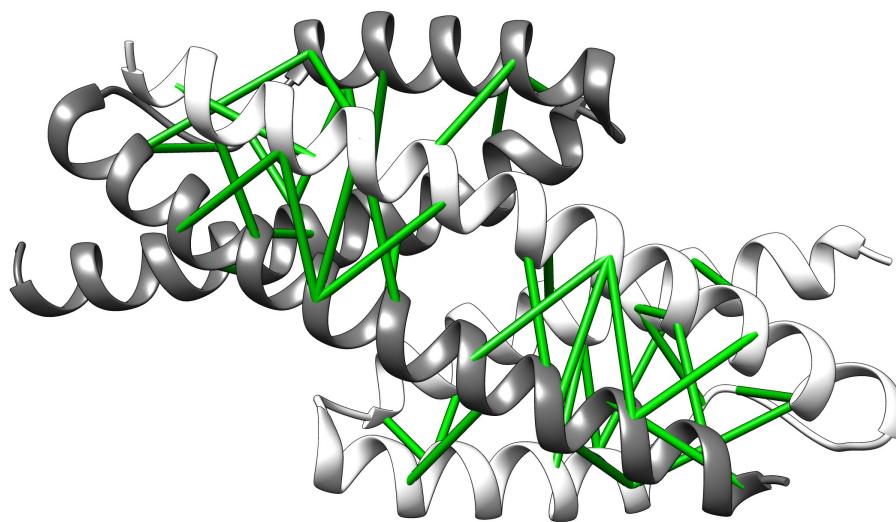
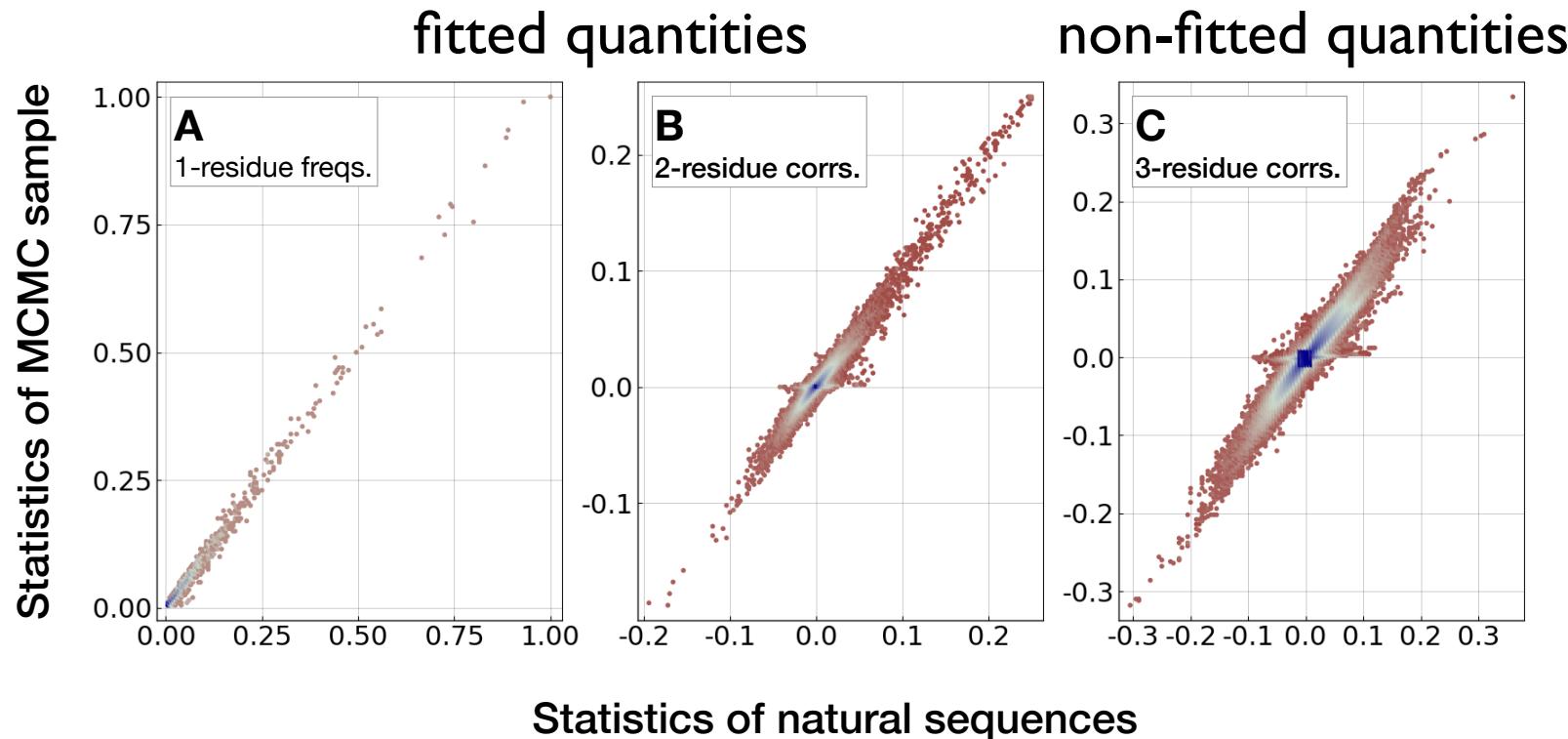
→ very good generative models are needed

Is the DCA model generative ?

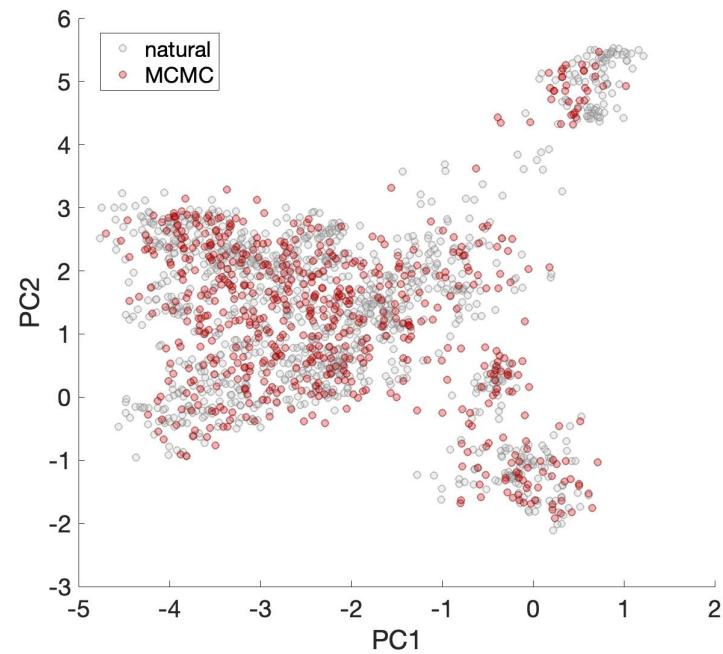
fitted quantities



Is the DCA model generative ?

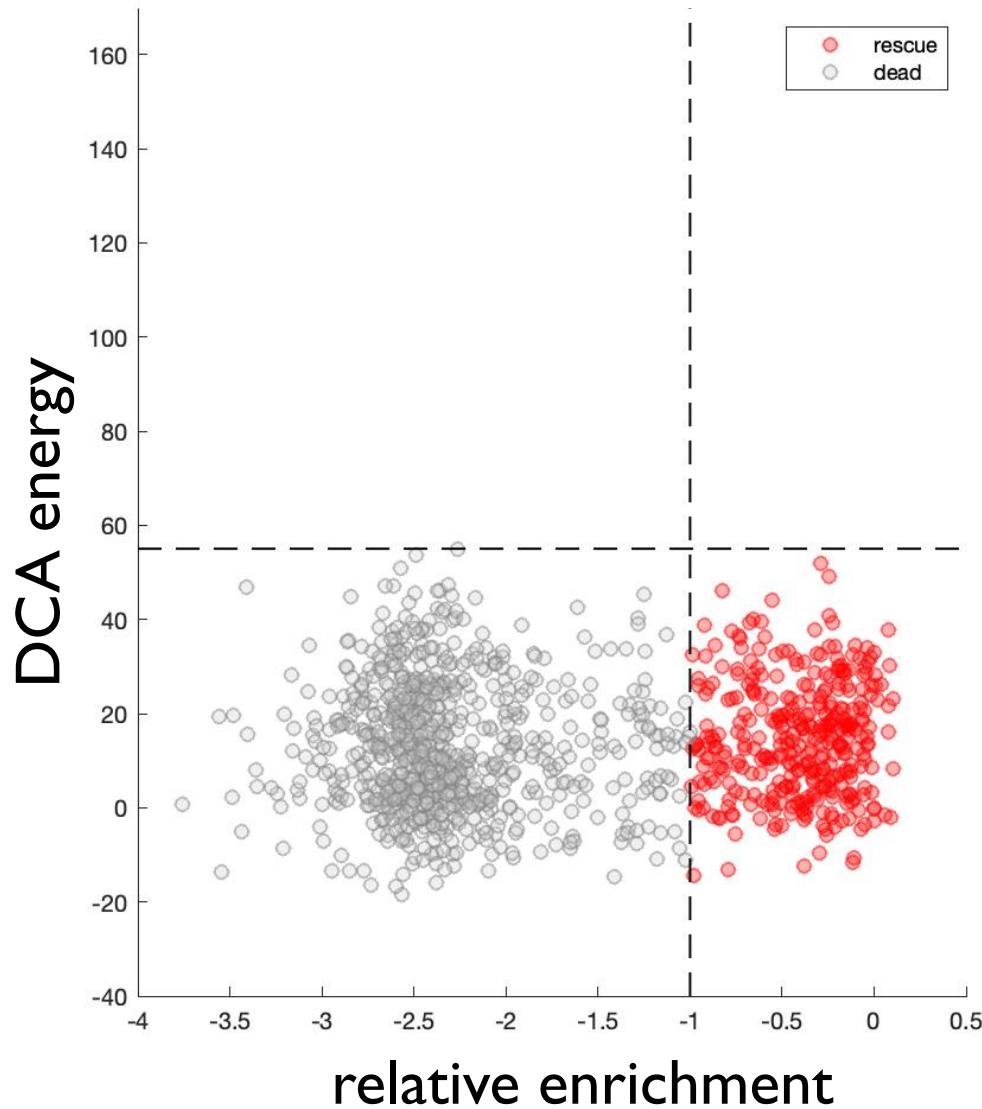


Strongest couplings = 3D contacts



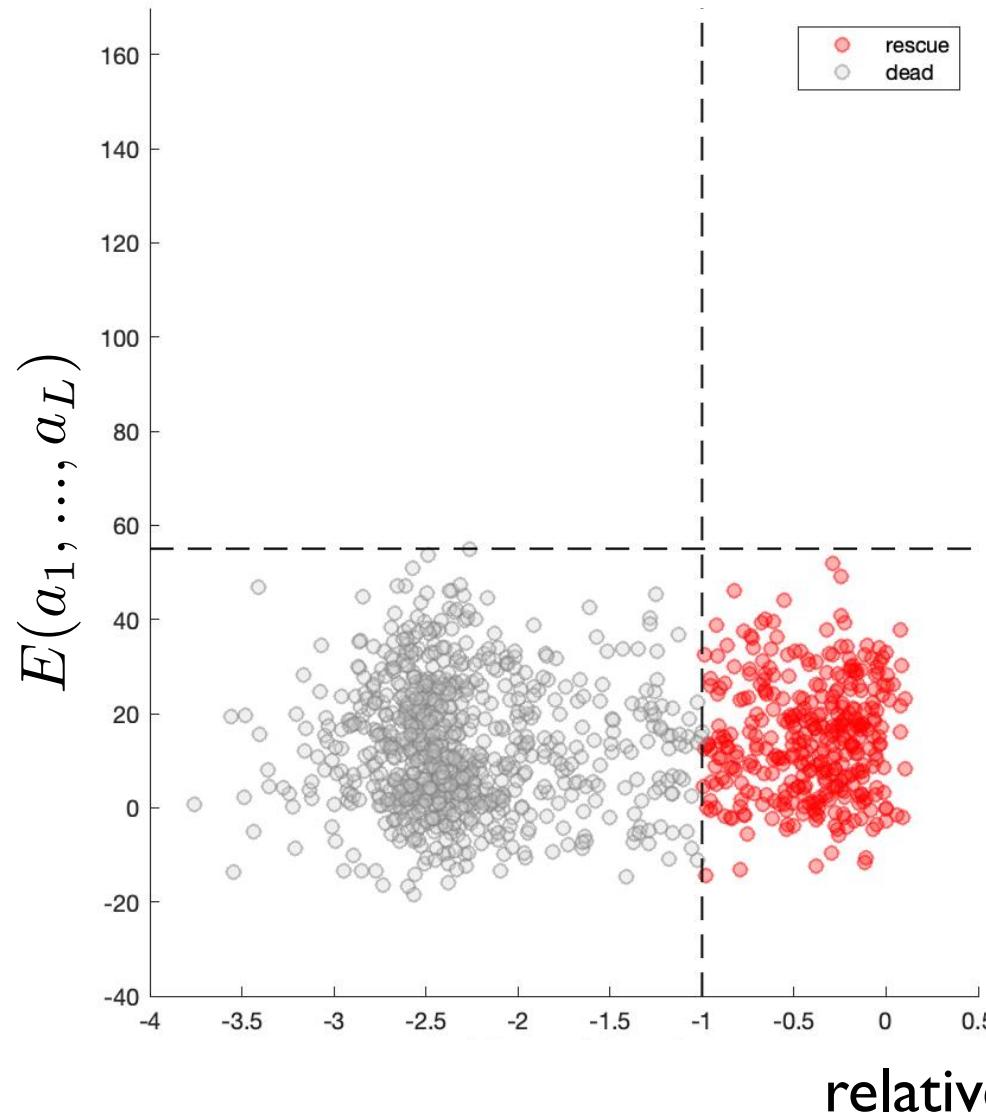
Chorismate mutase design

Natural sequences

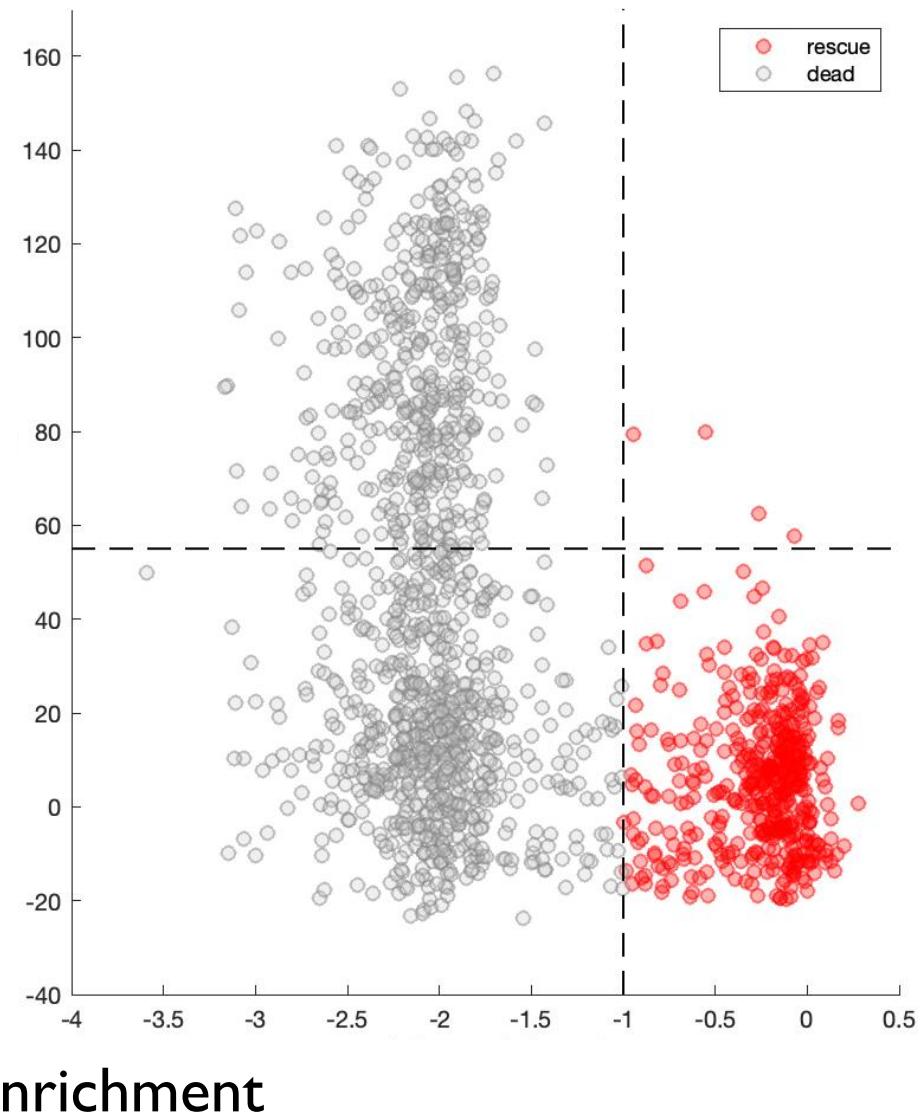


Chorismate mutase design

Natural sequences

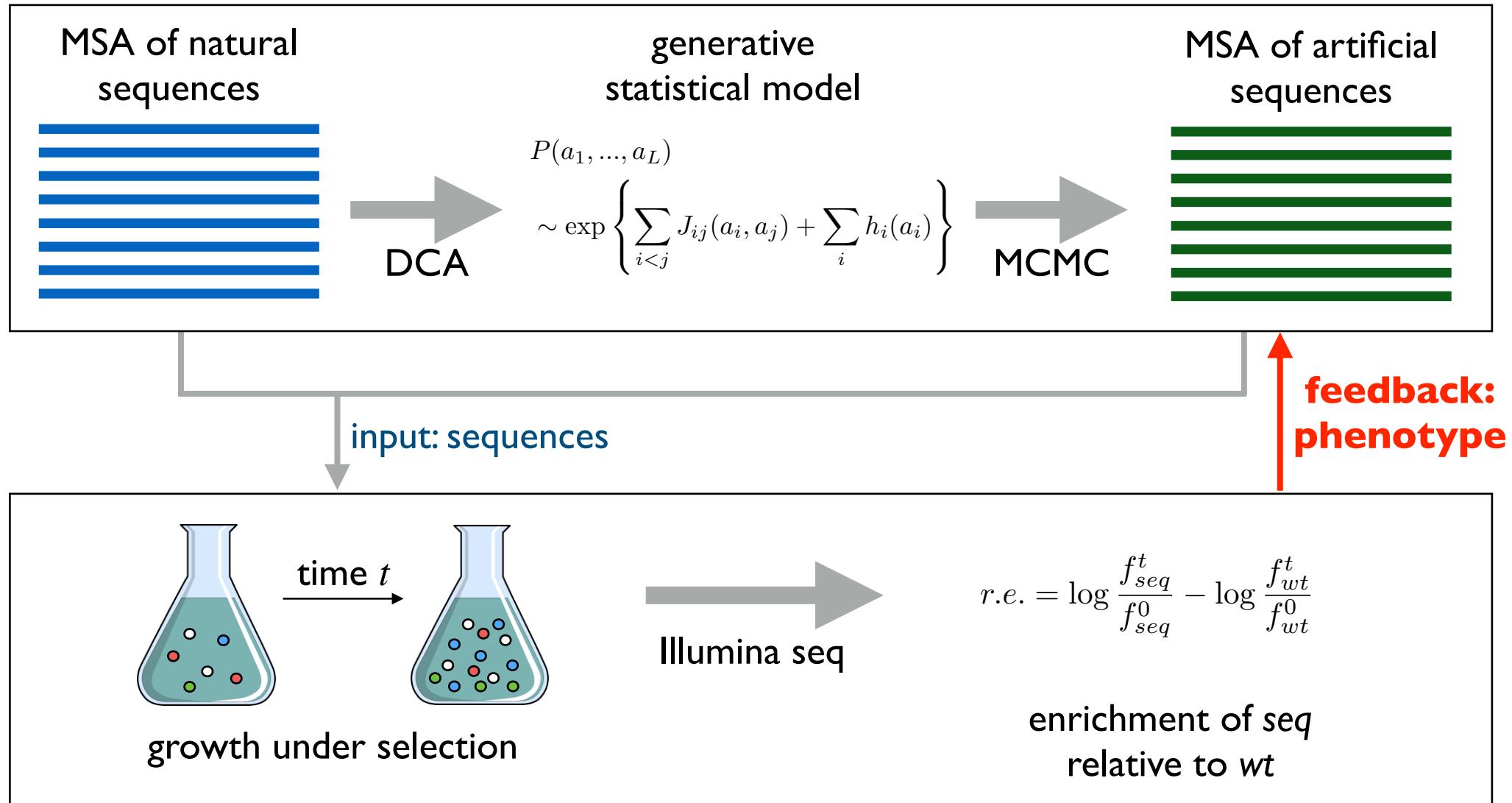


Designed sequences



- ▶ **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:

MSA of >1000 natural sequences	phenotype dead / rescue
	0
	1
	1
	0
	0
	0
	1
	0

supervised problem:

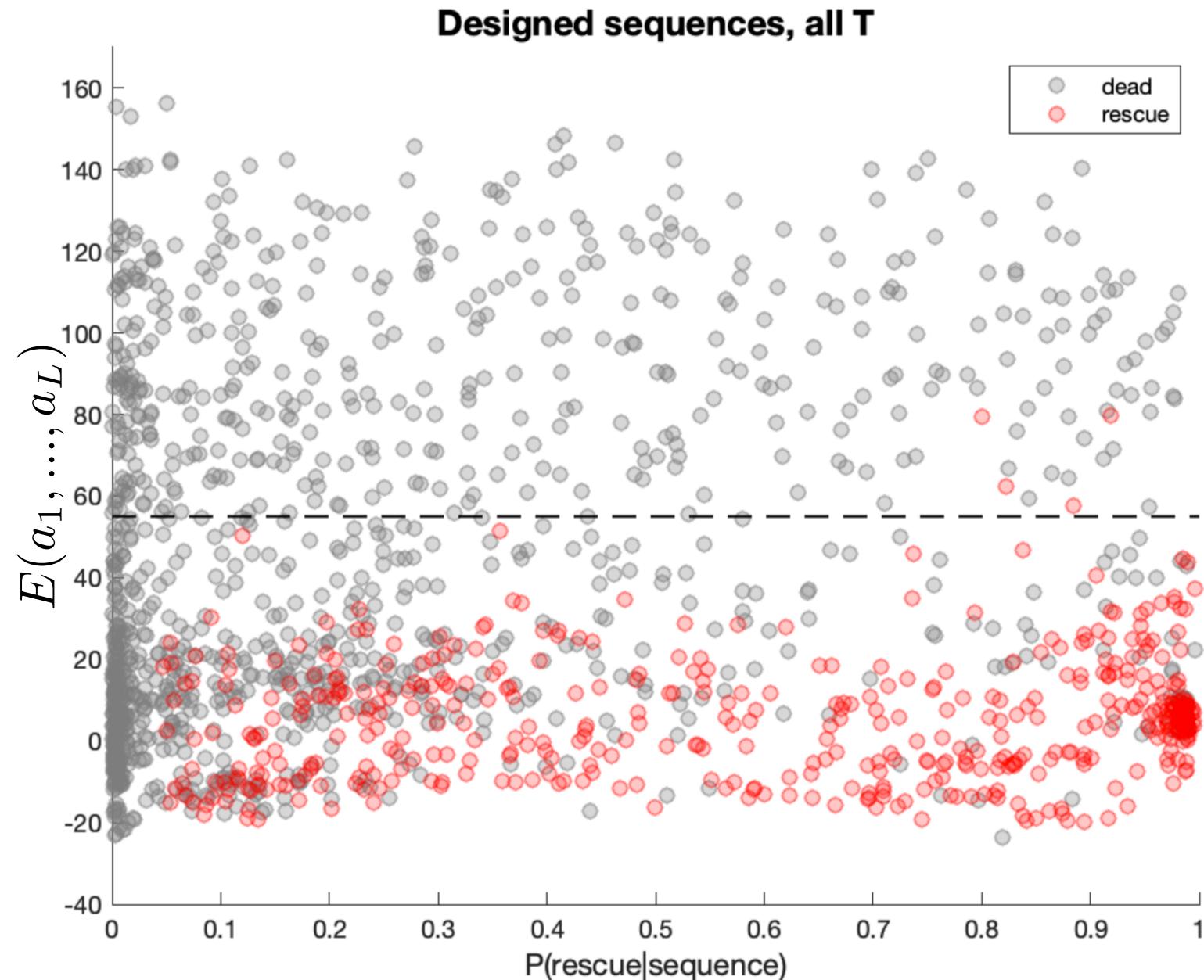
- predict phenotype x from sequence (a_1, \dots, a_L)

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

- infer parameters from natural sequences & phenotypes

Integrating phenotypic information

- validate on designed sequences

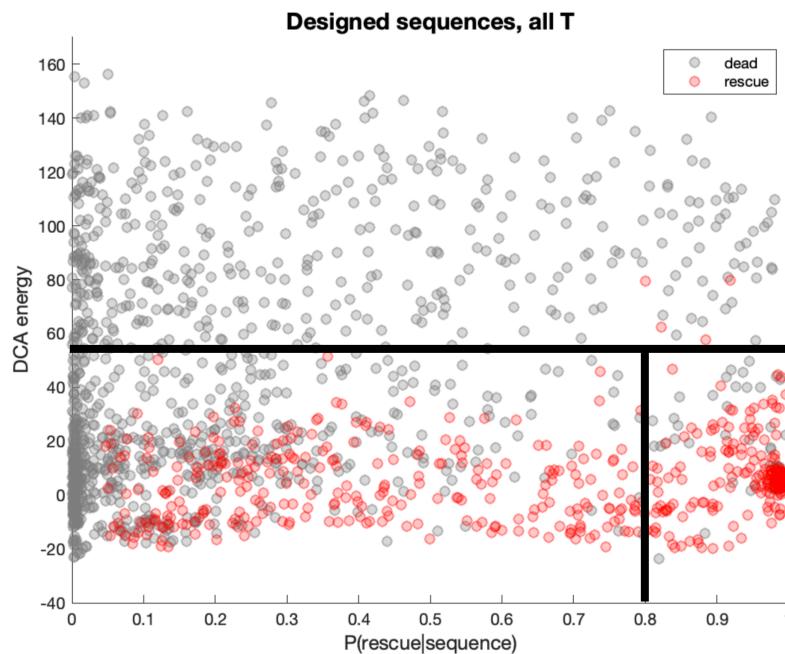


Should we be astonished ?

Sequence spaces:

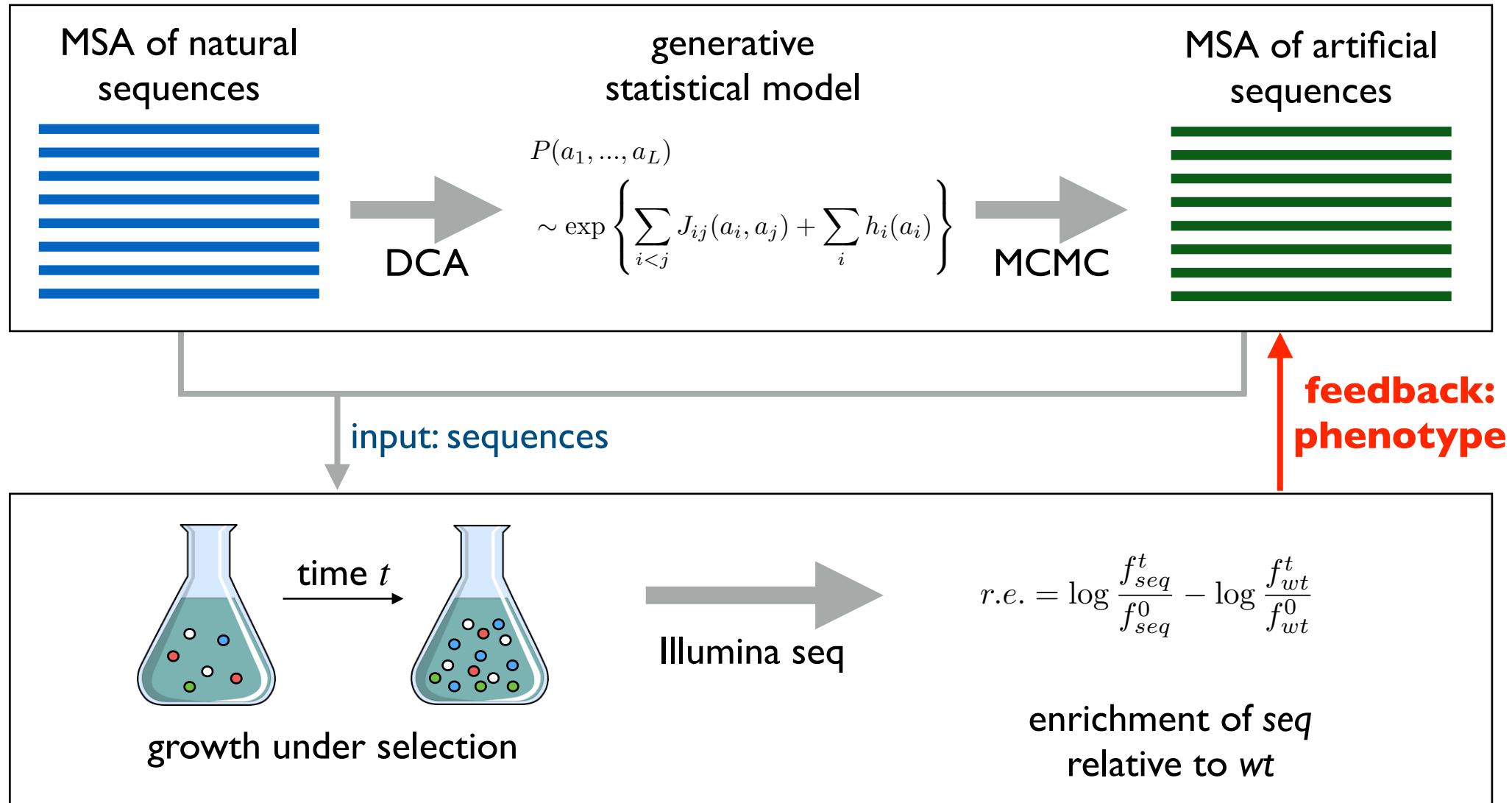
- sequences of 20 amino acids, length 96 : $\simeq 10^{125}$
- sequences coherent with conservation : $\simeq 10^{86}$
- sequences coherent with coevolution : $\simeq 10^{54}$

Success probabilities:



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

Experimental feedback for better design



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