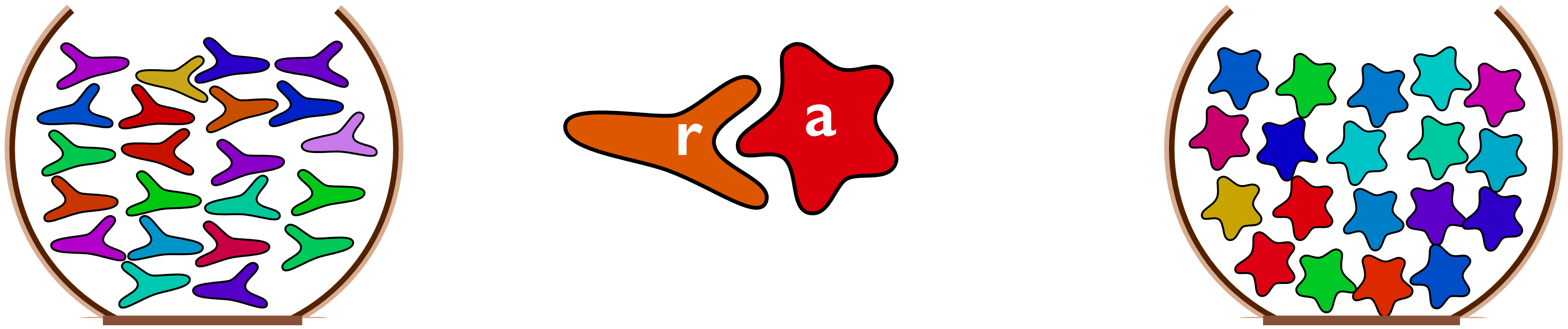


# Predictability and diversity in immune repertoires

Aleksandra M Walczak

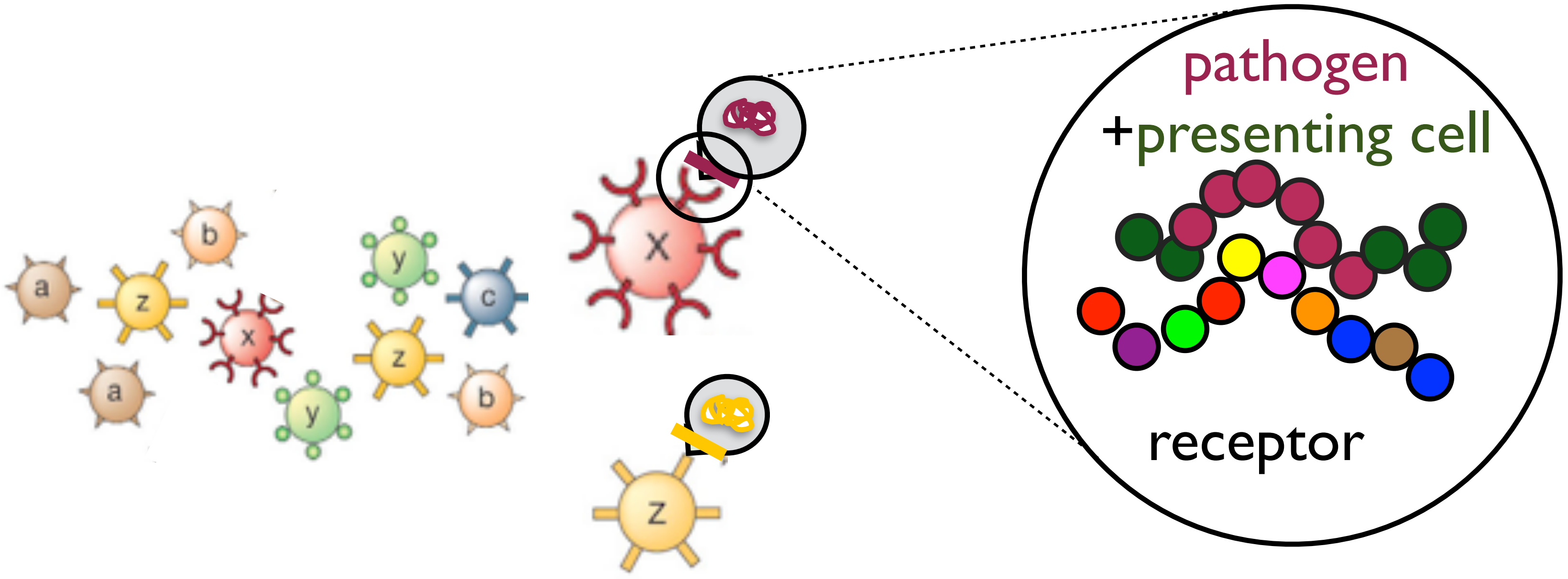
Laboratoire de Physique Théorique - ENS, CNRS

work with
Yuval Elhanati, ENS
Quentin Marcou, ENS
Maximilian Puelma Touzel, ENS
Andreas Mayer, ENS
Anand Murugan, Princeton
Vijay Balasubramanian, UPenn
Olivier Rivoire, UJF
Curt Callan, Princeton
Thierry Mora, ENS

and
Misha Pogorely, RAS Moscow
Dmitriy Chudakov, RAS Moscow
Ilgar Mamedov, RAS Moscow
Yury Lebedev, RAS Moscow

# Immune receptors

- T-cells important actors of immune system



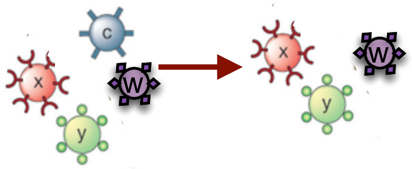
many unique receptors ↔ many different pathogens (virus...)

- triggers immune response

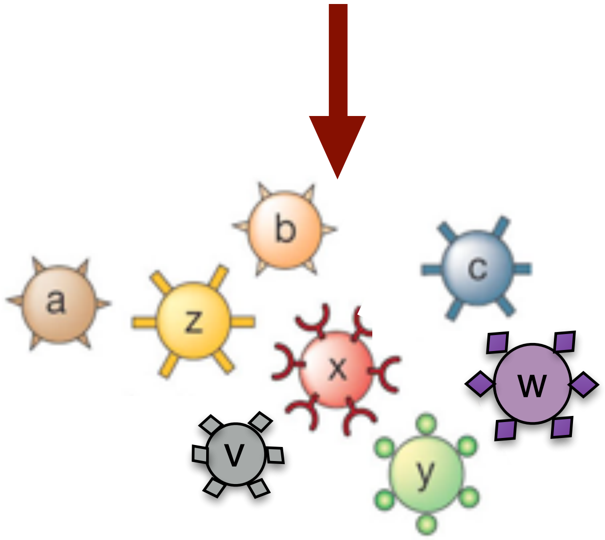
natural, healthy (=“normal”) diversity of immune receptors?

optimal distribution ?

# Repertoire evolution



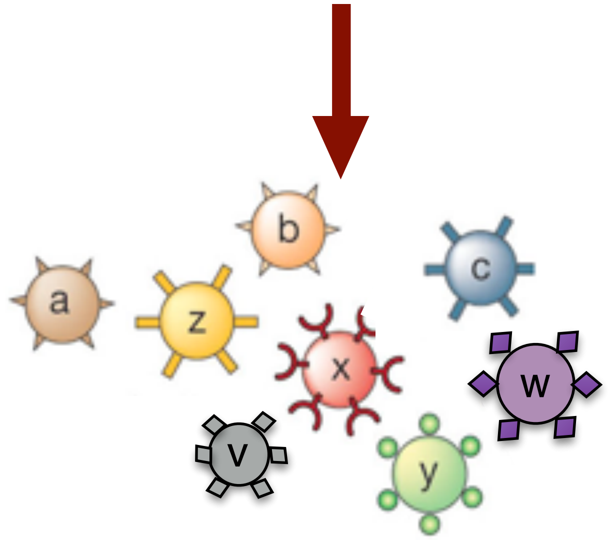
**RECEPTOR GENERATION**



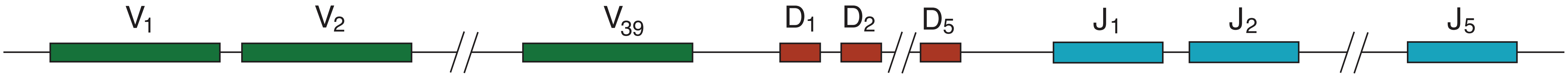
combinatorics + *randomness* → diversity

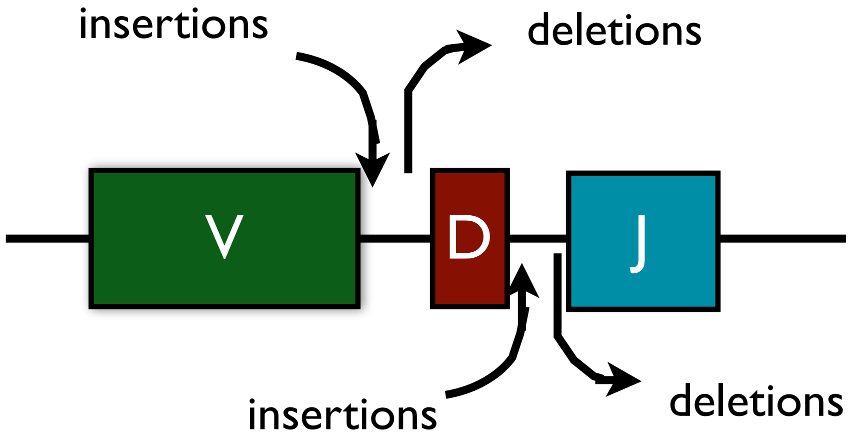
# Repertoire evolution

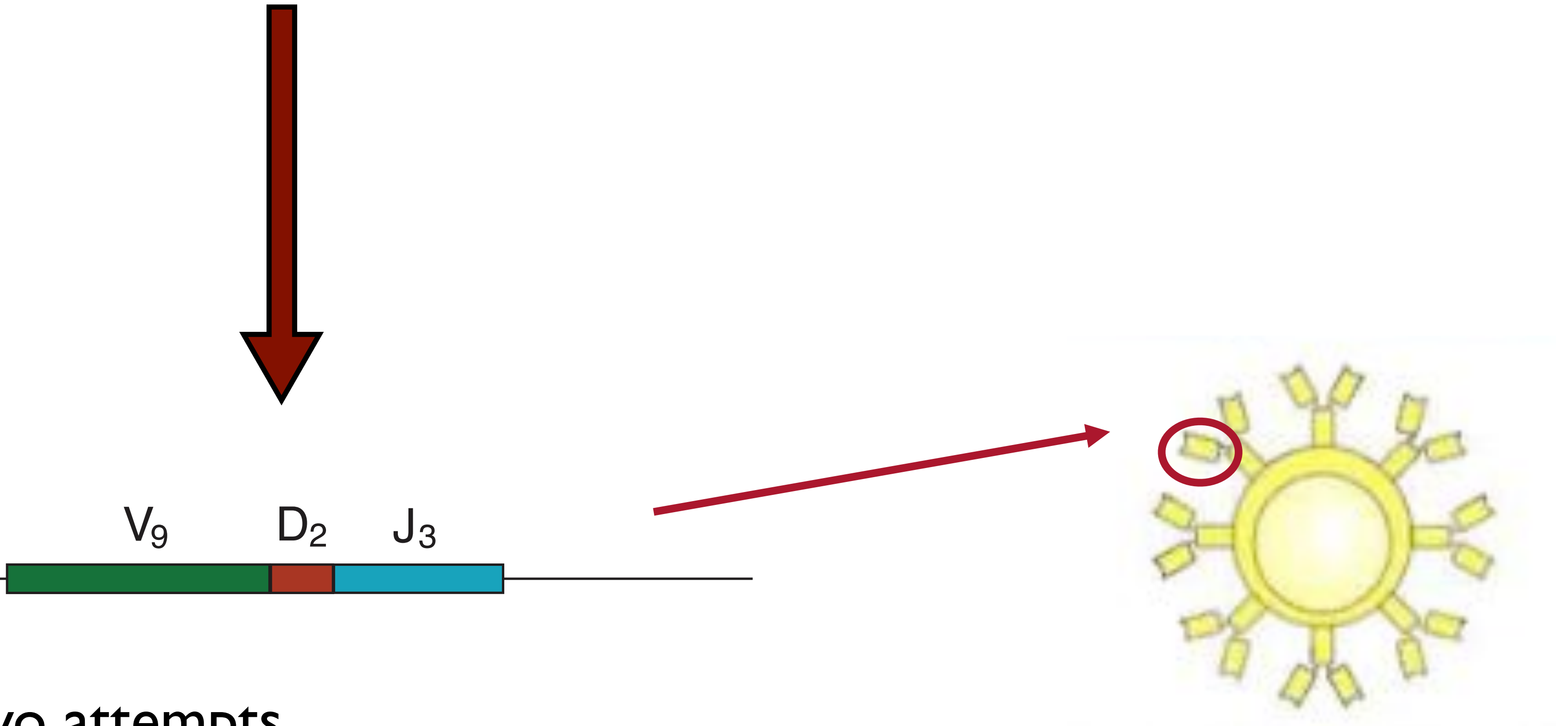
**RECEPTOR GENERATION**



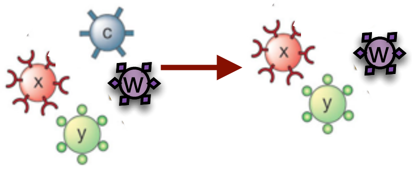
combinatorics + *randomness* → diversity



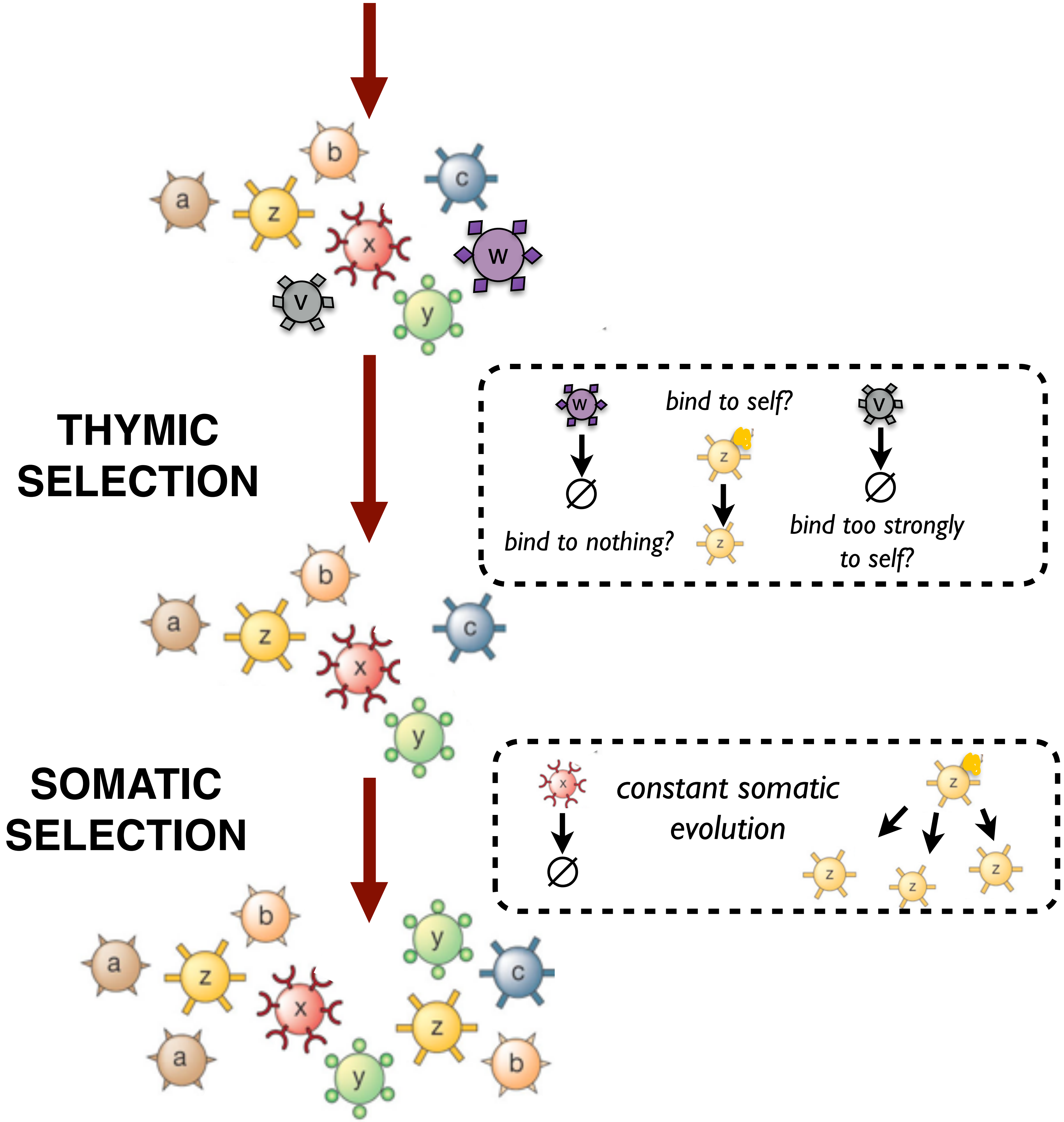




- two chromosomes per cell = two attempts



## RECEPTOR GENERATION



THYMIC SELECTION

*bind to self?*

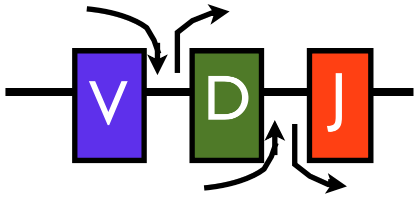
*bind to nothing?*

*bind too strongly to self?*

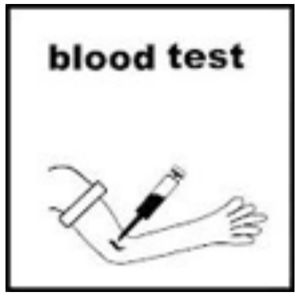
SOMATIC SELECTION

*constant somatic evolution*

# Sequence data



*new data*

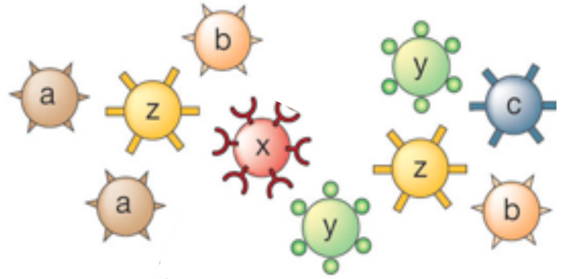


+

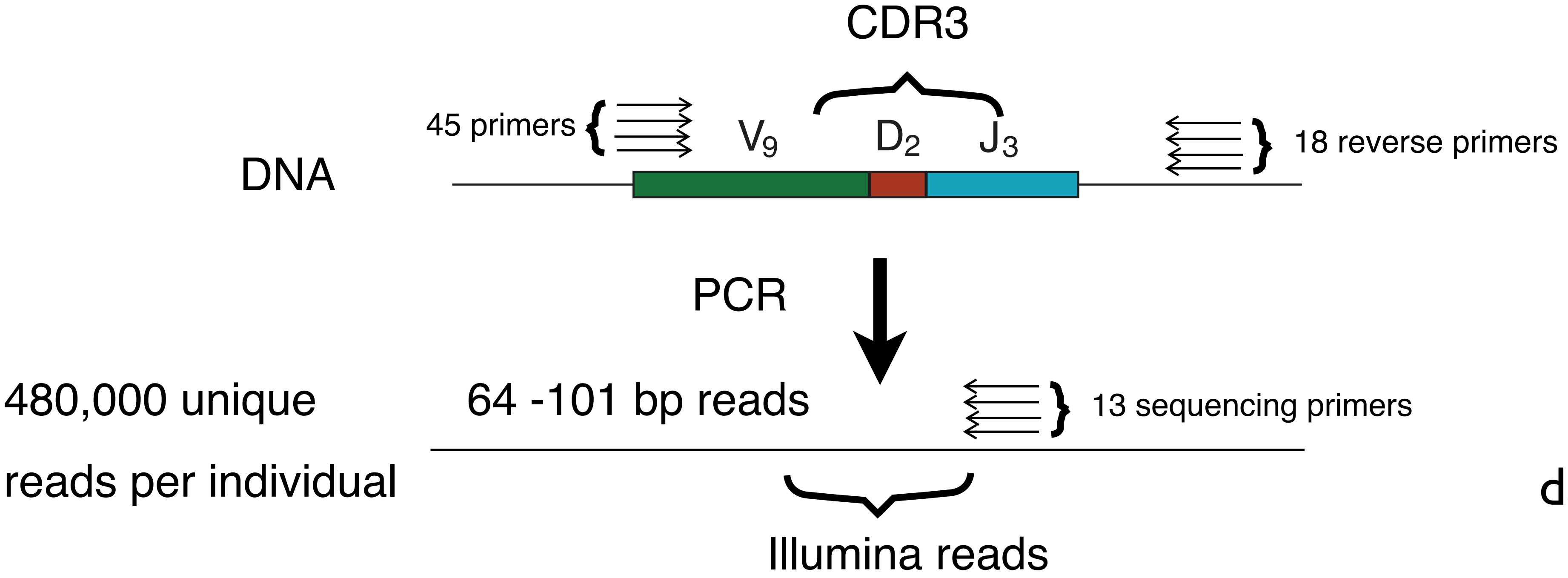
sequencing machine

=

*natural diversity* distribution



- human T-cell beta chain receptor sequences
- 9 people
- out of frame reads (~14% of each type of cells) = 35,000 unique reads ⟶ *generation*
- in frame reads  (~235,000 unique reads) ⟶ *selection*



CDR3

45 primers

$V_9$ $D_2$ $J_3$

18 reverse primers

DNA

PCR

480,000 unique

reads per individual

64 -101 bp reads

13 sequencing primers

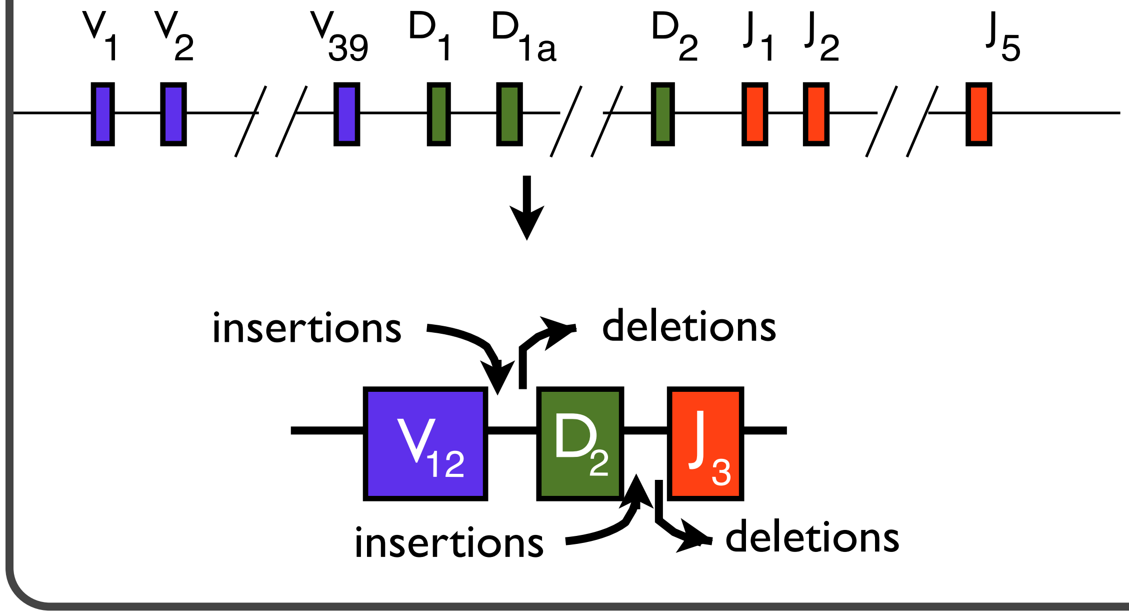
Illumina reads

data from Robins lab and
Chudakov lab

# Probabilistic VDJ recombination annotation

sequence generation:

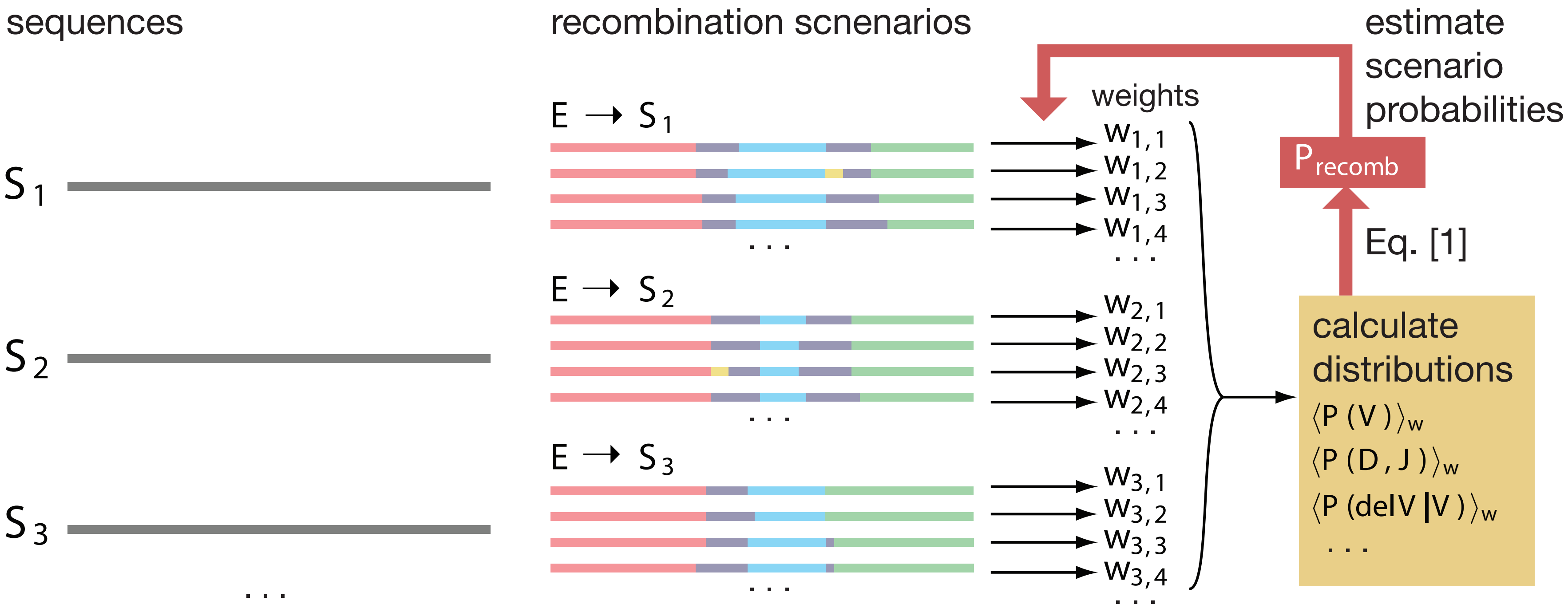
combinatorics + *randomness* → diversity



$\vec{\sigma}$ - receptor DNA sequence

```
TAGGACCTCGGAAACCTCTTCTGCTGGCGCAGAGATACAGGGCTGAGTCTTCTT
TATCTGCCGATGTCGCGAAGGCCCTCCCGCTAAGATCACTGGTGGCACAGAAGT
TAGGGAGAGGTGCCTAACTGCTGGCACAGAAGTACAGAGAGGTCTGGTTGGGGT
TCCGCCGCTAGTCCCTGAAACTACTGGCACAGAGATAGAAAGCTGTCGGGTTCT
CGAAACTGCTGGCACAGAAGTACACAGATGTTTGGGAGGGAGCAGCCGACTCCA
TCTTGGCCGCTAGTCCGAGAAACTGCTGGCACAGAAGTACACAGATGTTTGGGA
TTGTAGGAGCCGGACCGGCCCCCTGTCCCCTTGGCTGCTGGCGCAGAGATACAG
TCATTTAACGTGCGGCCCGCCTGGCACAGAAGTAAAGAGCTGTCTGGTTGTGGT
TAGTAACTCCGCTTCACTGCTGGCACAGAAGTACACAGATGTCTGGGAGGGAGC
TCCCTCCGGTTTGAAGGTCTGCTGGCACAGAAGTACACAGATGTTTGGGAGGG
CCGGTGTTCGCACAGCCCTGGGGACCCTGGCGCAAACCCCGCTTCCCTCGAGGA
```

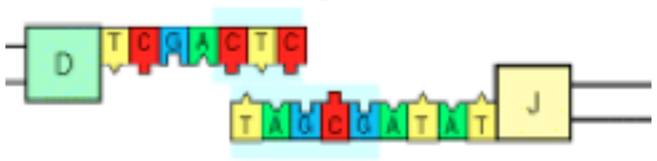
- probabilistic assignment of:



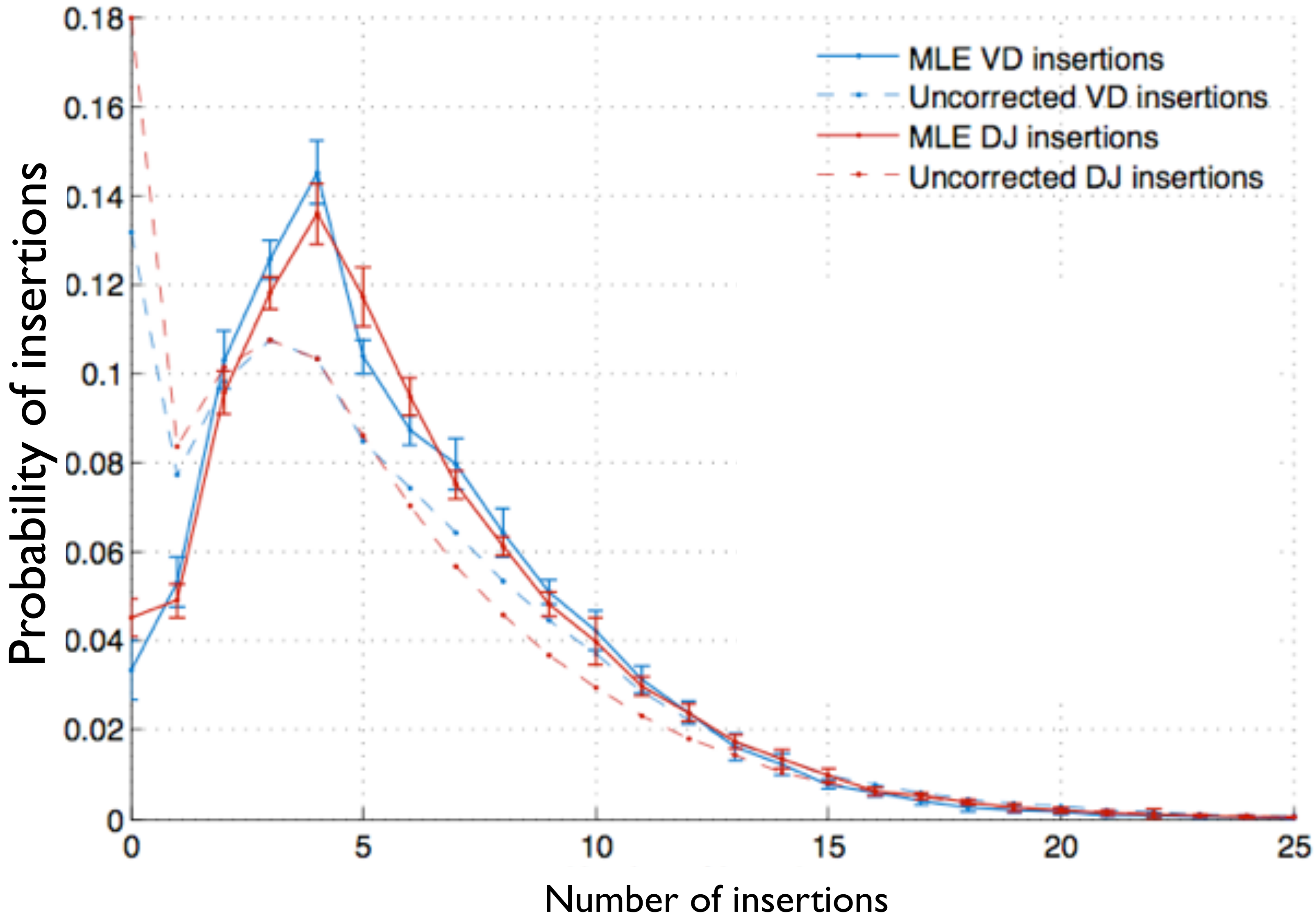
$$P^{\mathrm{recomb}}(\mathrm{scenario}) = P(V)P(D,J)P(\mathrm{deletions}V|V)P(\mathrm{insertions}DJ)... \qquad [1]$$

# Universal insertion profiles

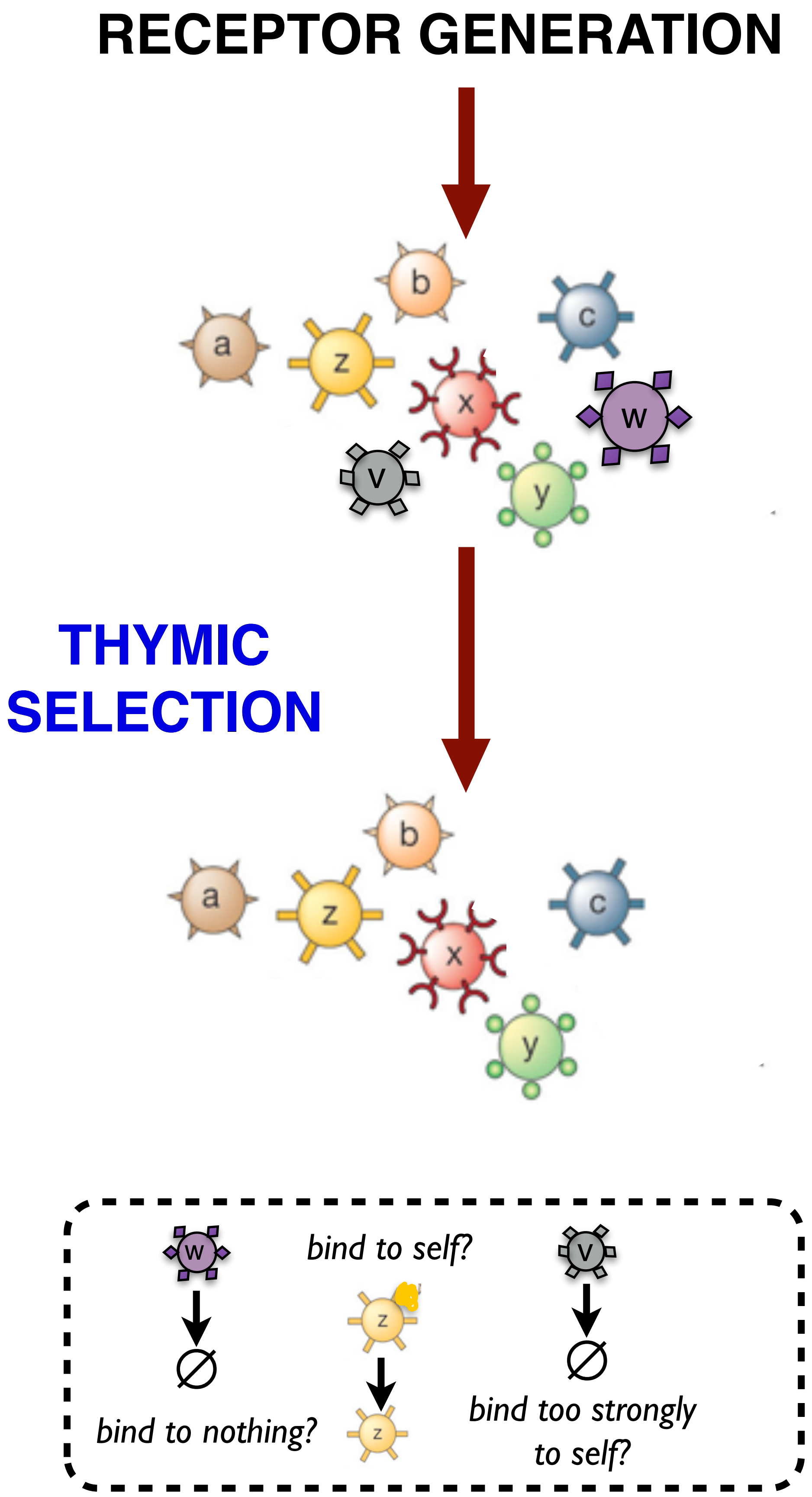
→ universal mechanism for generation



- VD and DJ insertion profiles are identical



# Receptor sharing



RECEPTOR GENERATION

THYMIC SELECTION

bind to nothing?

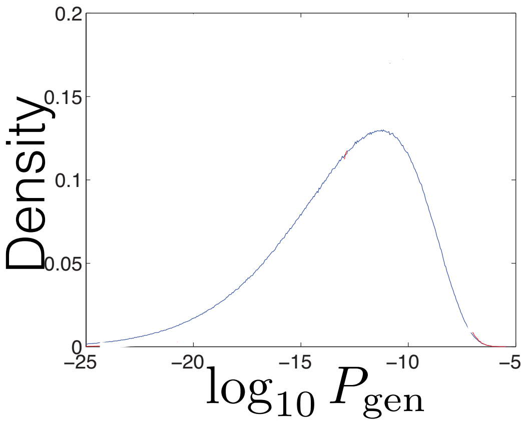
bind to self?

bind too strongly to self?

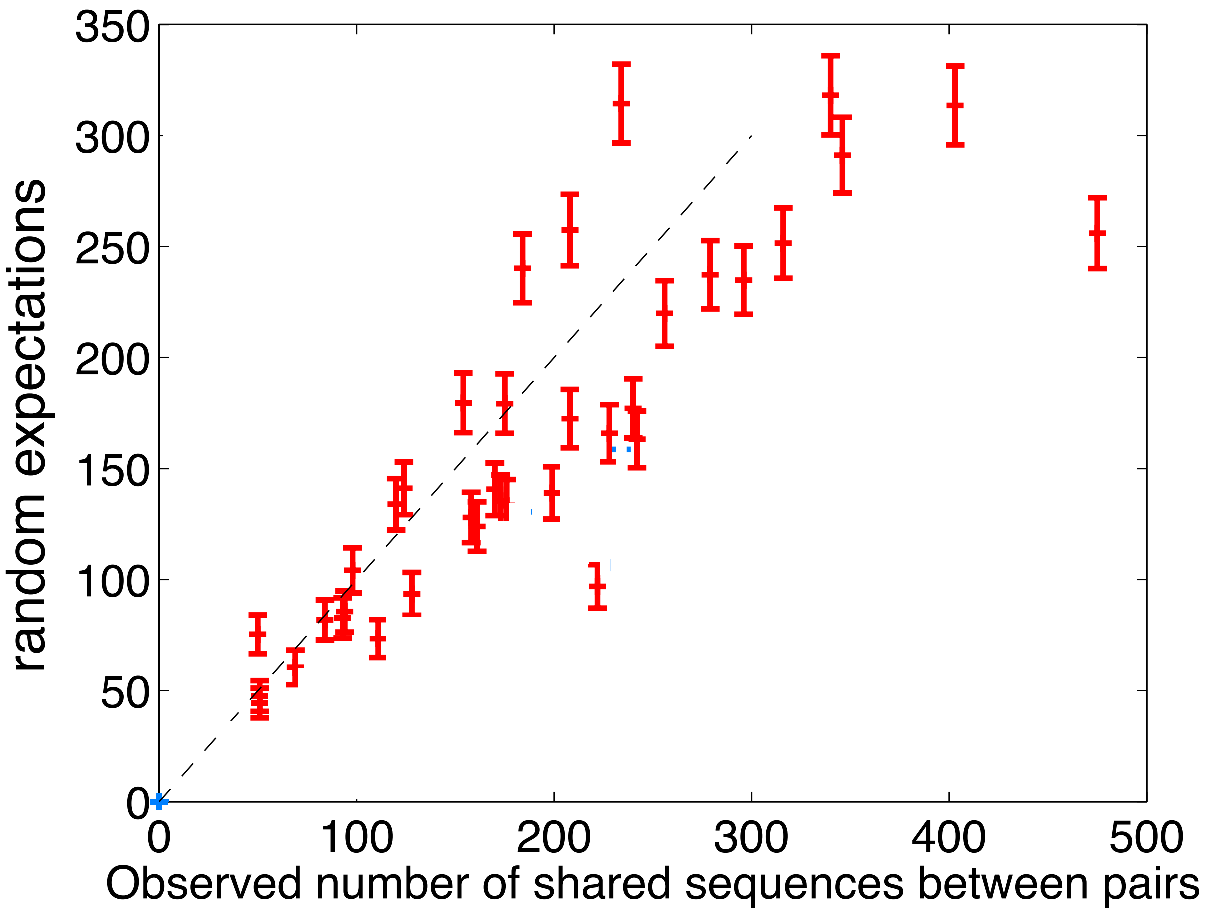
- quantify using selection factors

$$Q(\{\sigma\}) = \frac{P_{\mathrm{post-sel}}(\{\sigma\})}{P_{\mathrm{gen}}(\{\sigma\})}$$

human beta TCR data from Robins lab



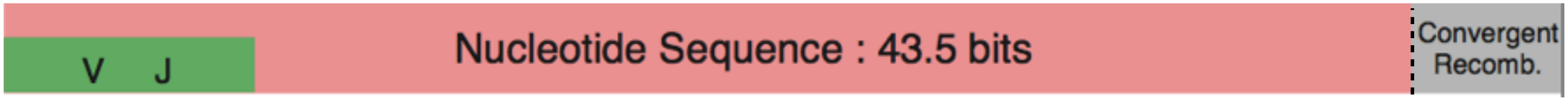
- how many shared receptors between 2 people?



→ close to random expectations

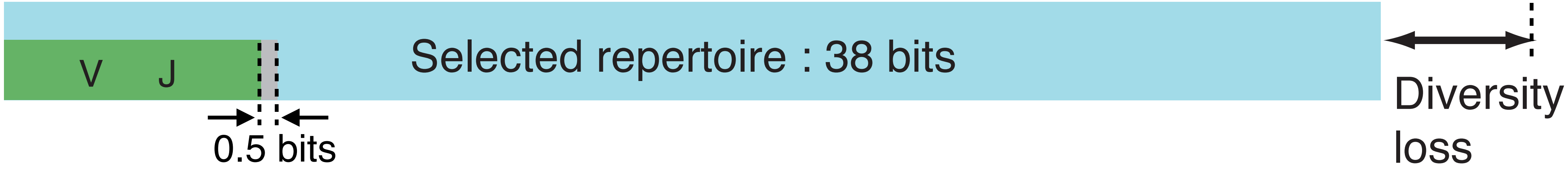
# Entropy of distribution

- entropy of generated repertoire



$\Longrightarrow$ repertoire size $10^{13}$ sequences

- entropy of post-thymic selection repertoire



$\Longrightarrow$ repertoire size $10^{11}$ sequences

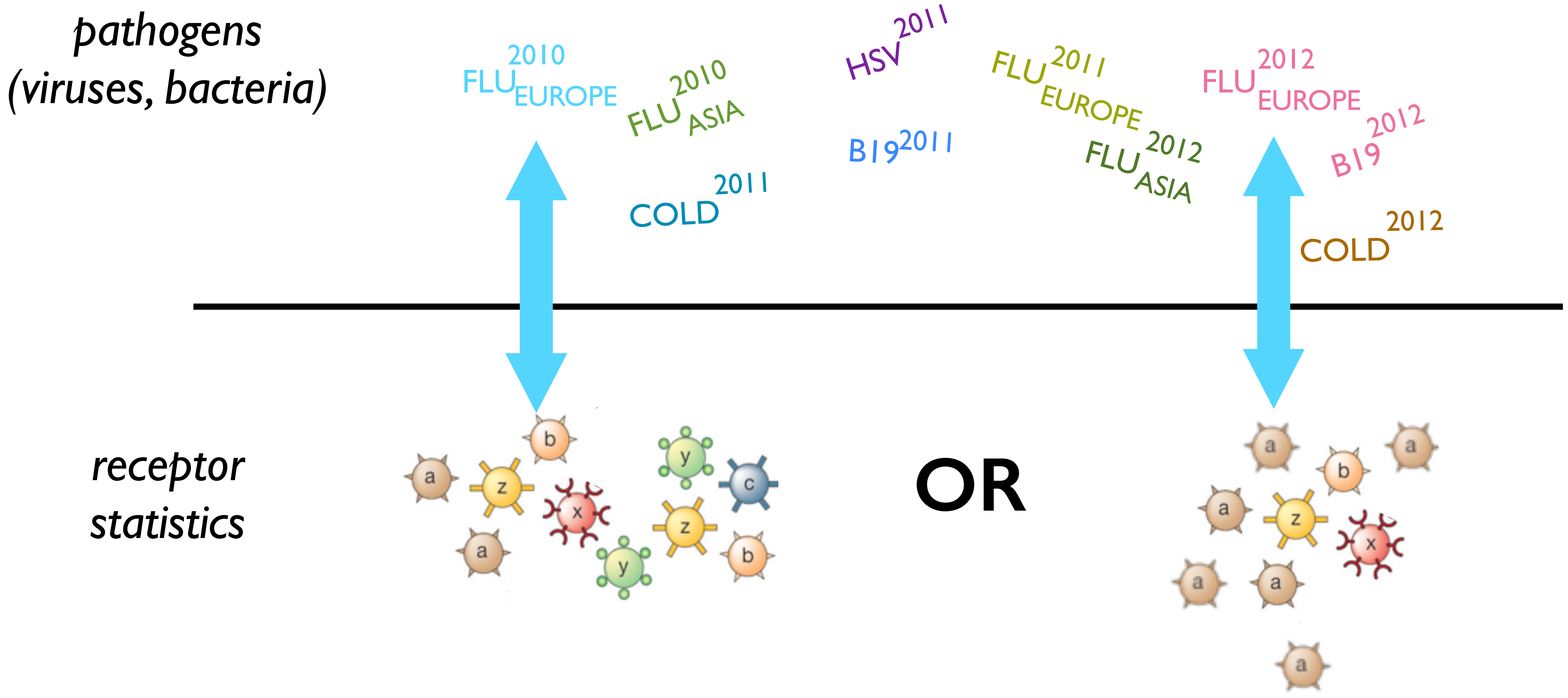
$\longrightarrow$ thymic selection gives 50-fold reduction in diversity

- thymic selection keeps ~15% of sequences but only 2% of diversity

$\longrightarrow$ thymic selection gets rid of rare clones

selection favours clones that are likely to be generated

# Receptor distributions



*pathogens (viruses, bacteria)*

FLU$_{\text{EUROPE}}^{2010}$

FLU$_{\text{ASIA}}^{2010}$

COLD$^{2011}$

HSV$^{2011}$

B19$^{2011}$

FLU$_{\text{EUROPE}}^{2011}$

FLU$_{\text{ASIA}}^{2012}$

FLU$_{\text{EUROPE}}^{2012}$

B19$^{2012}$

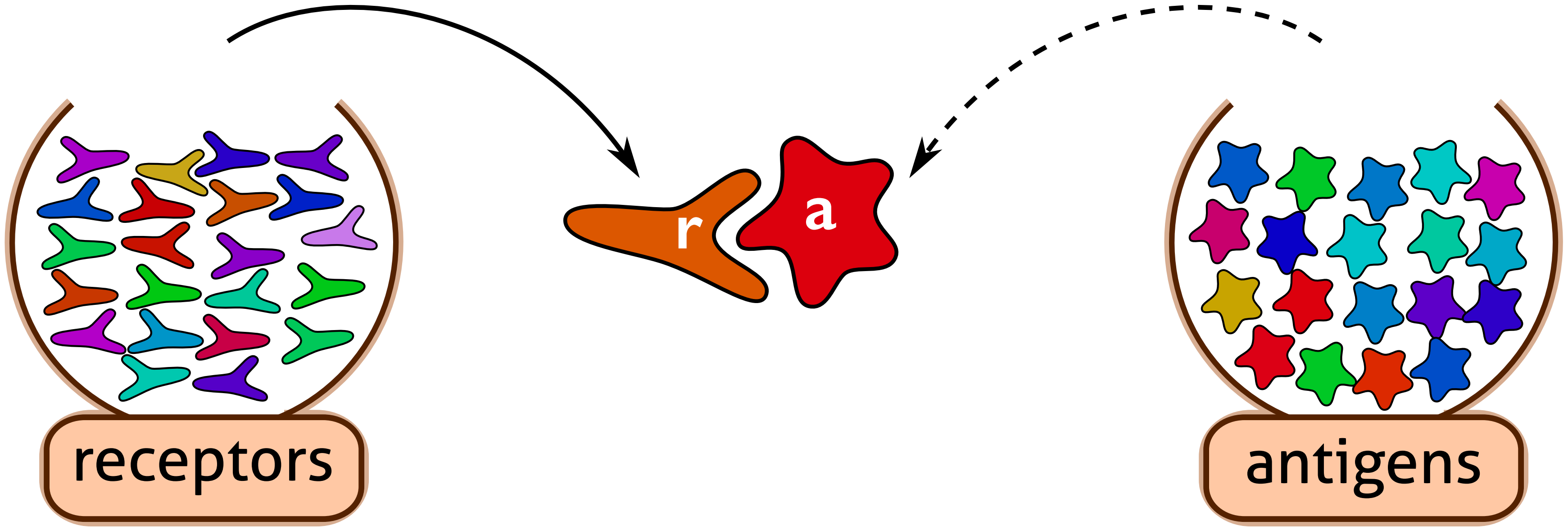
COLD$^{2012}$

*receptor statistics*

OR

optimal distribution ?

# The trade-off

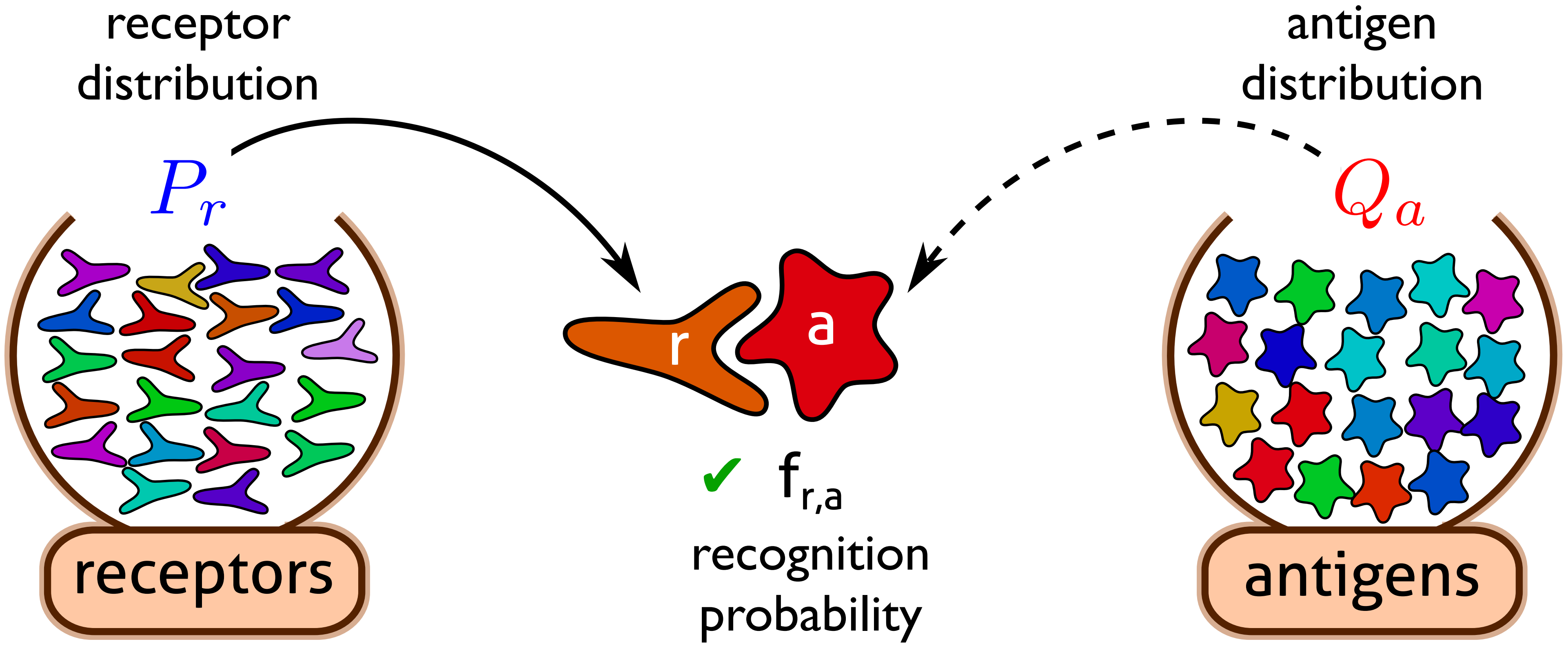


limited number of encounters

**How should immune receptors be distributed to minimize harm from infections?**

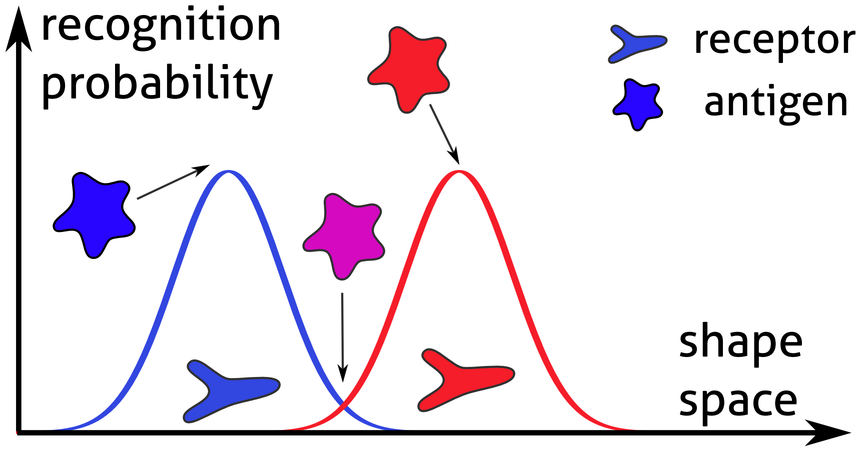
lymphocyte repertoire ↔ antigenic environment

# Cross-reactivity

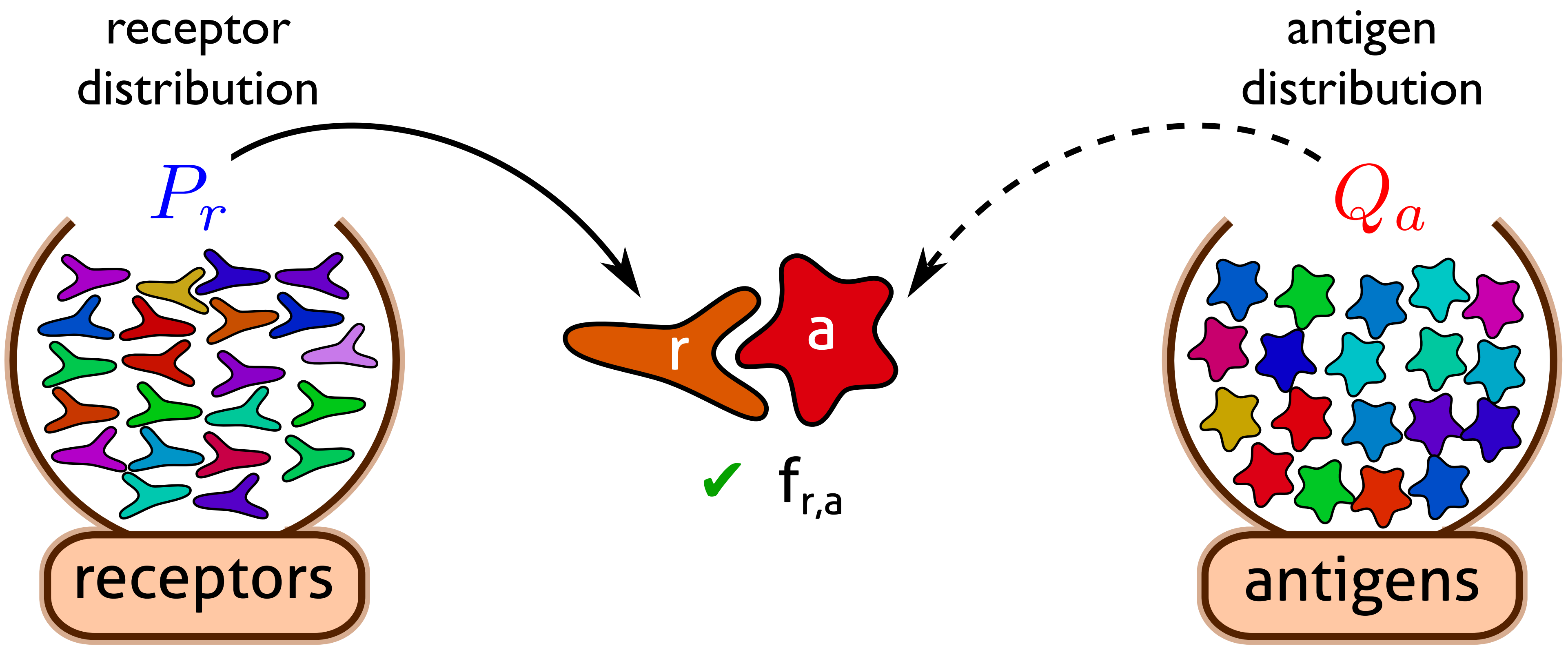


- cross-reactivity - recognition probability
- probability of immune response from encounter with a given antigen

$$\tilde{P}_a = \sum_r f_{r,a} P_r$$



# Receptors - antigens interactions



receptor distribution

antigen distribution

$P_r$

$Q_a$

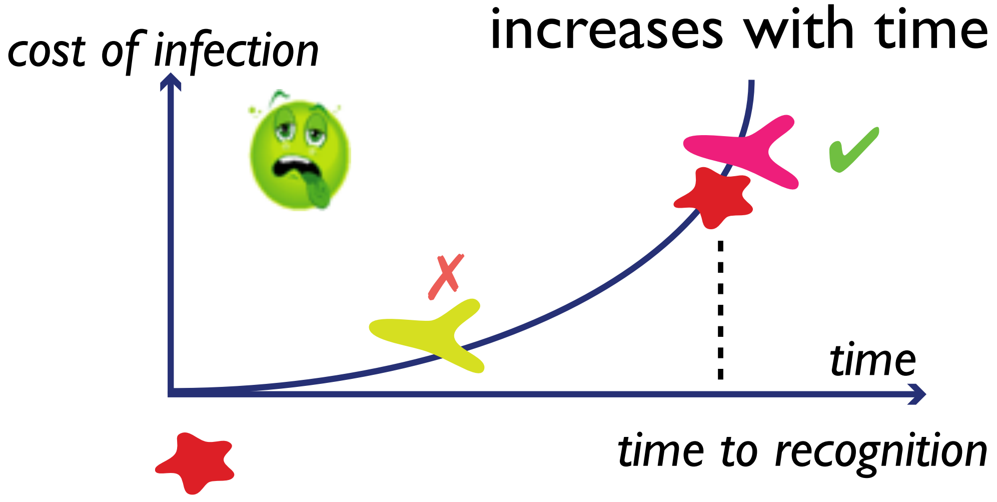
✔ $f_{r,a}$

receptors

antigens

- probability of immune response from encounter with a given antigen $\tilde{P}_a = \sum_r f_{r,a} P_r$
- time measured in mean number of encounters $m$

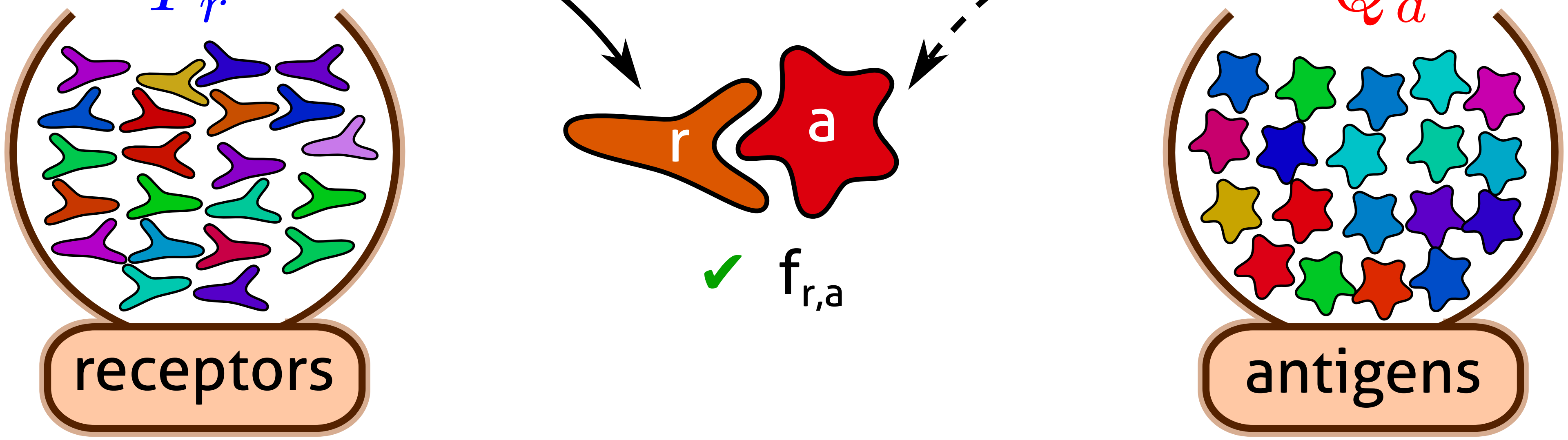
*harm* caused by a given antigen increases with time



$$\bar{F}_a(P_r) = \overset{\text{virulence}}{\mu_a} \int_0^{+\infty} dm\, \overset{\text{effective cost of infection}}{F_a(m)}\, \underset{\text{Poisson distributed recognition}}{\tilde{P}_a e^{-m\tilde{P}_a}}$$

$$\mathrm{Cost}(\{P_r\}) = \sum_a Q_a \bar{F}_a(P_r)$$

# Cost



receptor distribution

$P_r$

antigen distribution

$Q_a$

✓ $f_{r,a}$

receptors

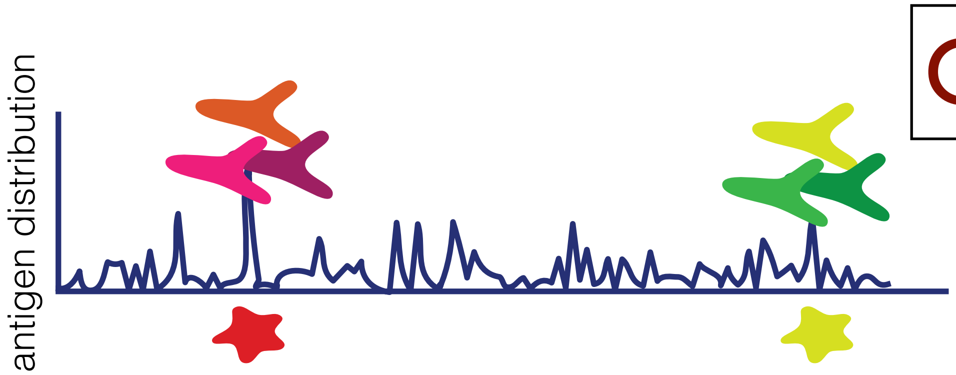
antigens

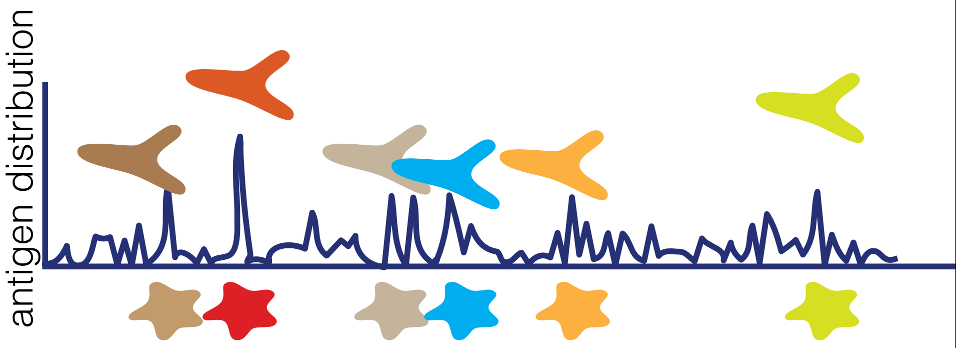
total harm caused by antigen increases with time

$$\mathrm{Cost}(\{P_r\}) = \sum_a Q_a \bar{F}_a(P_r)$$

trade-off: many antigens ↔ limited resources

**Optimal repertoire?**





→ minimize cost given fixed antigen distribution

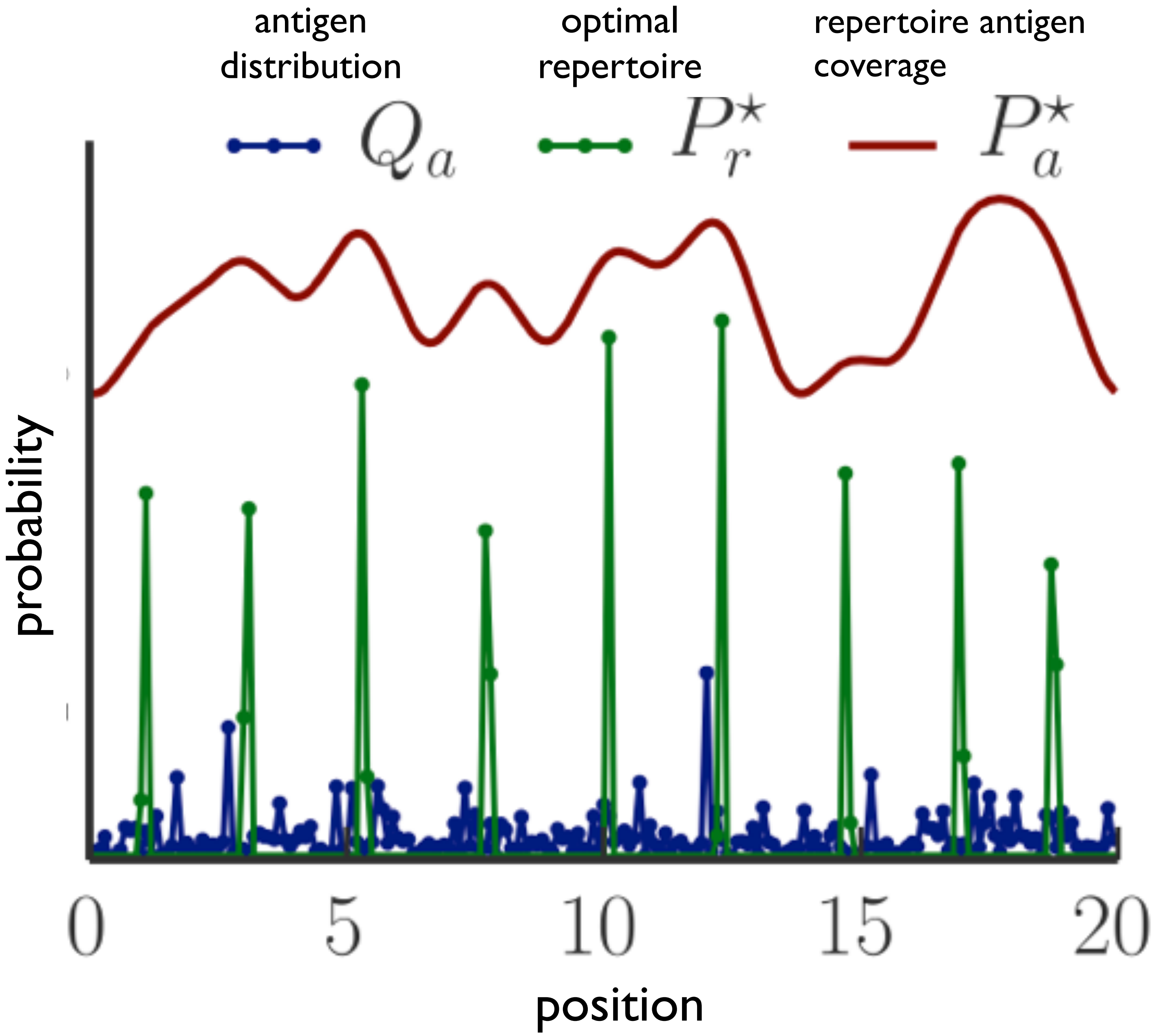
# Peaked optimal repertoires

- exponentially expanding antigen population
  + exponentially growing cost in time

$$F(m) = m$$

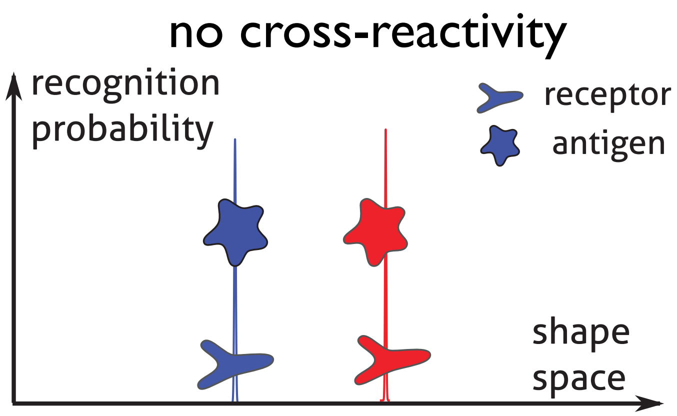
- peaked distributions
- tile space

- coverage follows antigen distribution
- but not exactly



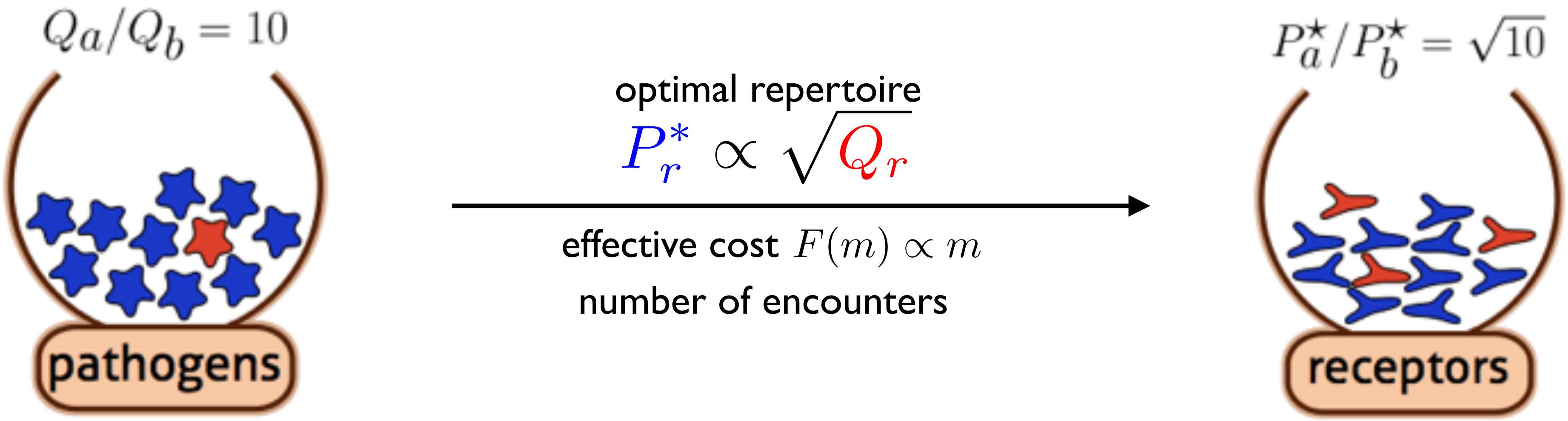
# Covering rare pathogens

How many resources aimed at common/rare antigen?



depends on cost of late recognition ⟶ effective cost function

- exponentially expanding antigen population



optimal repertoire

$$P_r^* \propto \sqrt{Q_r}$$

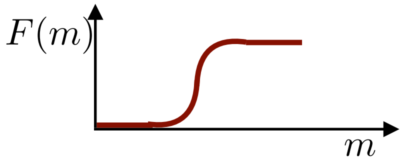
effective cost $F(m) \propto m$

number of encounters

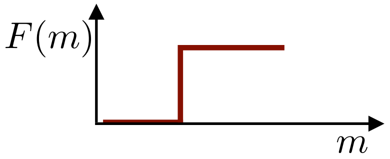
- exponentially expanding antigen population + exponentially growing cost in time

$$F(m) = m^\alpha \qquad \longrightarrow \qquad P_r^* \propto {Q_r}^{1/(1+\alpha)}$$

- saturated cost ⟶ low frequency cut-off

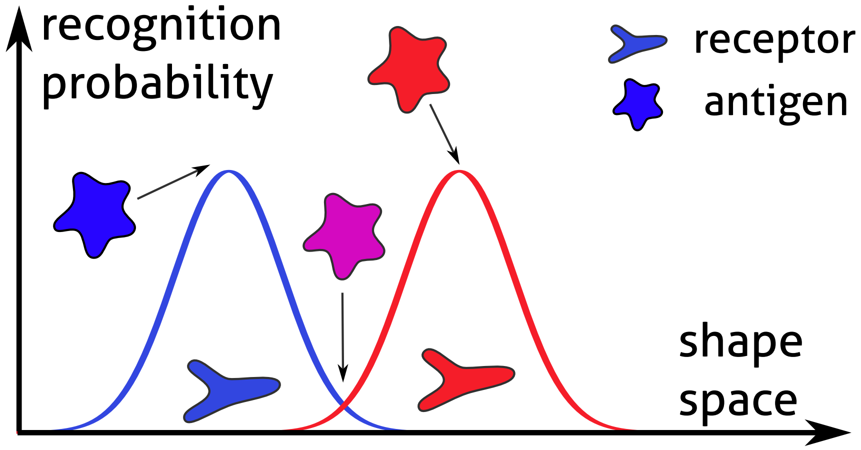


- harm past threshold ⟶ flattened receptor distribution

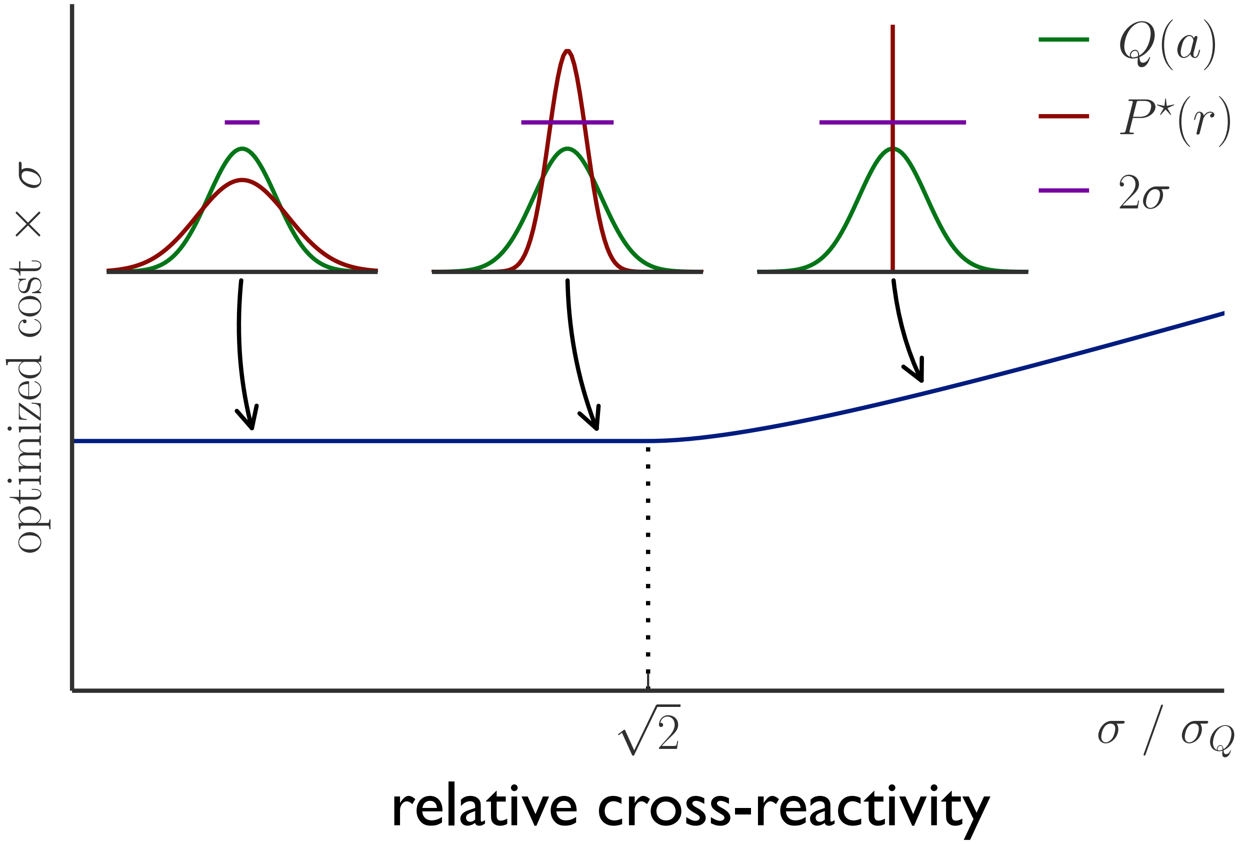


- $P_r^* \propto Q_r \Leftrightarrow$ very slowly increasing cost $F(m) \propto \ln m$

# Cross-reactivity



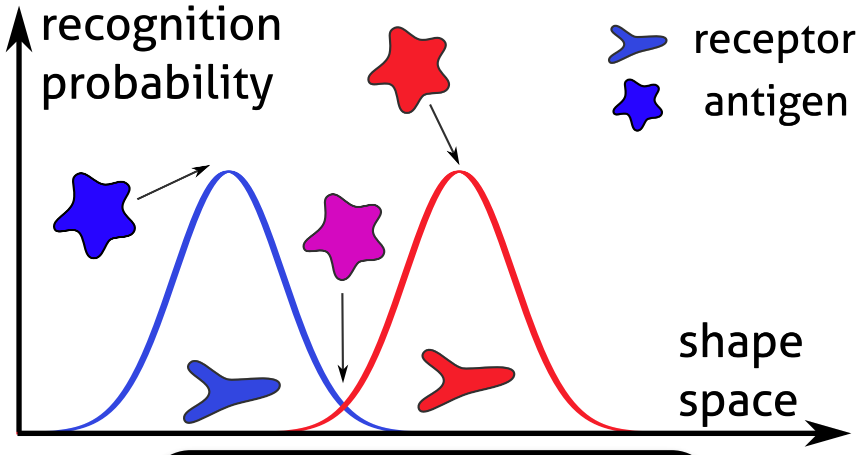
- antigen distribution + cross-reactivity Gaussian



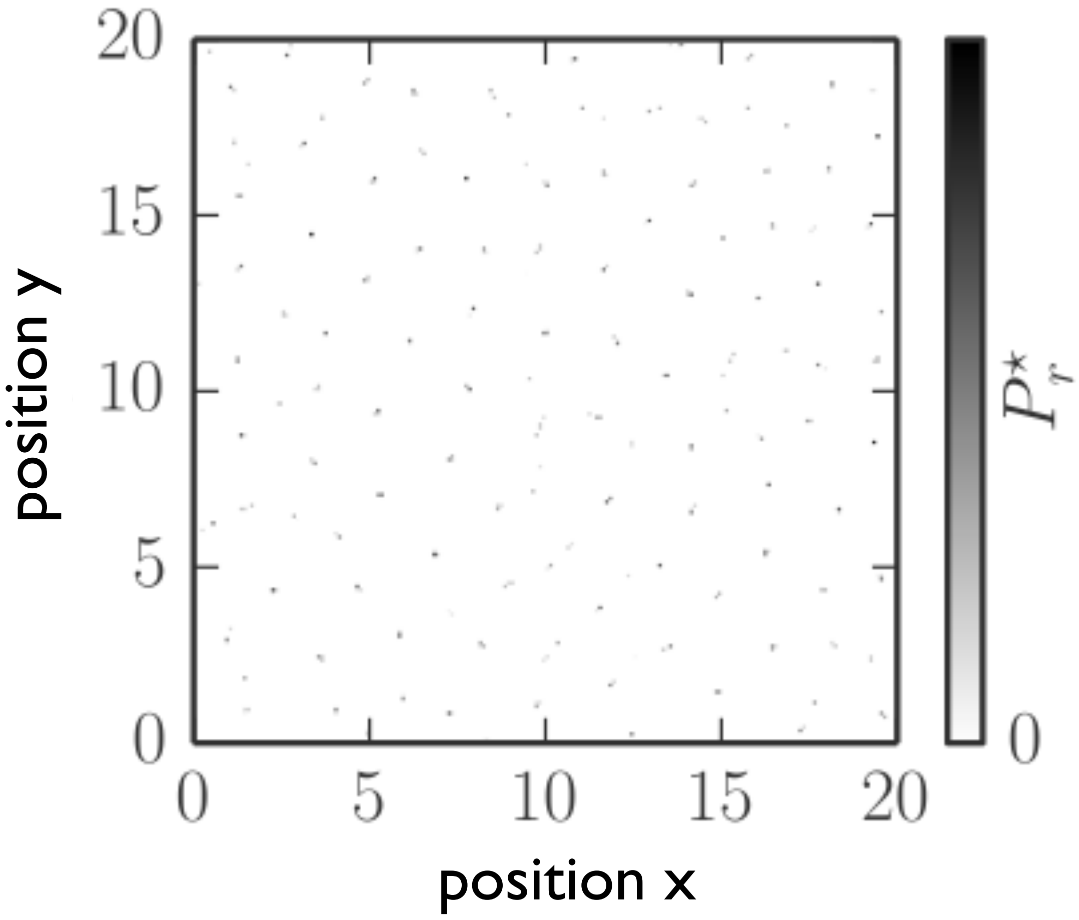
→ large cross-reactivity concentrates distribution

# Peaked optimal repertoires

rare antigens

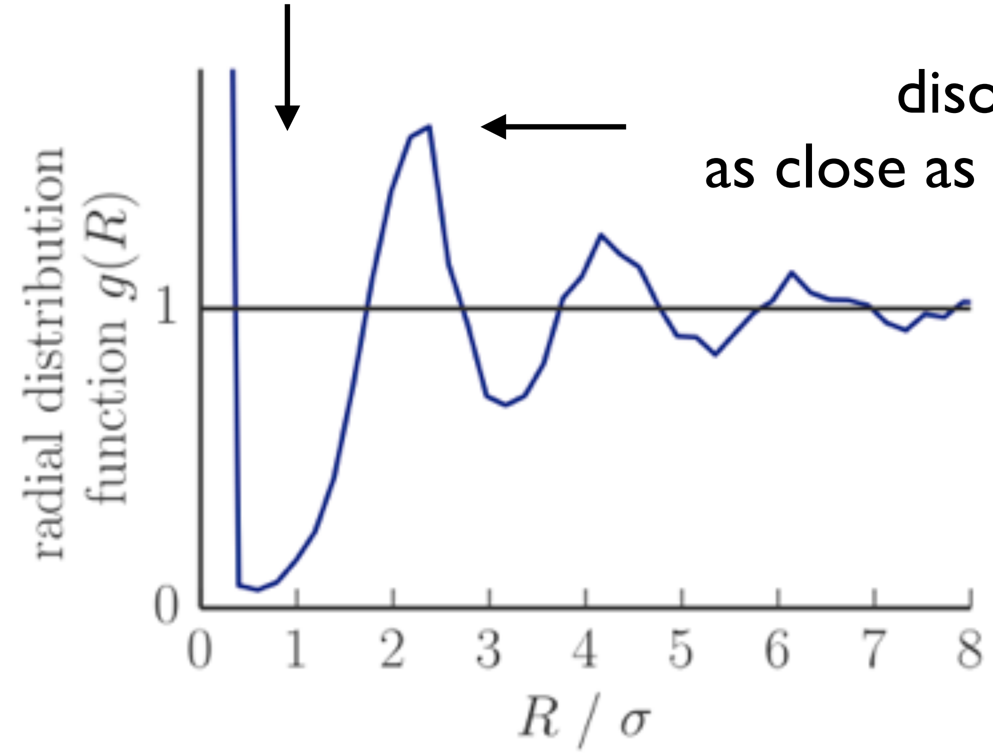


cross-reactivity

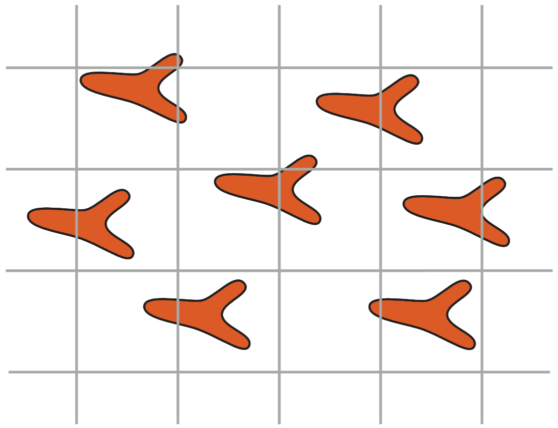


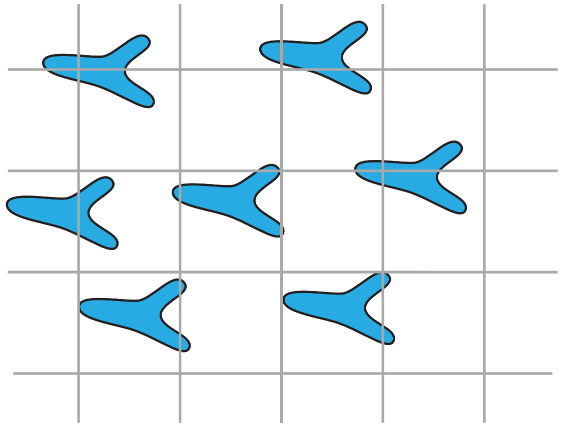
# Disordered hyperuniformity

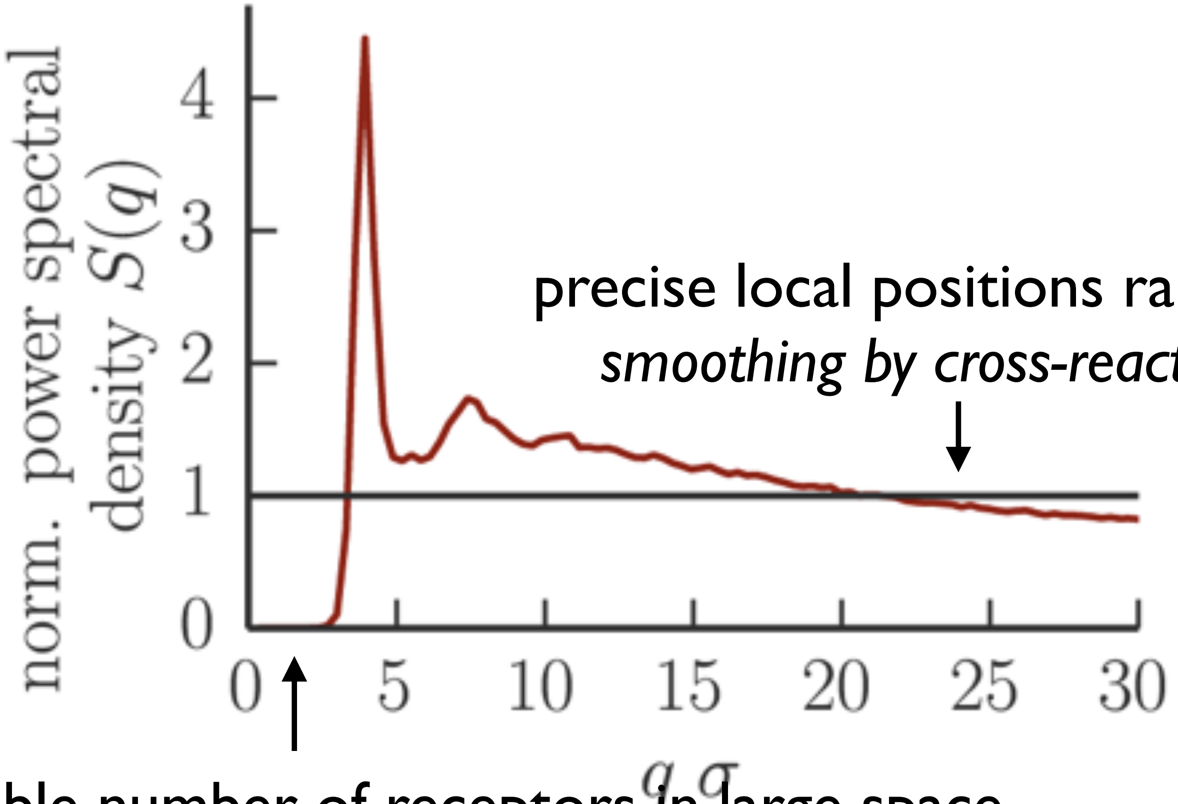
receptors cannot be close to each other



disordered tiling -
as close as possible but excluding





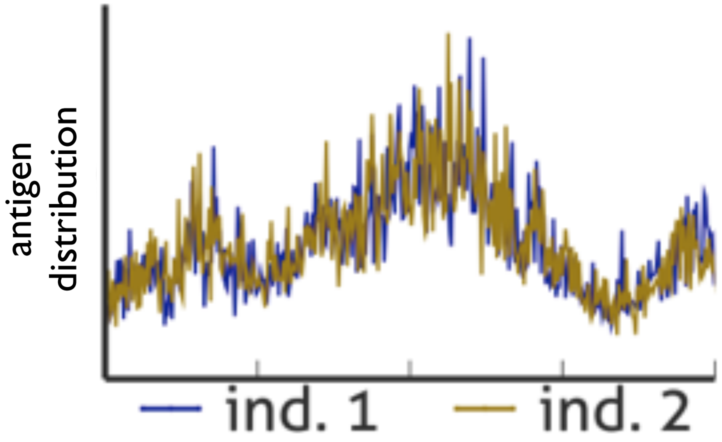


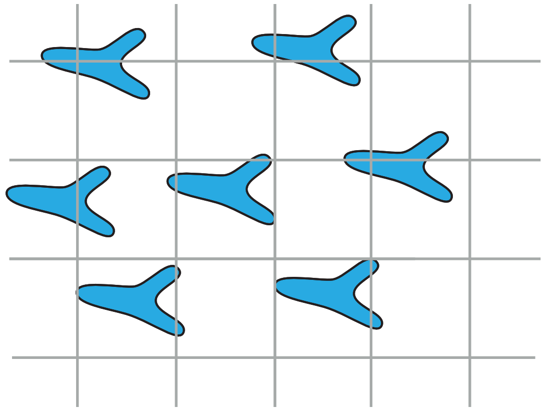
precise local positions random
*smoothing by cross-reactivity*

reproducible number of receptors in large space
*tracking of antigen*

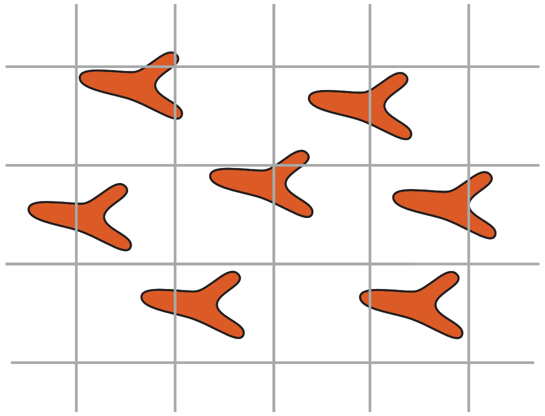
# Personalized responses

two individuals see the environment slightly differently





→ very different repertoires



# Self-organized dynamics

## Can optimal repertoires be reached via dynamics?

$$\dot{N}_r = N_r \left[ b \boxed{\sum_p Q_p f_{r,a} \boxed{A\left(\sum_{r'} N_{r'} f_{r',a}\right)}} - d \right]$$

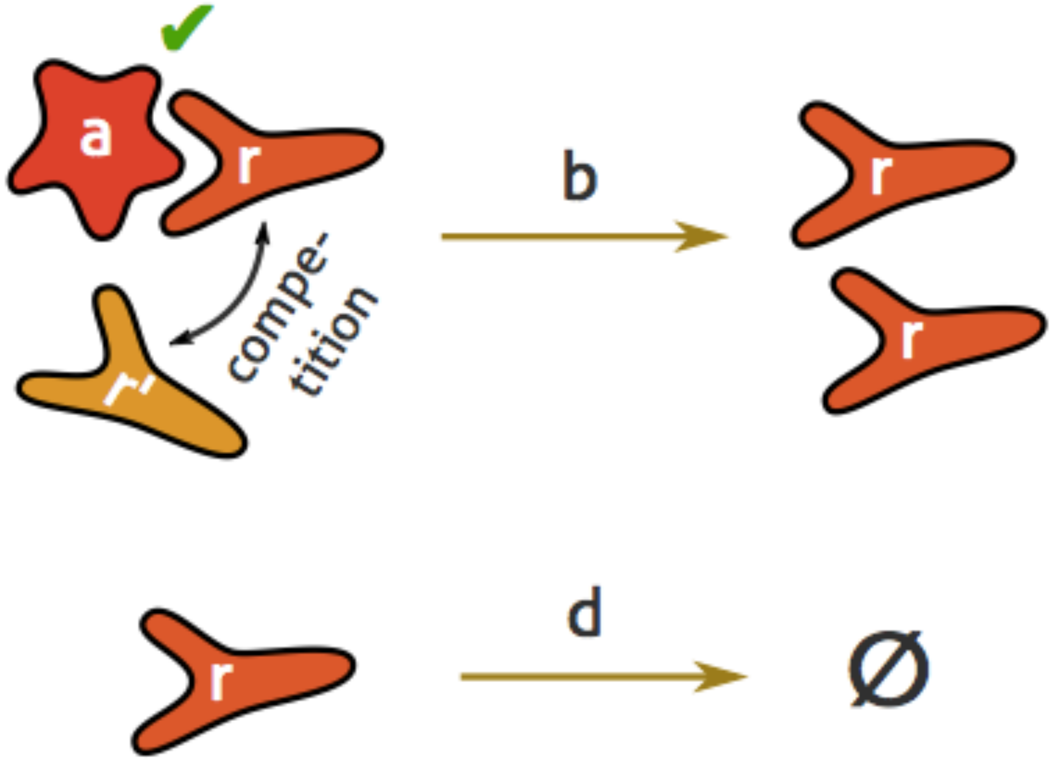
population size

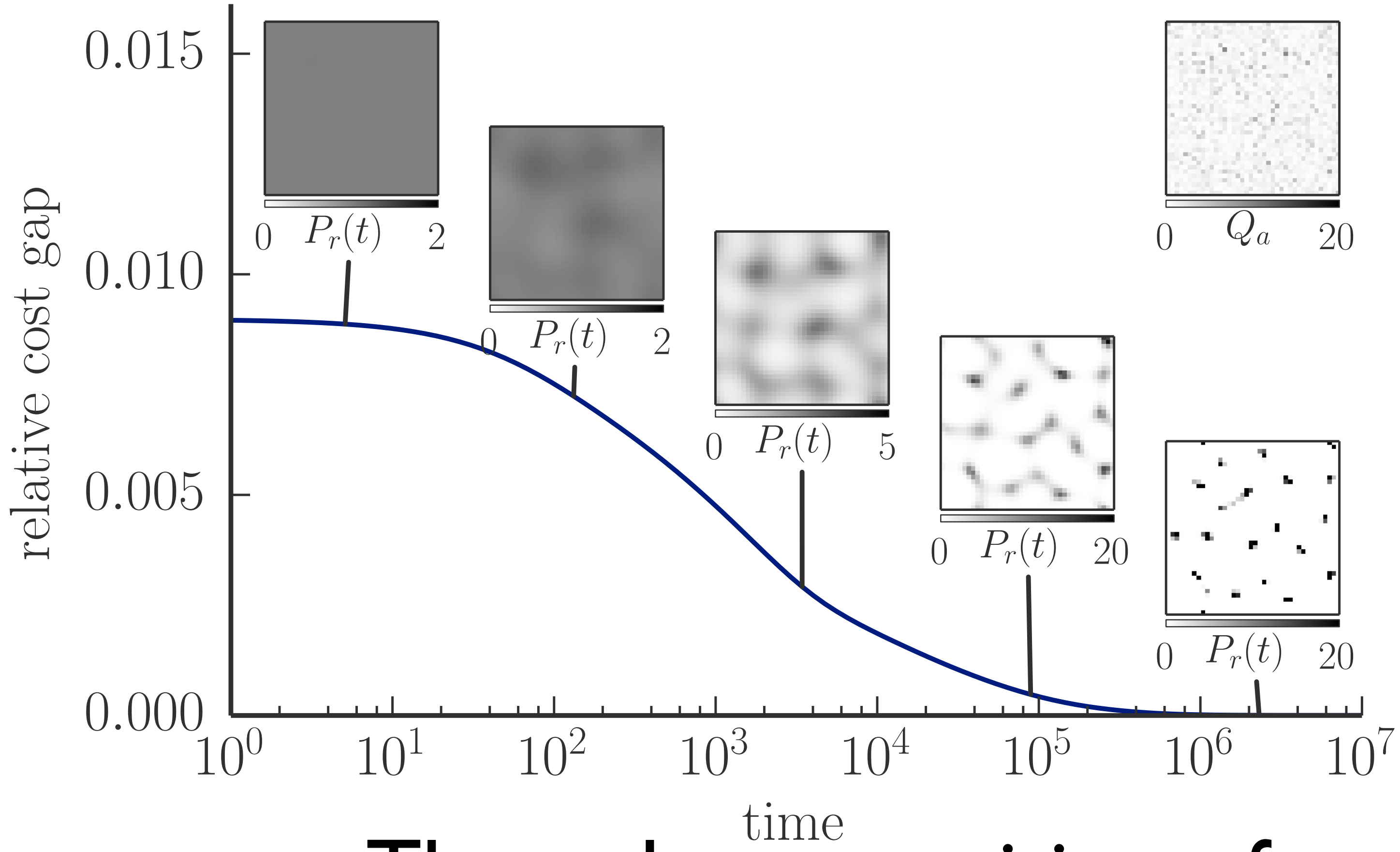
proliferation rate

detectable pathogen

availability of pathogen
→ reduced by competition
e.g. $A(\tilde{N}_a) = \frac{1}{(1+\tilde{N}_a)^2}$

death rate



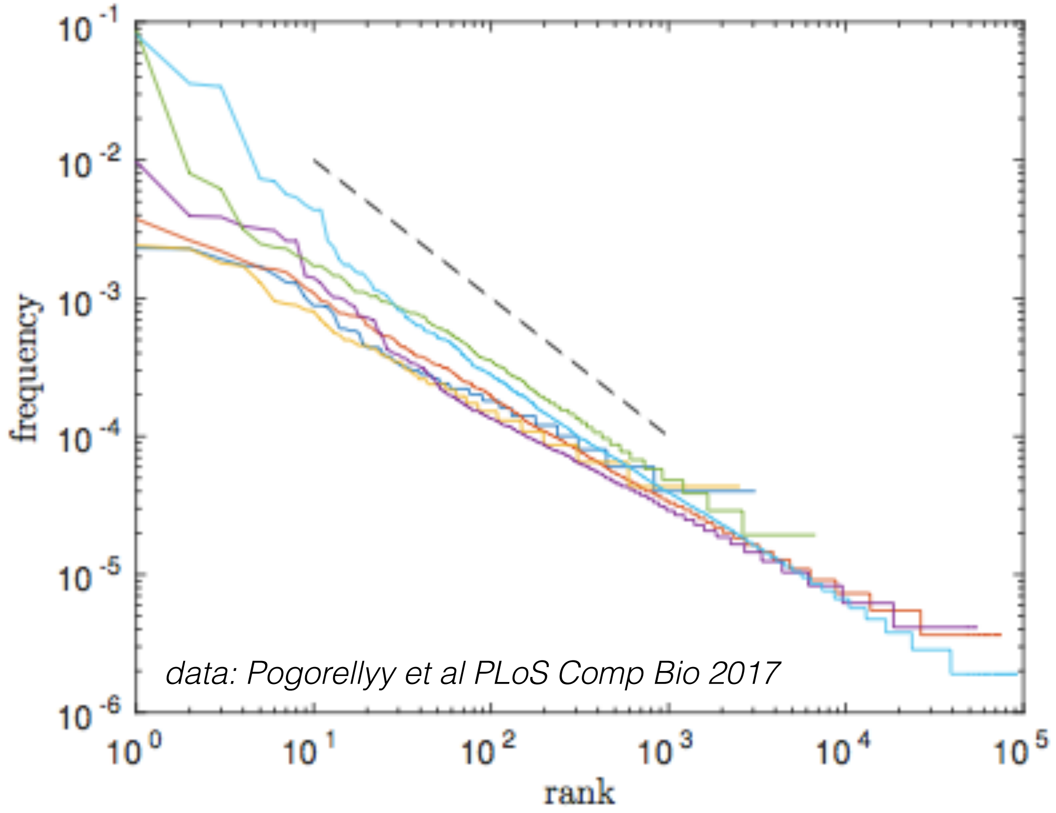


⟶ Through competition of receptors for antigen

# Estimating frequencies

- trying to infer species frequencies

*human T-cells:*

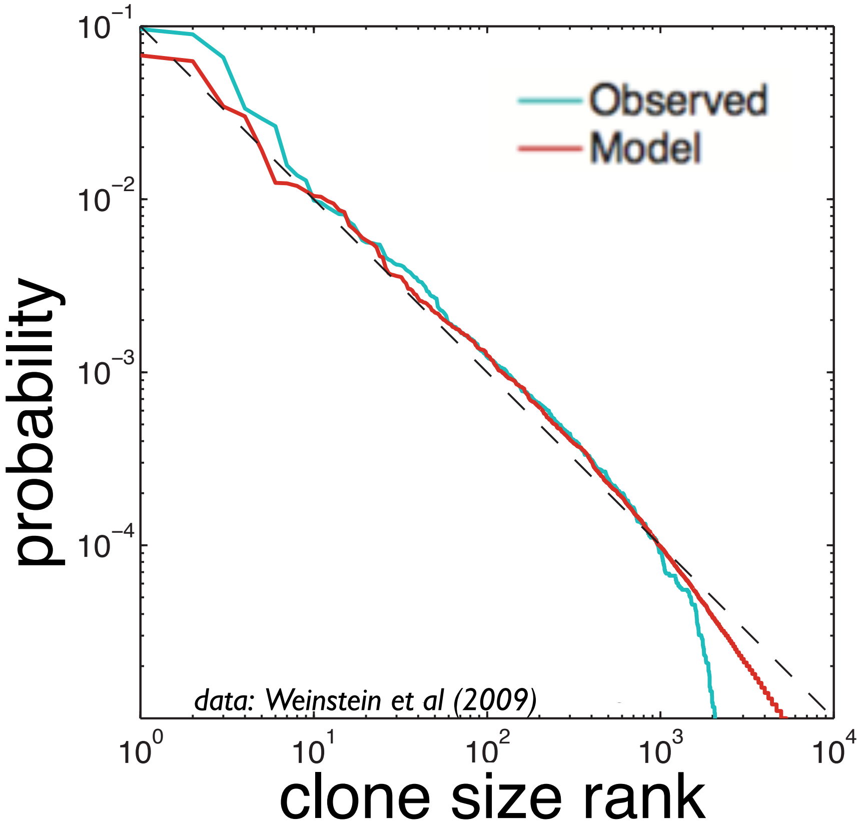


data: Pogorellyy et al PLoS Comp Bio 2017

beta chain

Post-selection: 38 bits

*zebrafish B-cells: clone size distribution*



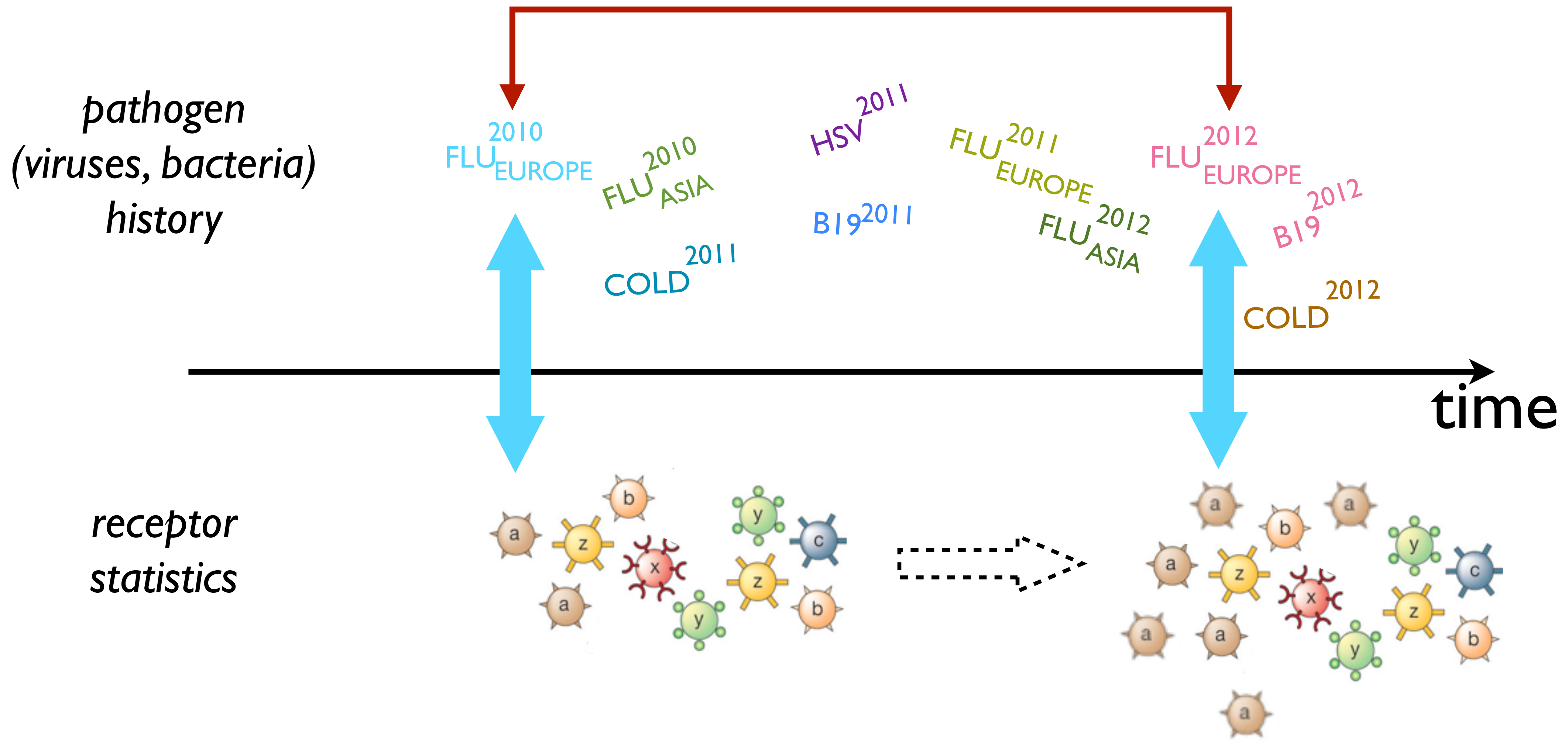
data: Weinstein et al (2009)

- also in other environments (microbiome, ponds, forests)

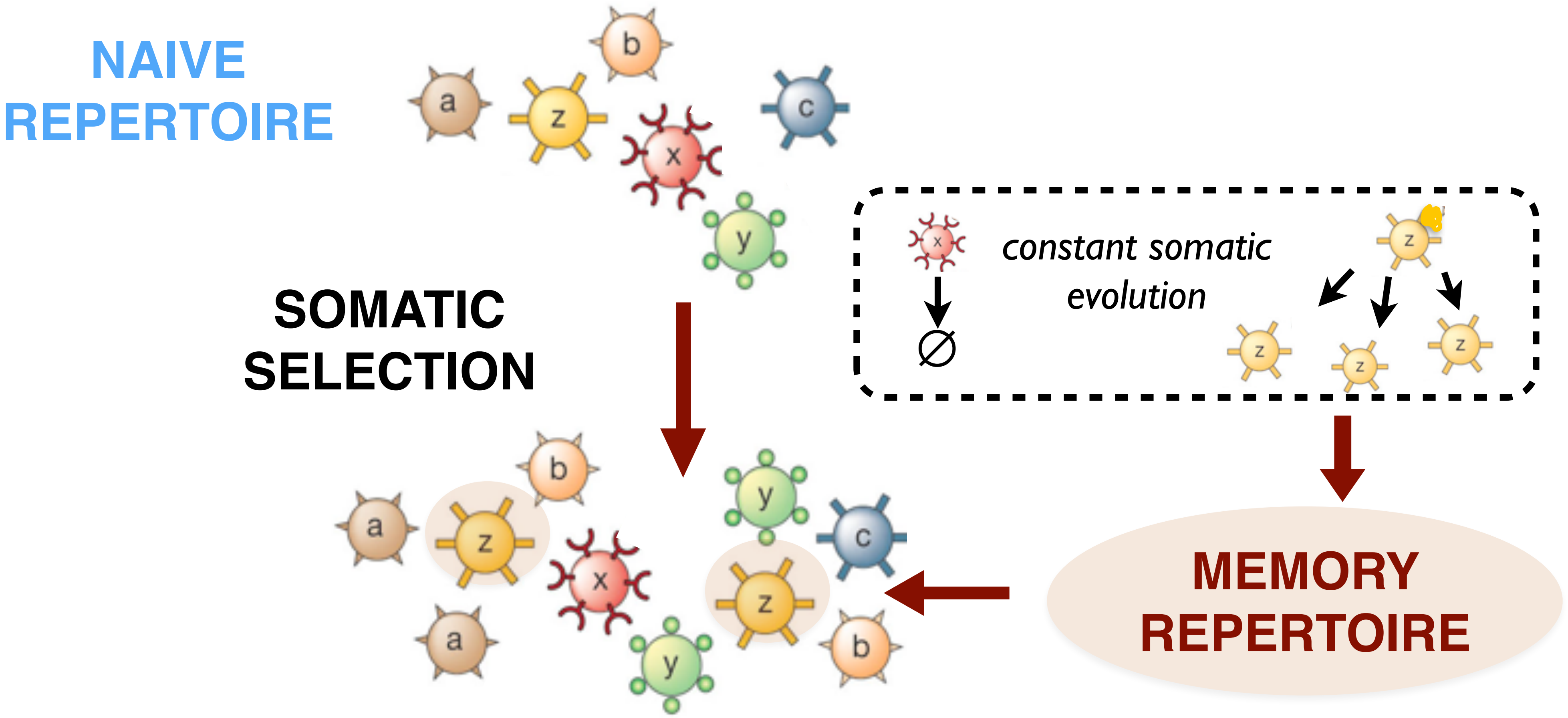
# How to update receptor frequencies?

- *adaptive* immune system - optimal predictor of future pathogens?

pathogen statistics is predictable because ***correlated in time***



# How to remember?

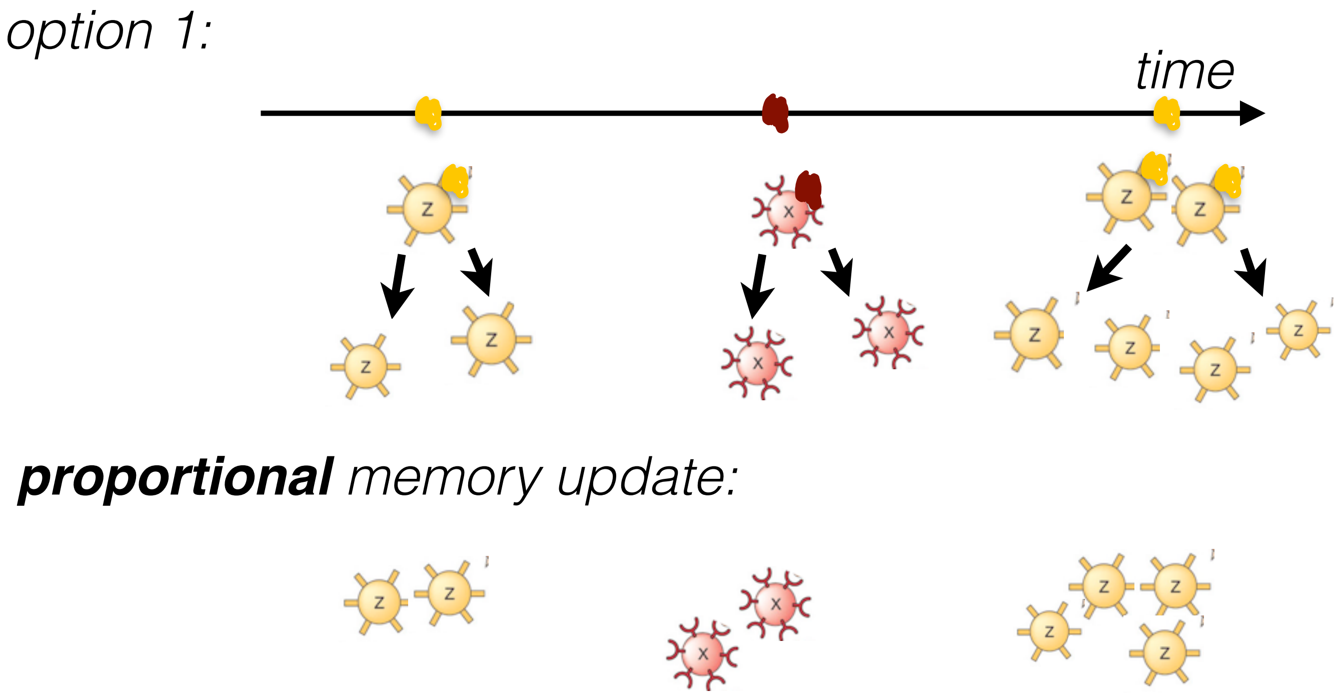

NAIVE REPERTOIRE

SOMATIC SELECTION

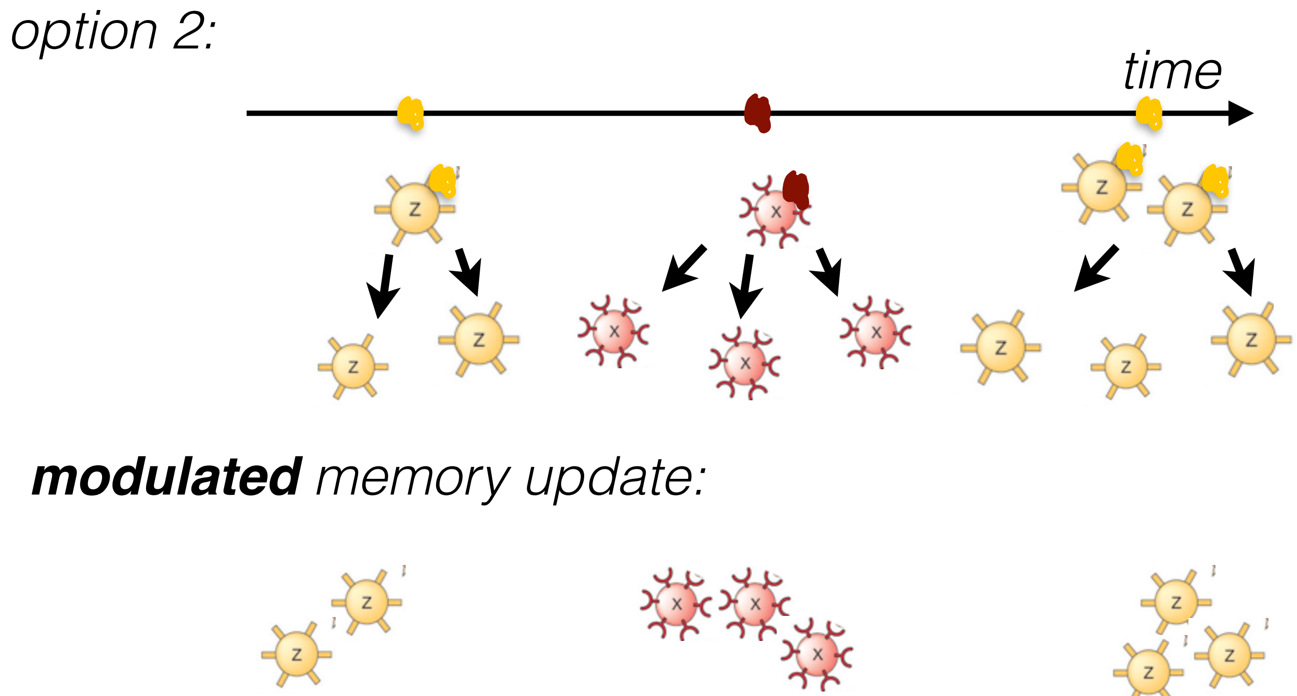
constant somatic evolution

MEMORY REPERTOIRE

is memory useful ?


option 1:

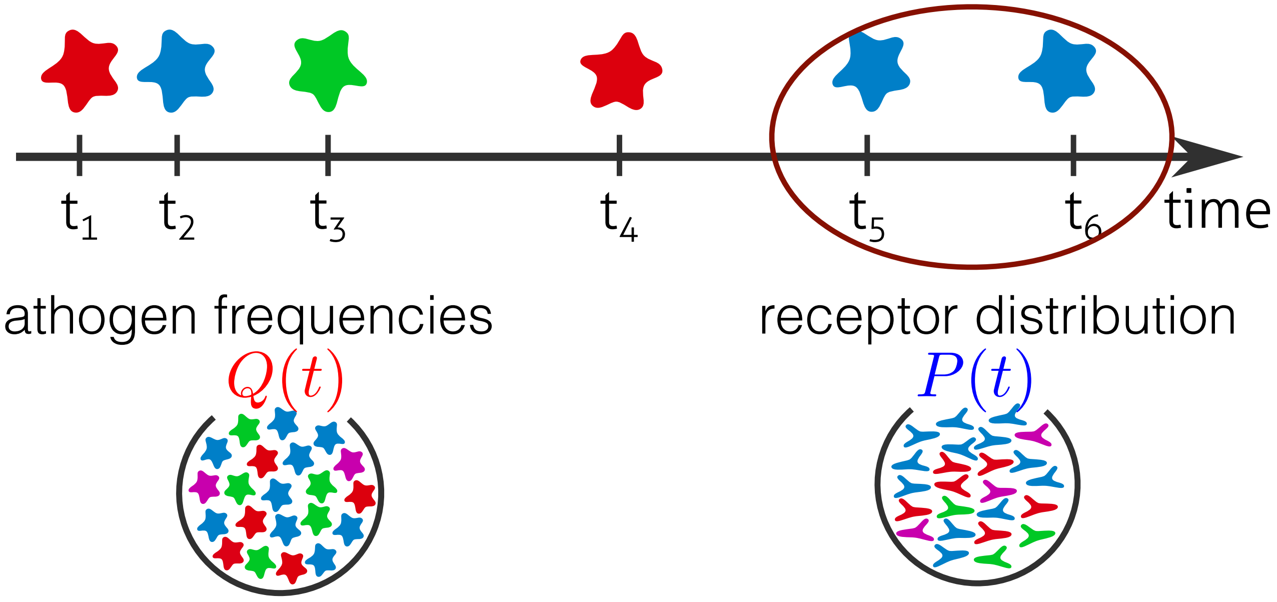
proportional memory update:


option 2:

modulated memory update:

# Estimate pathogen frequencies

$\lambda$ Poisson sampling rate of environment



what receptor distribution P(t) minimizes expected harm?

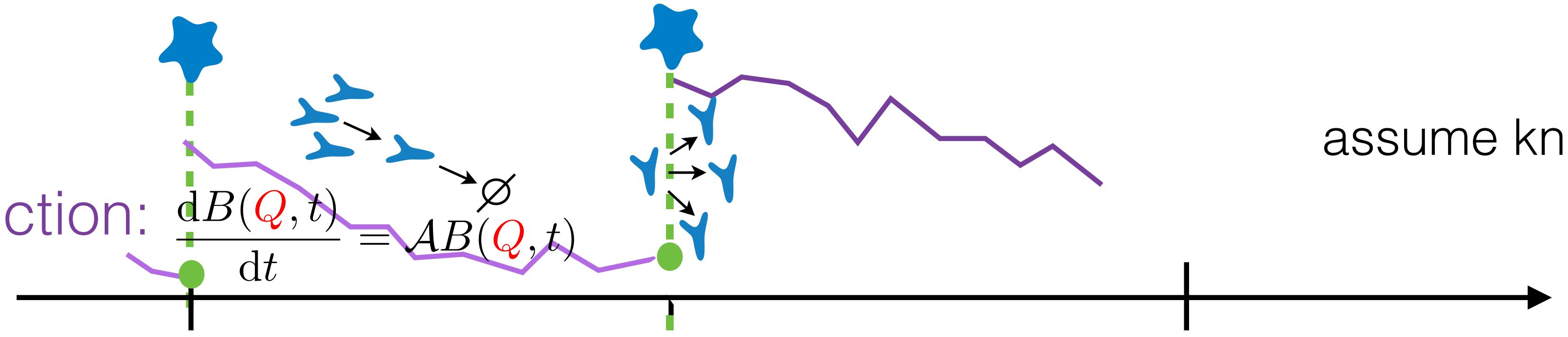
- Q(t) unknown → estimate it

belief of Q(t)

$$\langle \mathrm{Cost}\,(\mathrm{P(t)},\mathrm{Q(t)})\rangle = \int \mathrm{dQ}\ \mathrm{Cost}\,(\mathrm{P(t)},\mathrm{Q})\,\mathrm{B}\,(\mathrm{Q},\mathrm{t}) \xrightarrow{\text{min Cost}} P^{\star}(t) = f\left(\langle Q(t)\rangle\right)$$

expected cost of an infection

- propagate belief in time = encounters + prior



prediction: $\frac{\mathrm{d}B(Q,t)}{\mathrm{d}t} = \mathcal{A}B(Q,t)$

assume know real dynamics $\mathcal{A}$

update: $B(Q,t^{+}) = \frac{\mathcal{L}(a|Q)B(Q,t^{-})}{Z}$

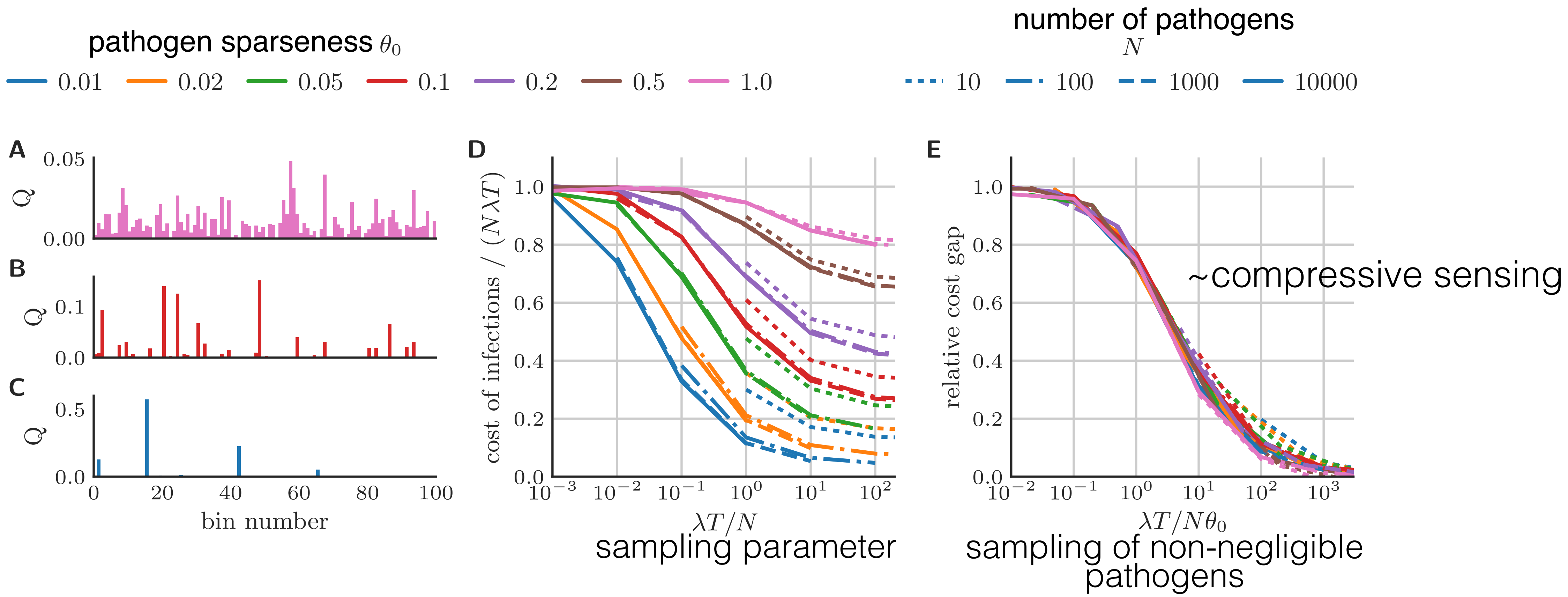
$Q_a(t^{+})$

MEMORY REPERTOIRE

# Memory helps

- memory helps in sparse environments

  → fast detection of few pathogens



~compressive sensing

sampling parameter

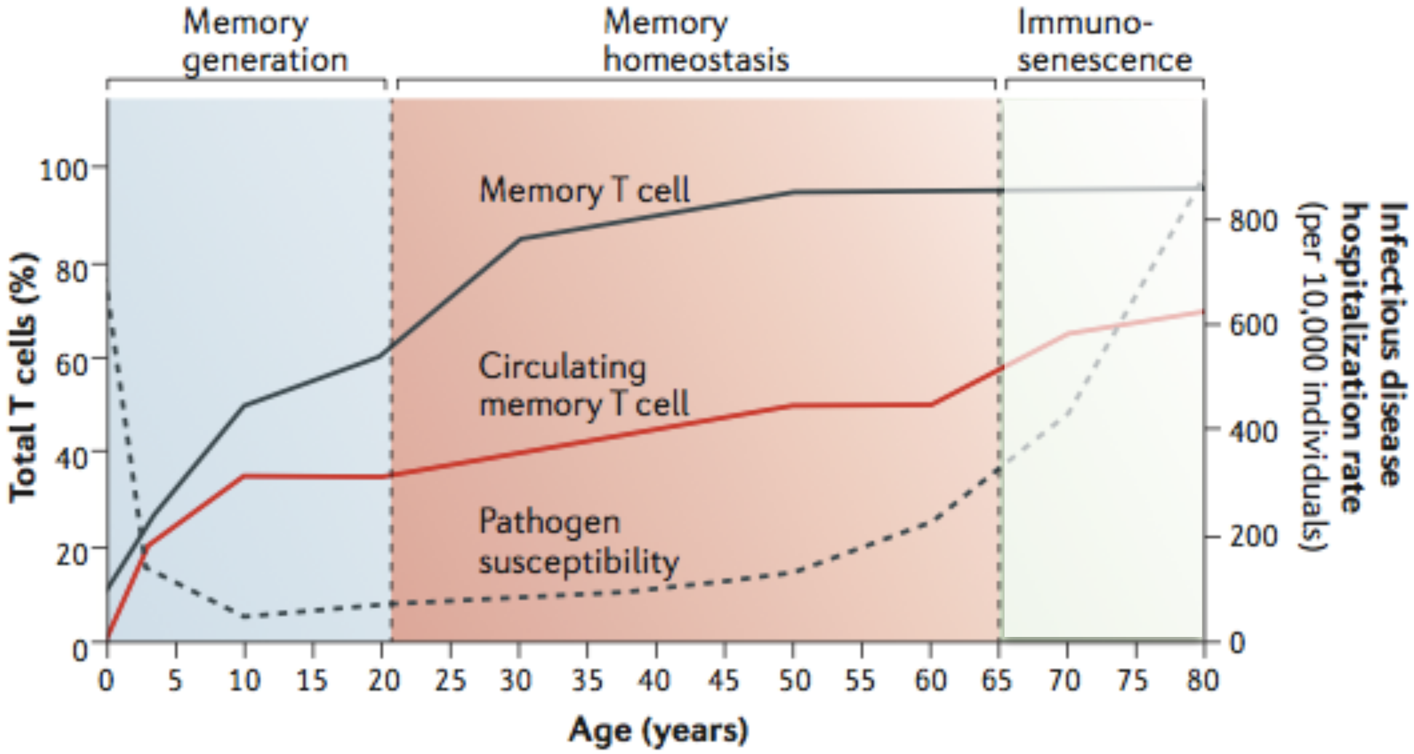
sampling of non-negligible pathogens

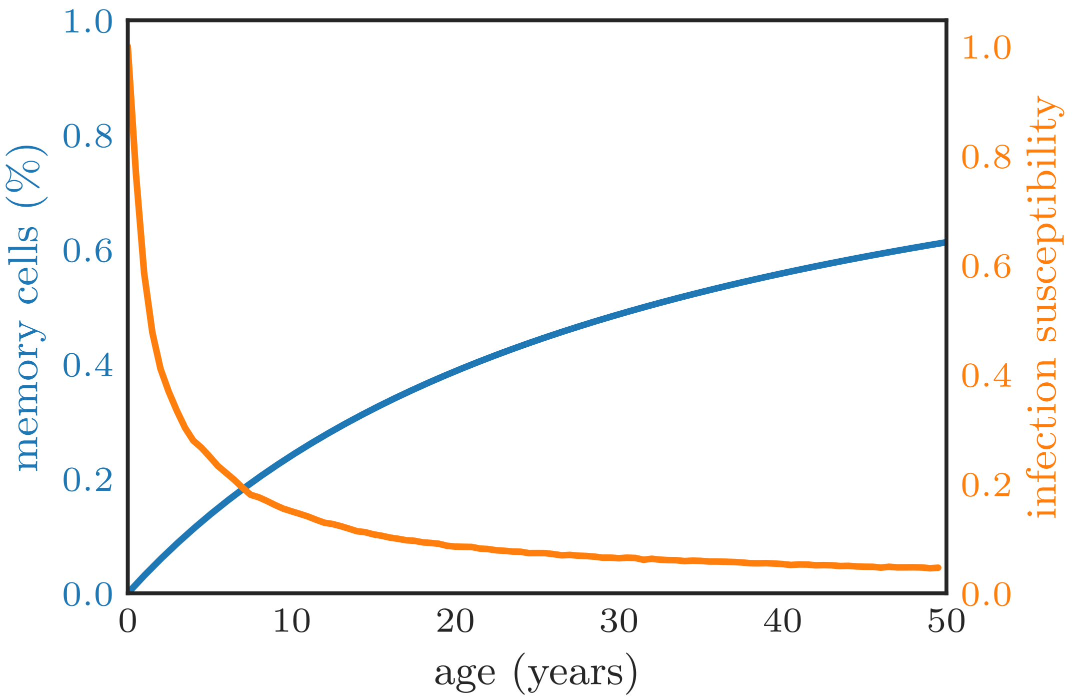
- advantage of memory - depends on sampling

→control theory rationalizes existence of immunological memory

# Rapidly acquired memory

- quickly learn global features of the distribution



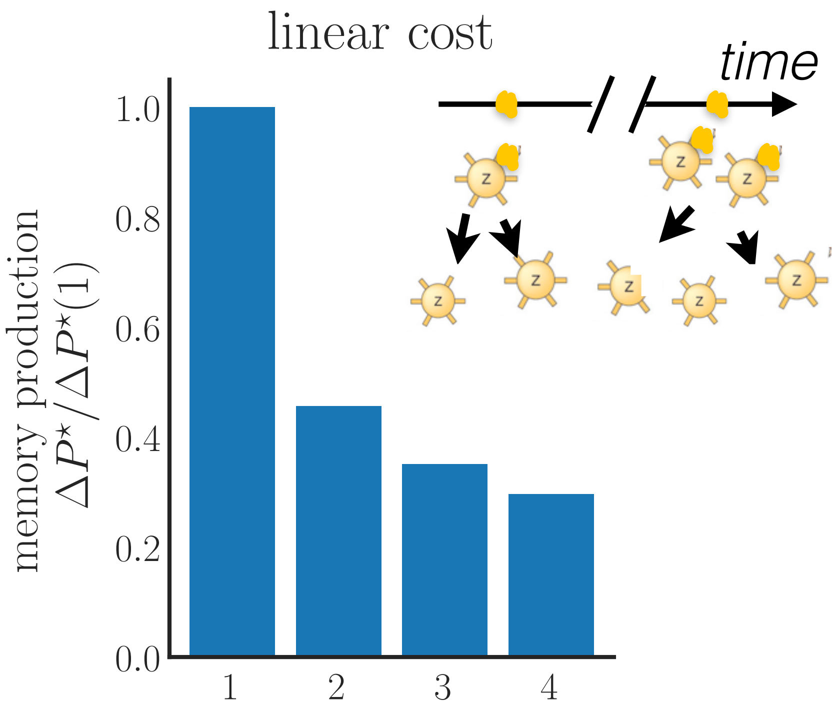


- predictive learning reproduces experimental features

*D. L. Farber, N. A. Yudanin· N. P. Restifo,NATURE REVIEWS | IMMUNOLOGY 2013*

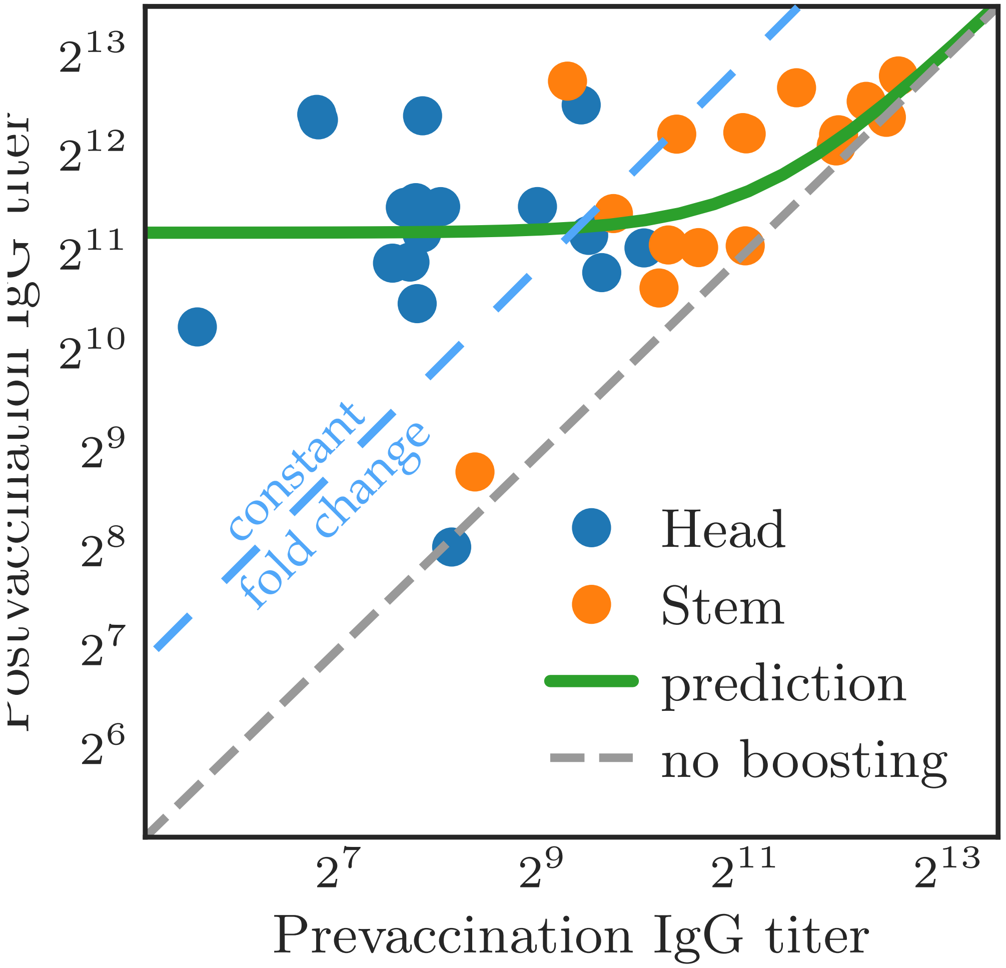
# Learning

- subsequent observations count as less evidence



→ *vaccination*

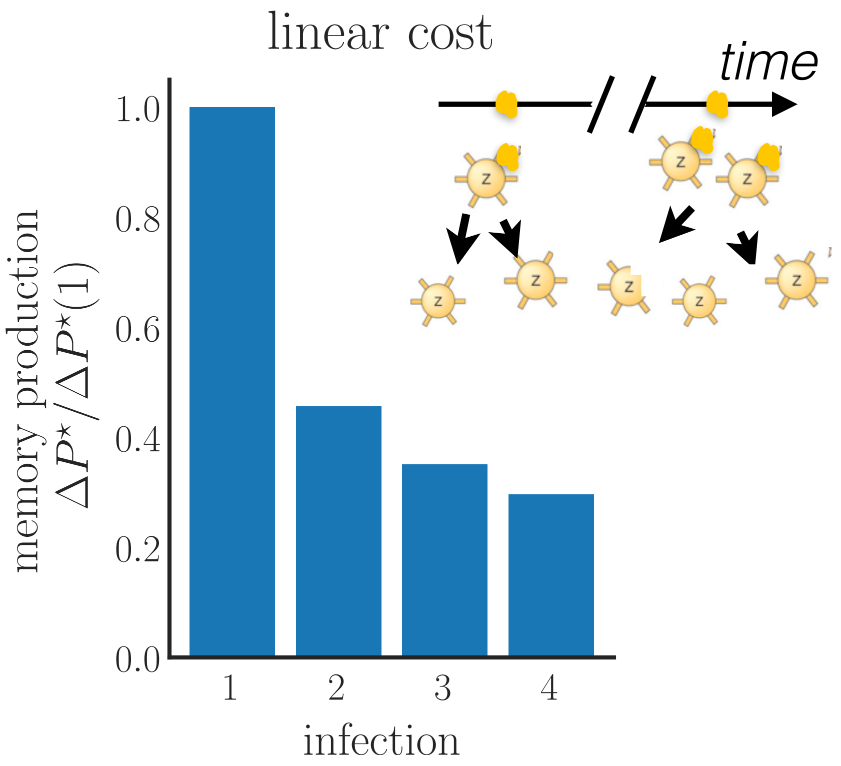
- booster vaccination titers for epitopes of hemaglutanin following vaccination with inactivated H5N1



*data from Ellebedy et al. PNAS 2014*

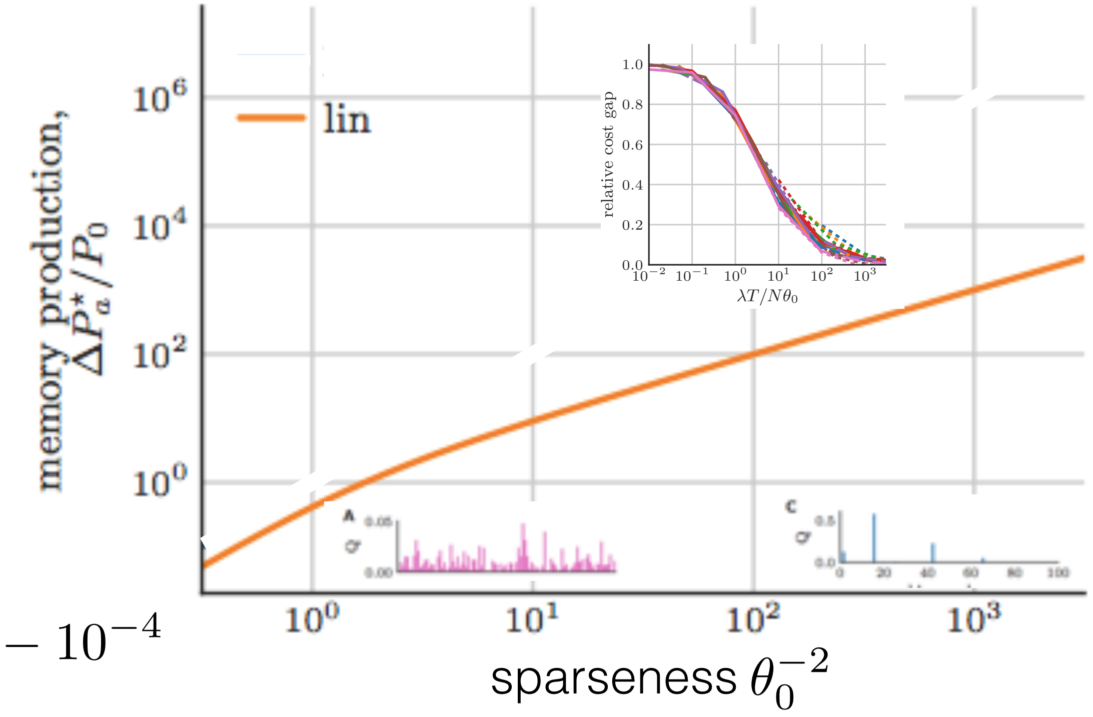
# Learning

- subsequent observations count as less evidence



- stronger response in sparse environments

memory increase ~100-1000*



→ very sparse environment $\theta_0 \sim 10^{-6} - 10^{-4}$

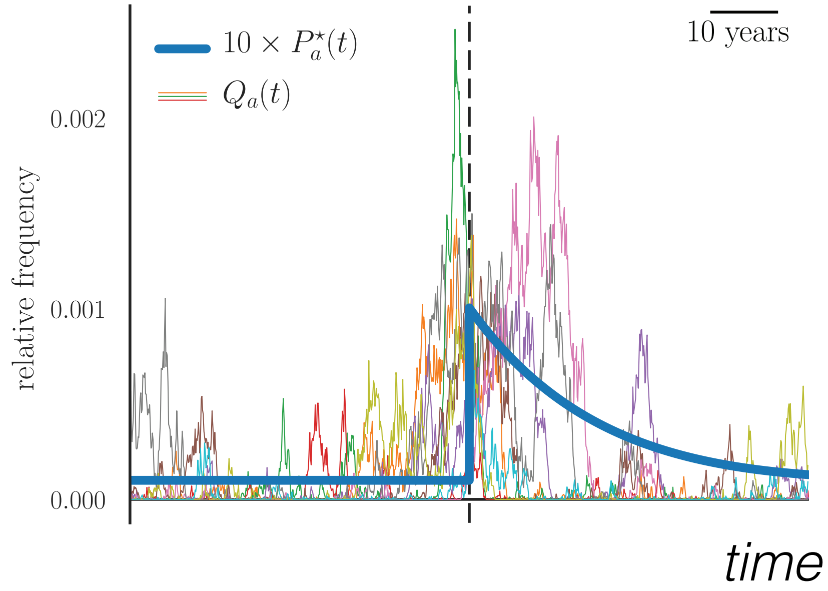
→ pathogen seen once - memory gives ~2 fold cost decrease

*Buchholz et al (2016) An. rev. imm.

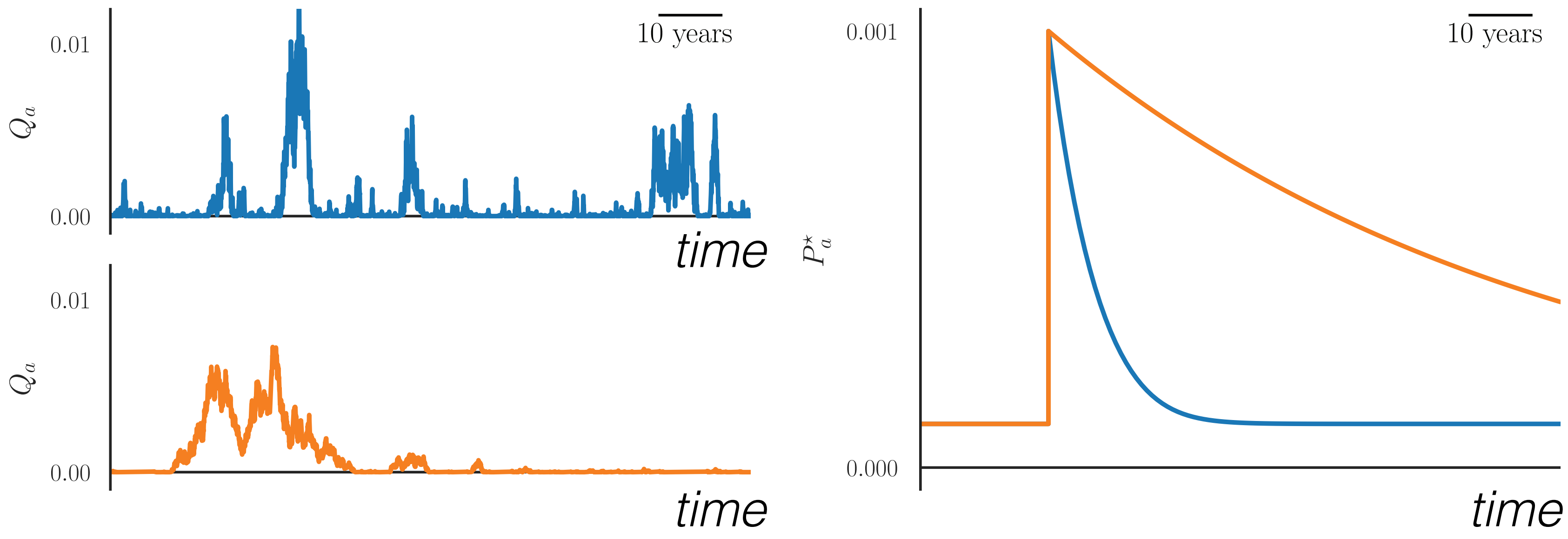
# Forgetting

- changing environment → propagate belief between sampling

- Bayesian prediction of changing environment



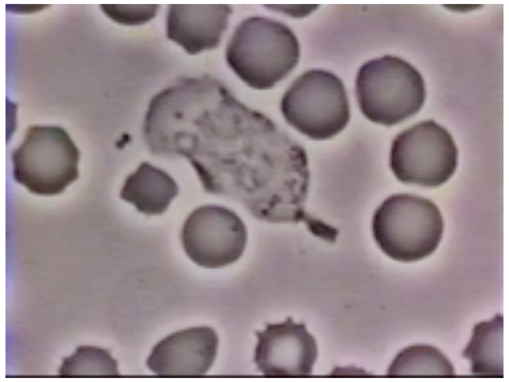
- forget faster in rapidly changing environments



# Different immune strategies

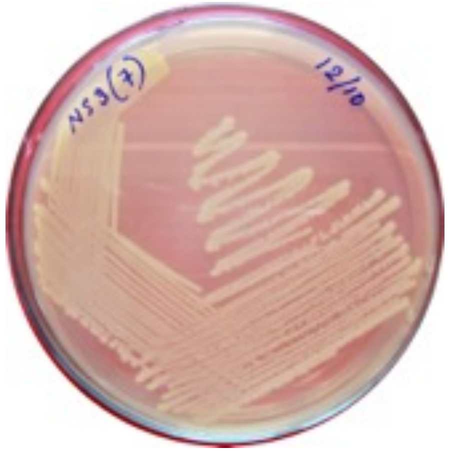
adaptive immunity

# Other immune strategies



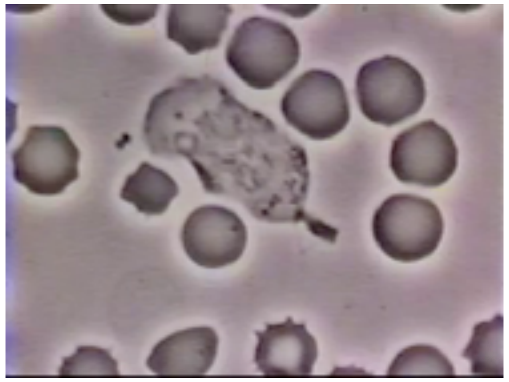
innate immunity

adaptive immunity

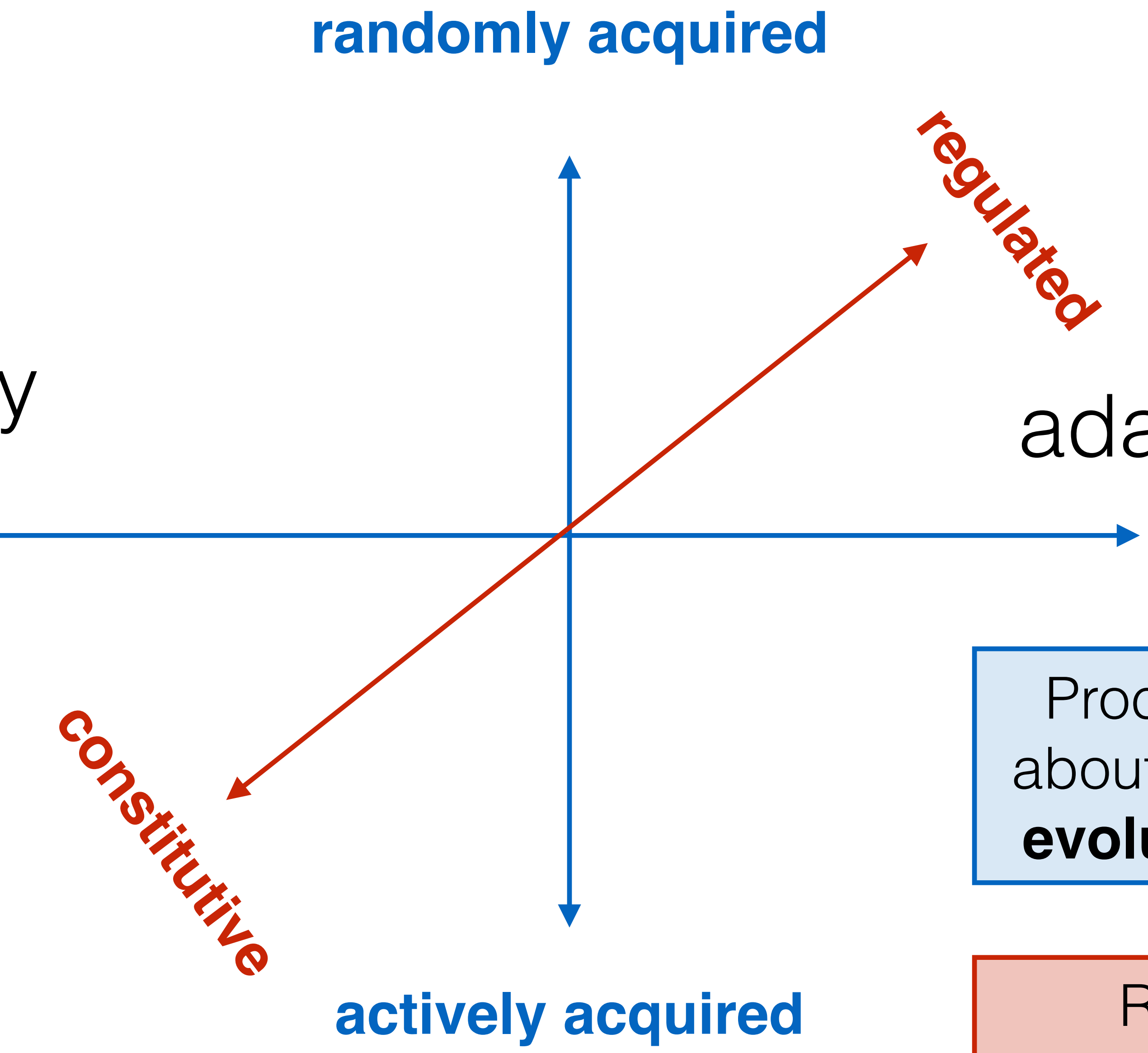


CRISPR immunity

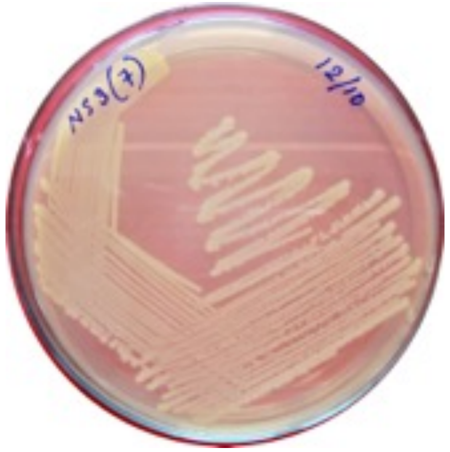
# Common strategic choices



innate immunity



adaptive immunity

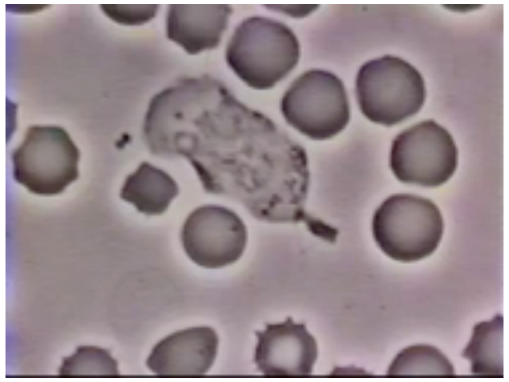


CRISPR immunity

Processing information about the environment on **evolutionary** timescales

Response during **organism** lifetime

# Common strategic choices



randomly acquired

regulate

in

optimal immunity?

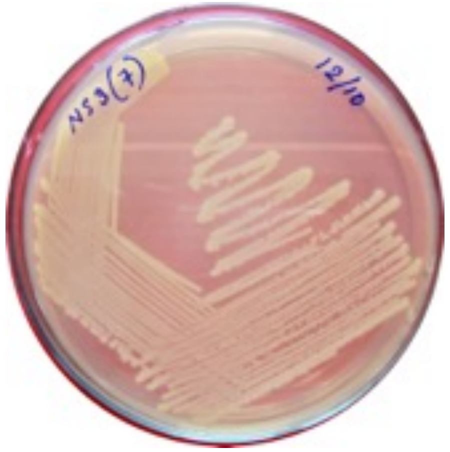
ty

constitutive

Processing information about the environment on **evolutionary** timescales

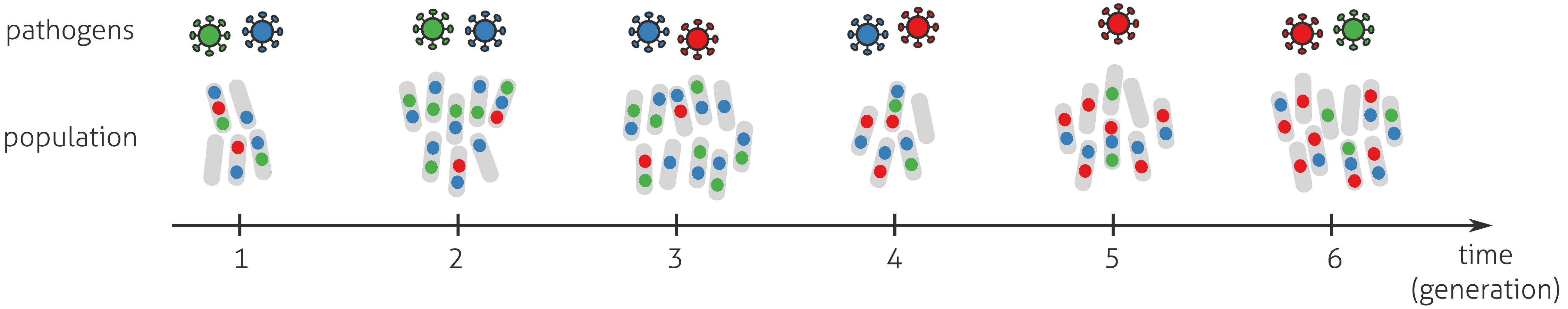
Response during **organism** lifetime

actively acquired



CRISPR immunity

# Optimal immunity

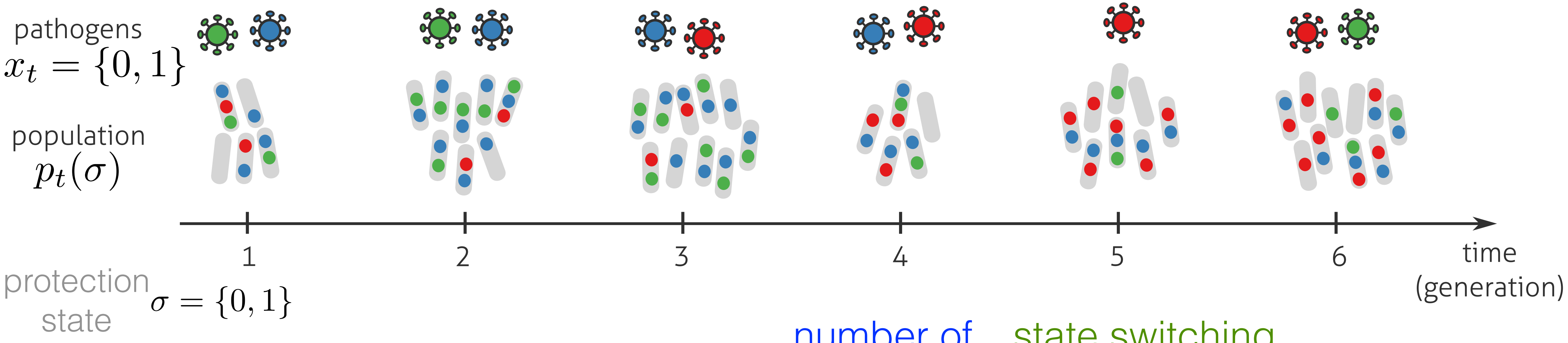


- match environment statistics
- ensure long term population growth

→ immunity as adaptation to pathogen statistics

- consider different strategies
- optimize long term population growth

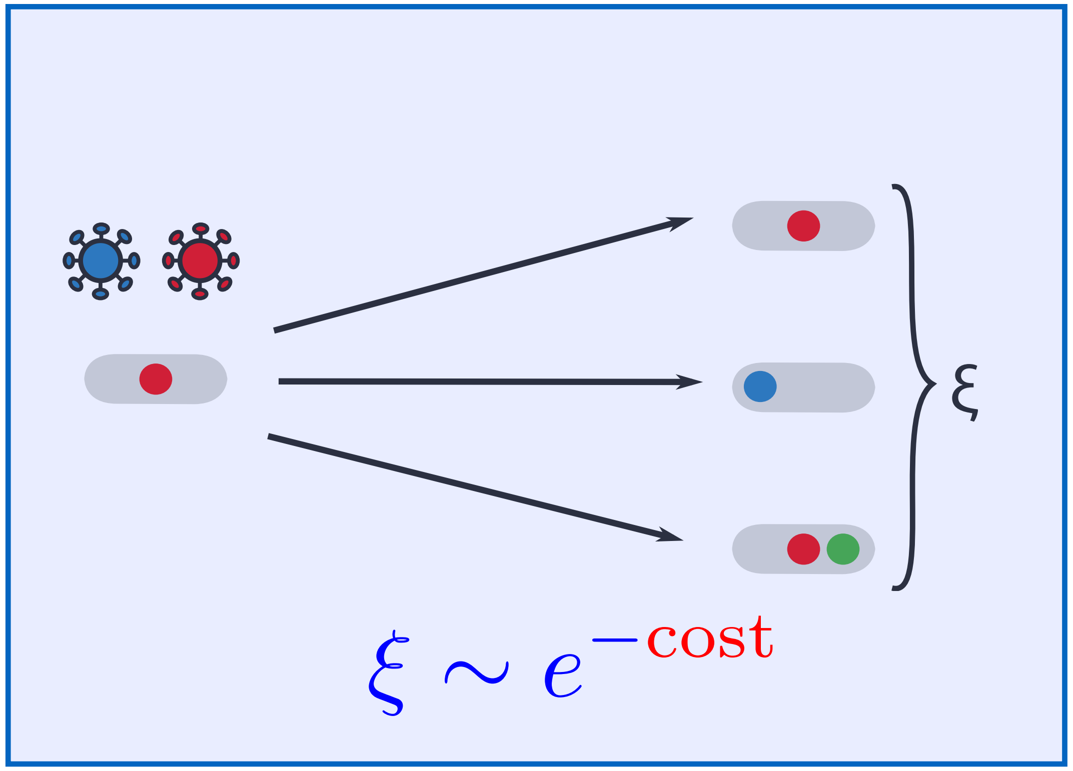
# Population growth



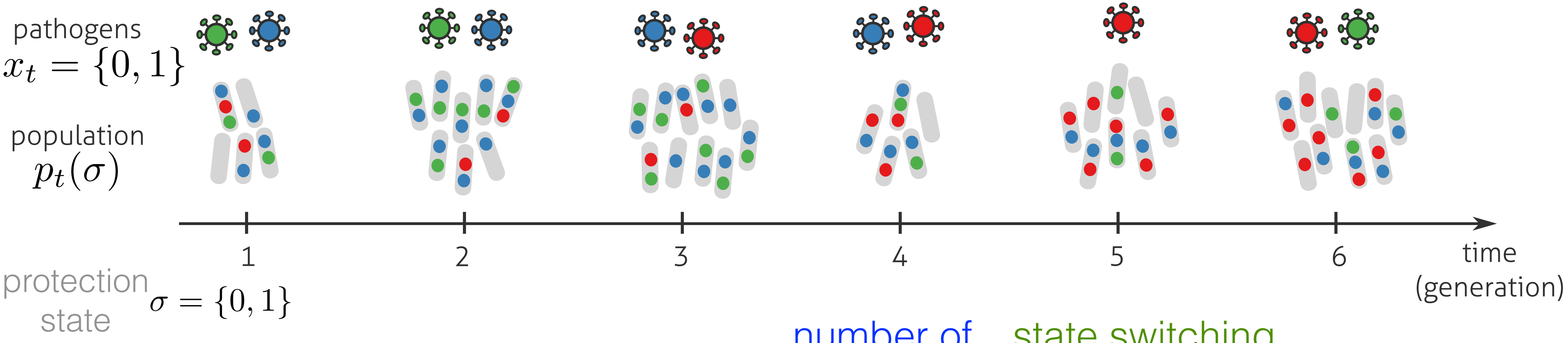
number of offspring

state switching probability

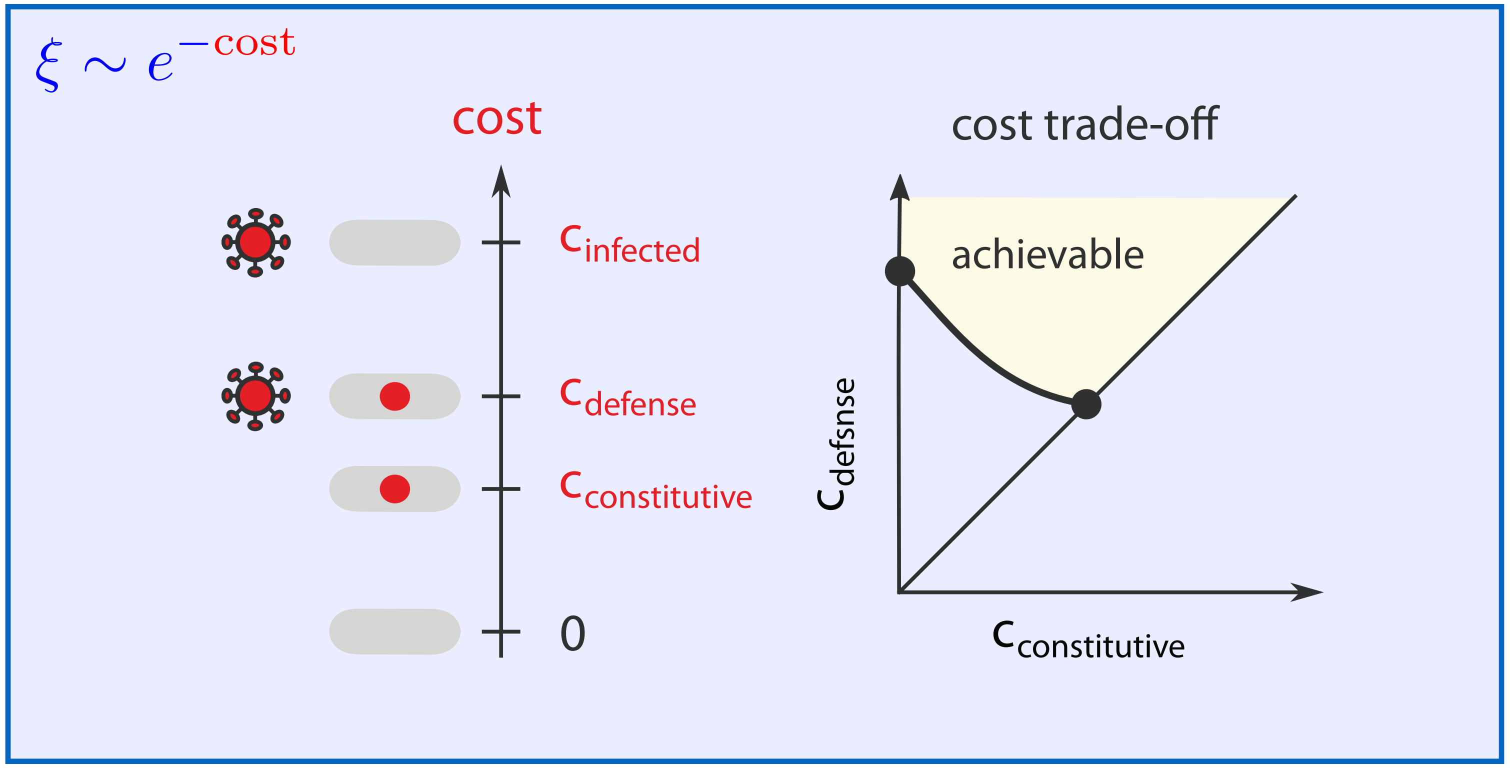
$$p_{t+1}(\sigma) = \frac{1}{Z_t} \sum_{\sigma'} \xi(\sigma', x_t)\, \pi(\sigma|\sigma', x_t)\, p_t(\sigma')$$



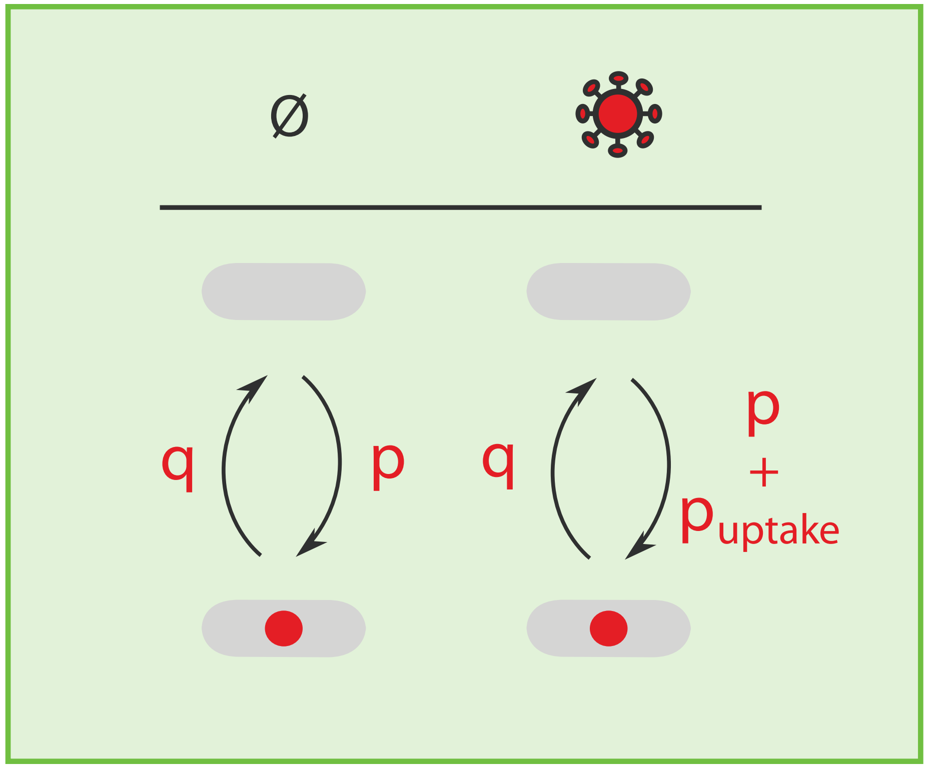
# Population growth



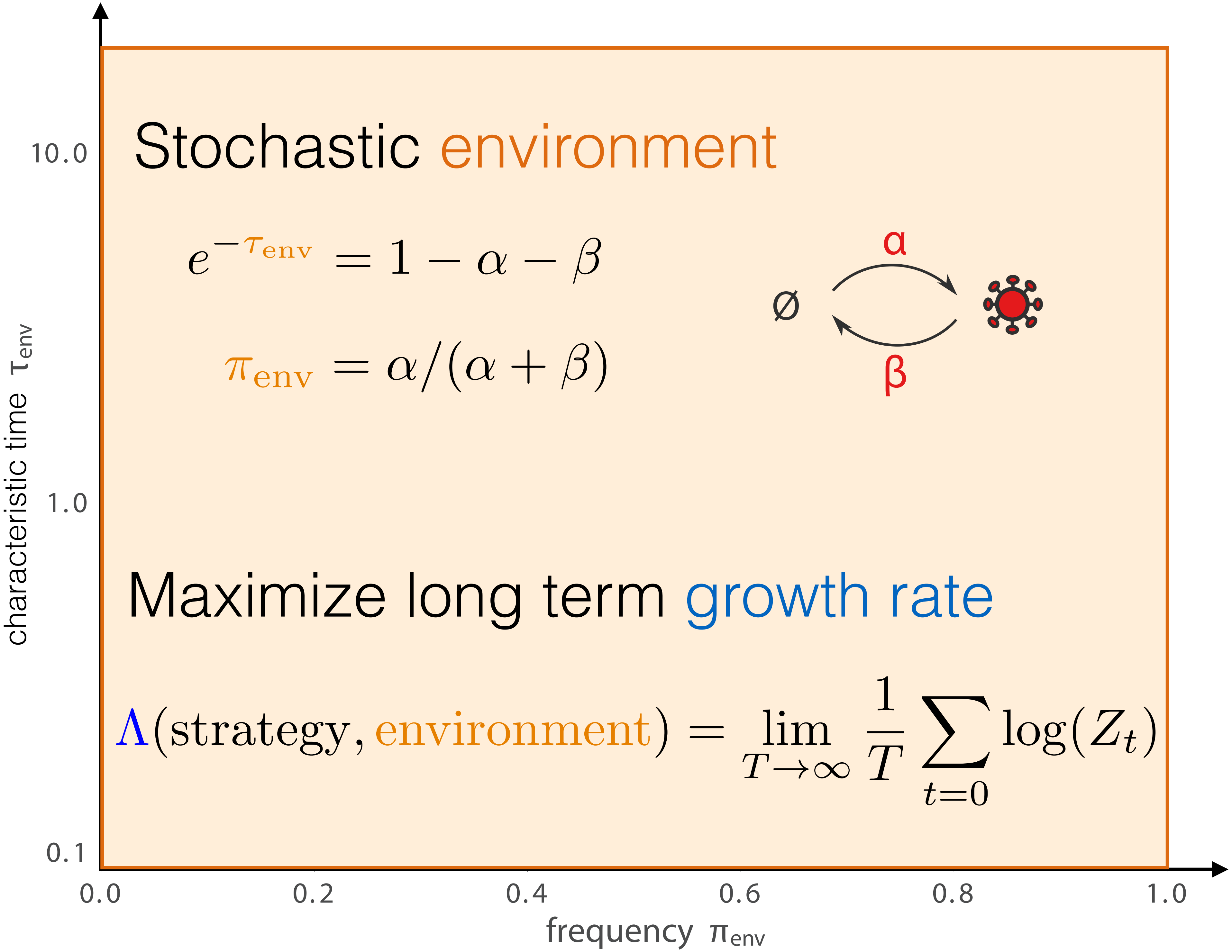
$$p_{t+1}(\sigma) \quad = \quad \frac{1}{Z_t}\sum_{\sigma'} \overset{\text{number of offspring}}{\xi(\sigma', x_t)}\ \overset{\text{state switching probability}}{\pi(\sigma|\sigma', x_t)}\ p_t(\sigma')$$



$\xi \sim e^{-\text{cost}}$



# The environment


10.0
1.0
0.1
characteristic time $\tau_{\mathrm{env}}$
0.0
0.2
0.4
0.6
0.8
1.0
frequency $\pi_{\mathrm{env}}$
∅
α
β

Stochastic environment

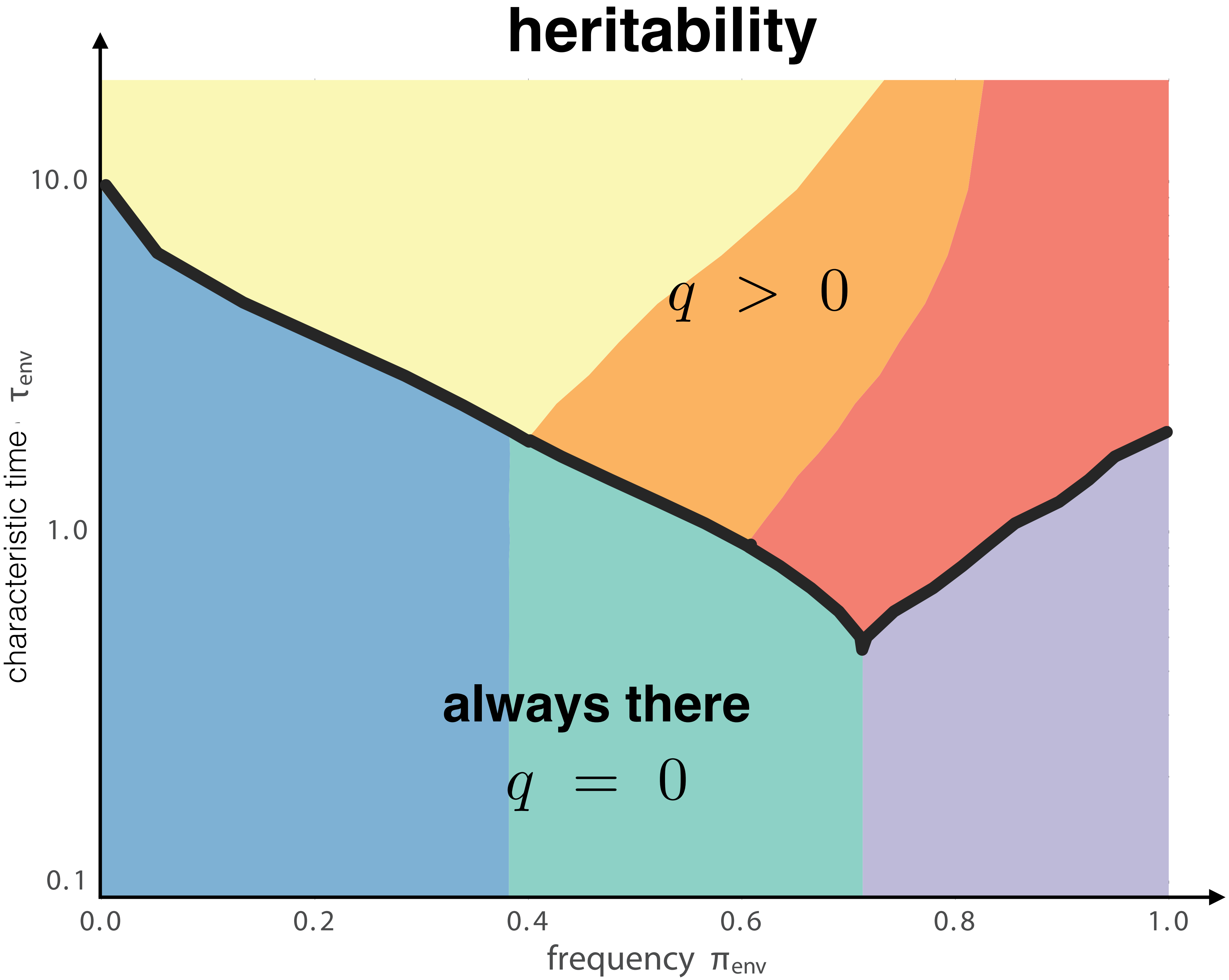
$$e^{-\tau_{\mathrm{env}}} = 1 - \alpha - \beta$$

$$\pi_{\mathrm{env}} = \alpha/(\alpha + \beta)$$

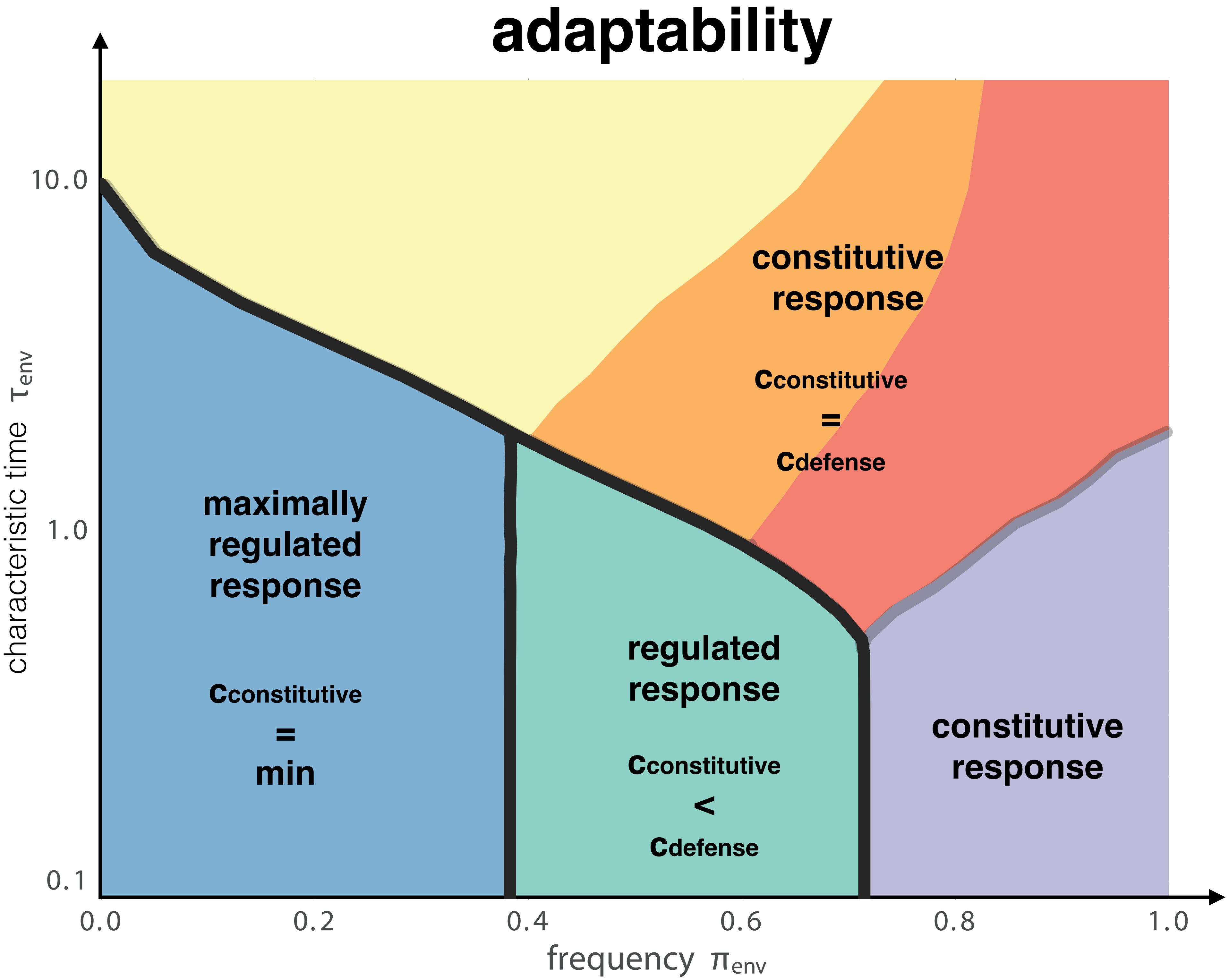
Maximize long term growth rate

$$\Lambda(\mathrm{strategy}, \mathrm{environment}) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0} \log(Z_t)$$

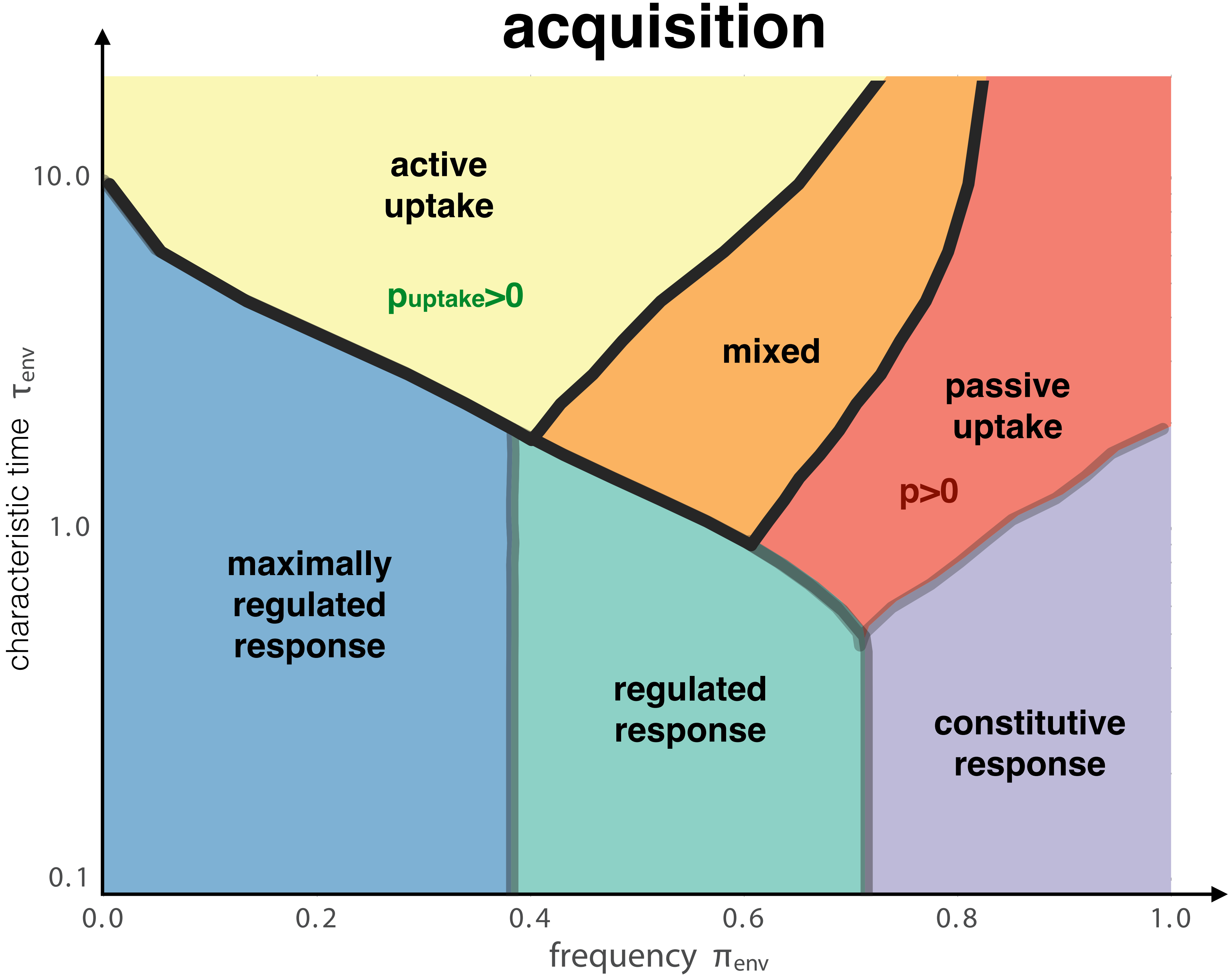
# Optimal strategies



# Optimal strategies

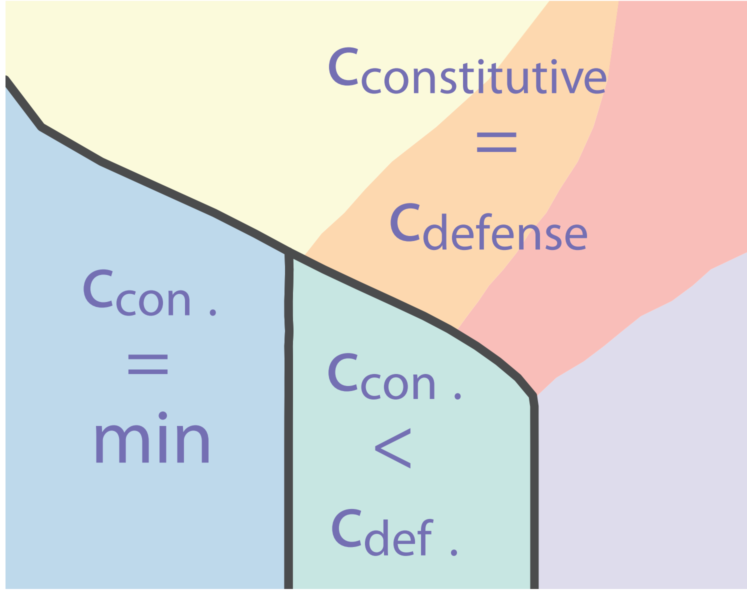


# Optimal strategies

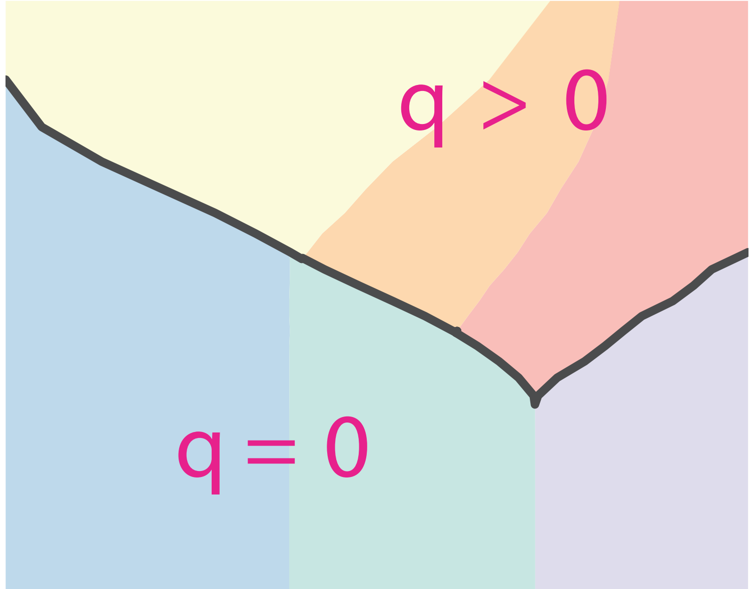


# Three strategy axes

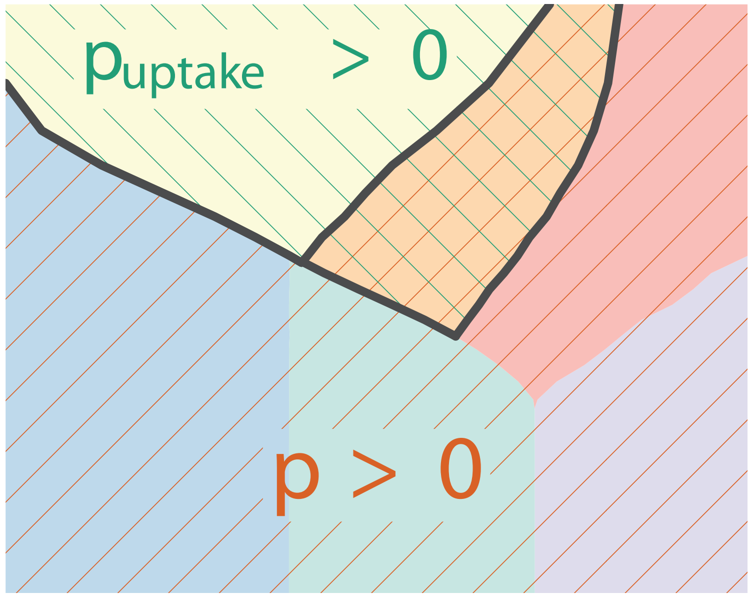
adaptability



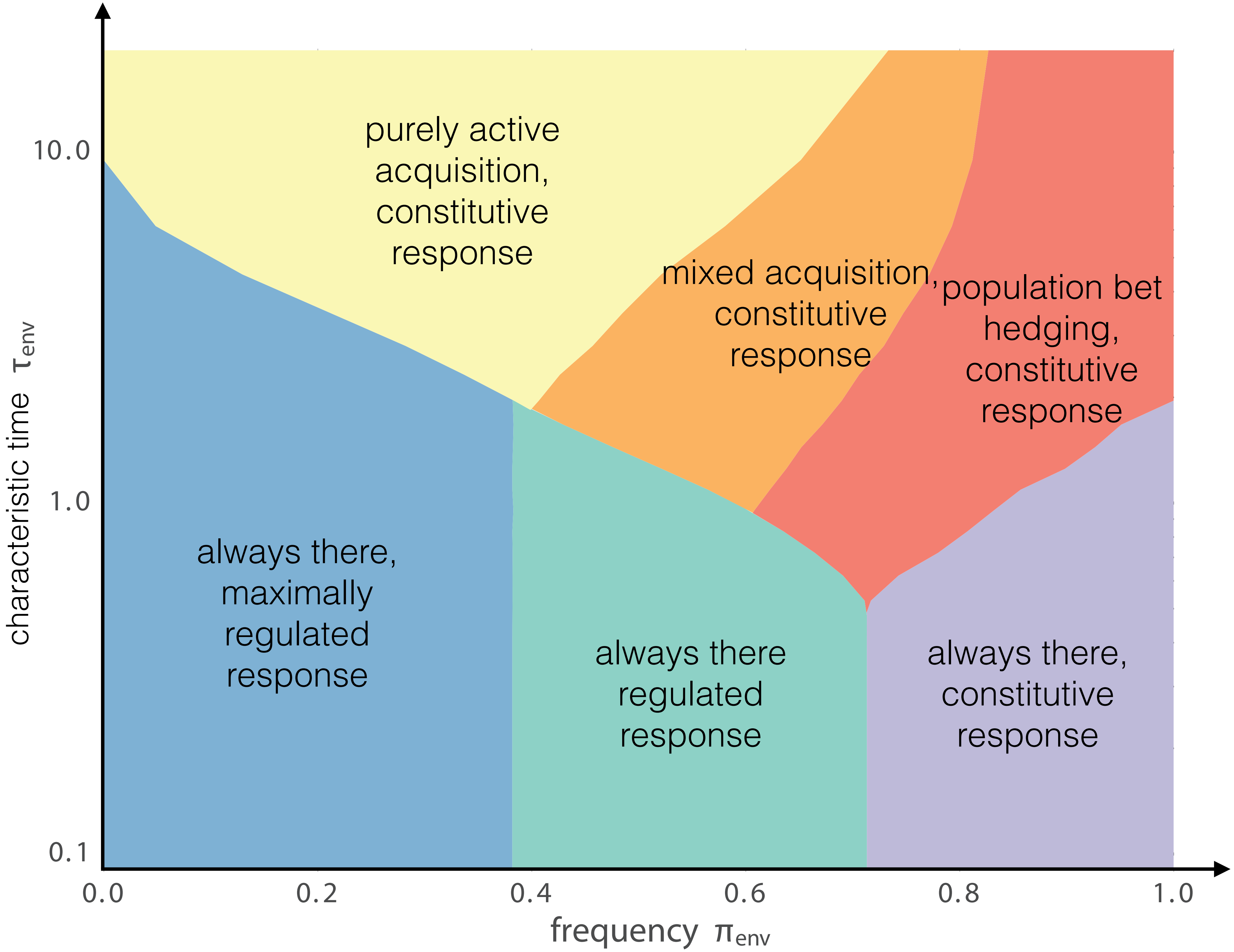
heritability



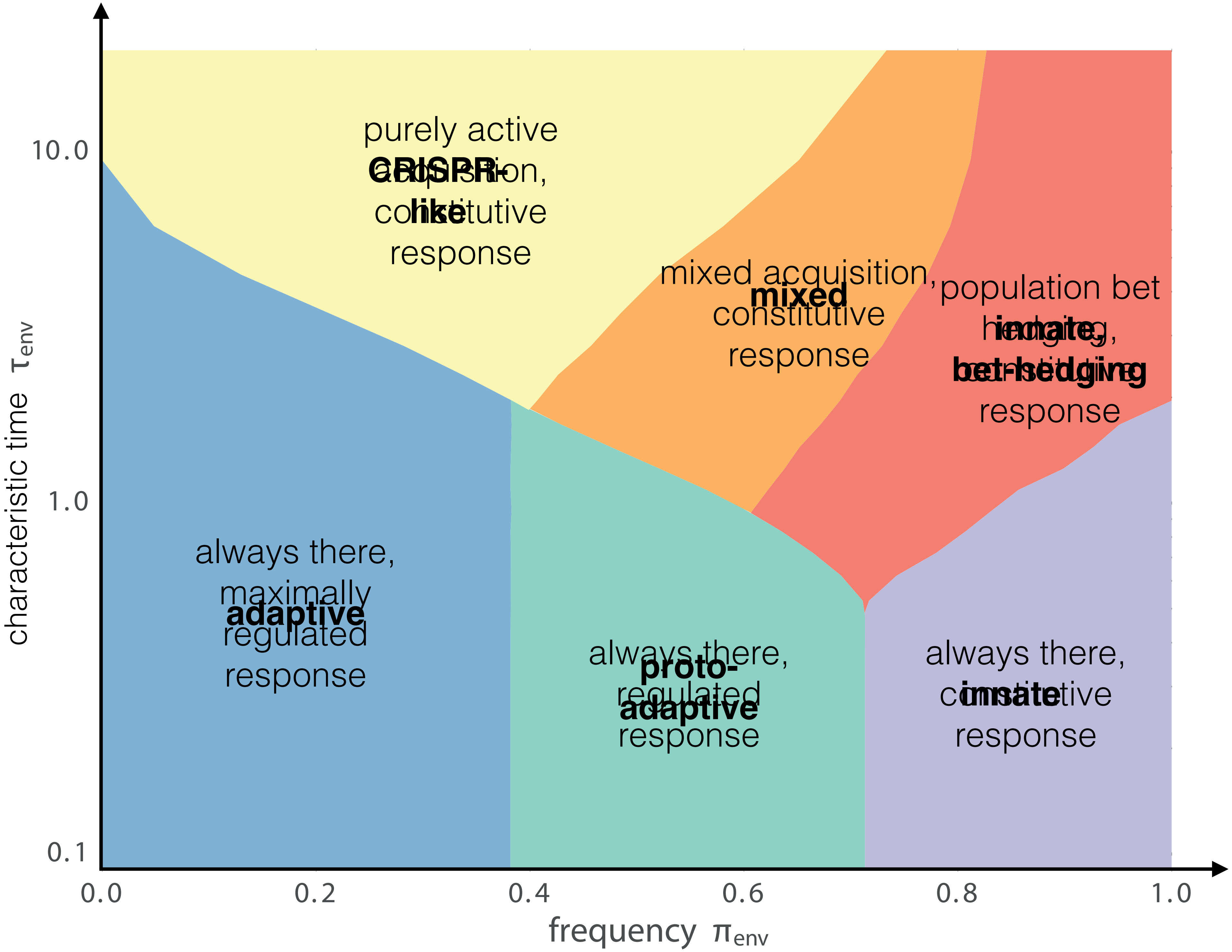
acquisition mode



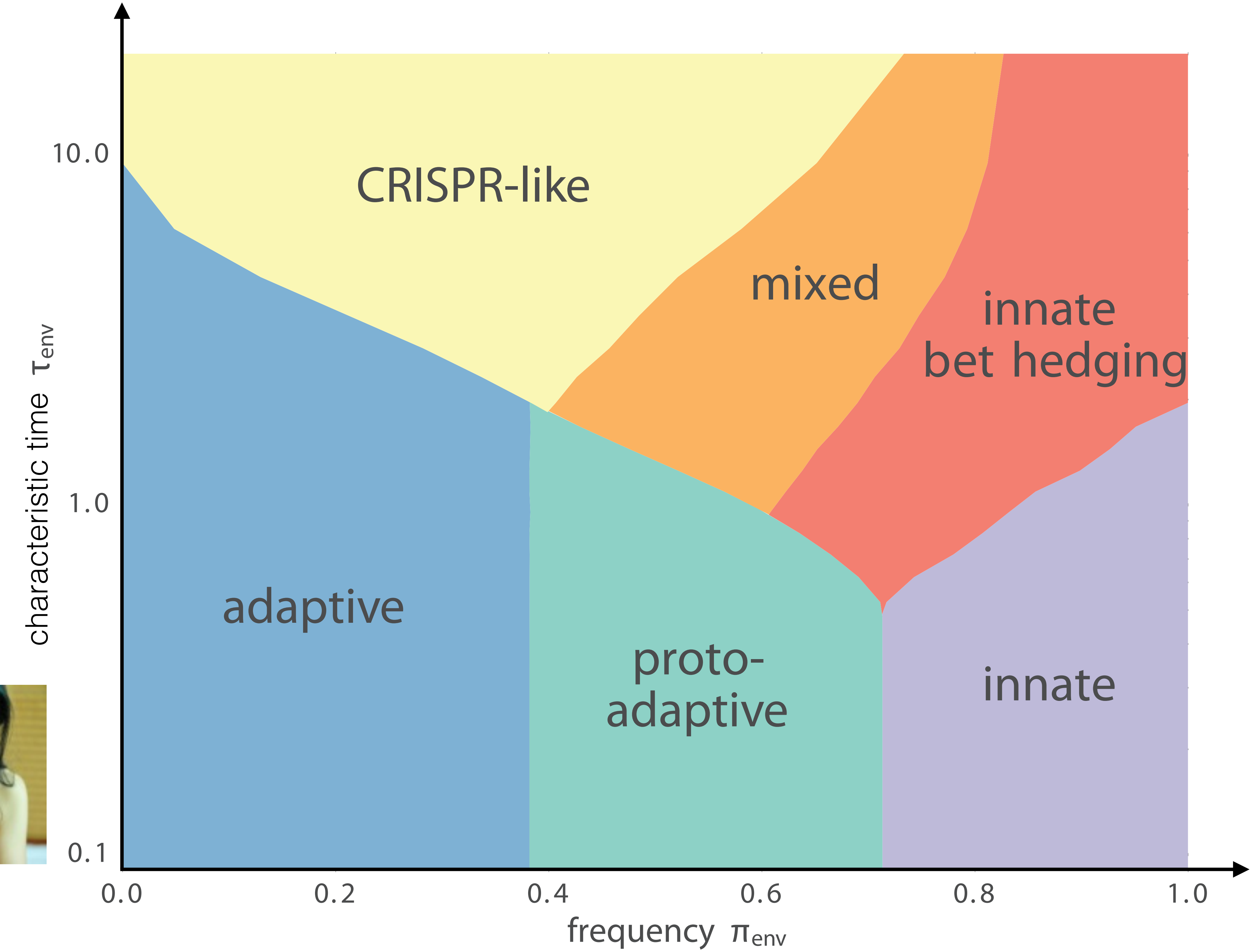
# Optimal strategies



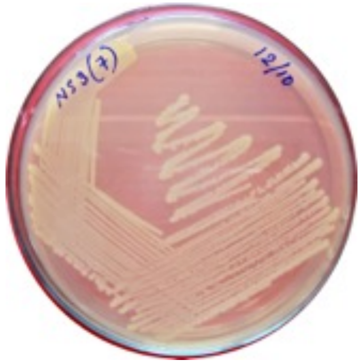
# Optimal immune systems

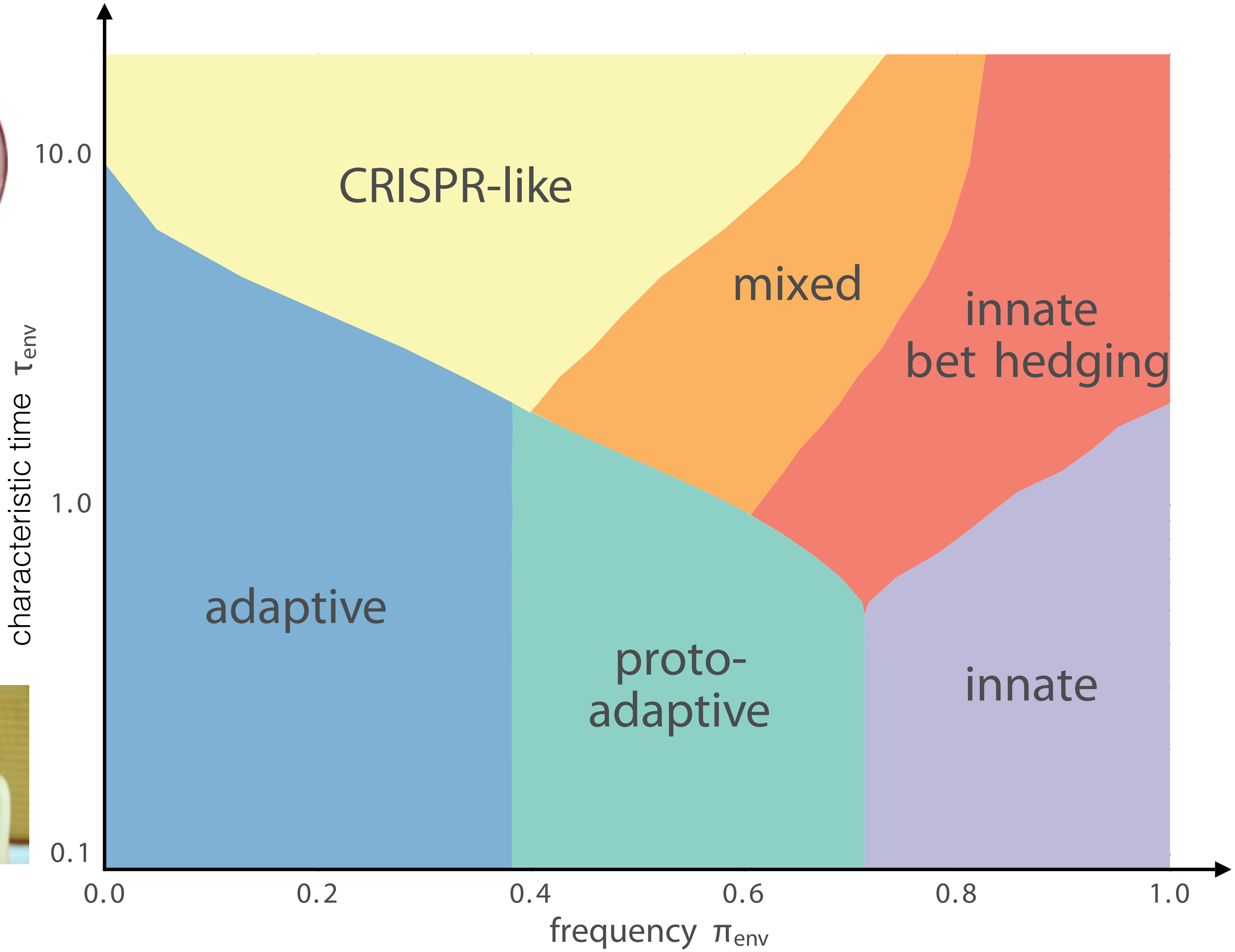


# Optimal immune systems



# Optimal immune systems





# Conclusions

## generating diversity and selection:

- random overlap between (most) individuals
- very long lived clones

## optimal repertoires:

- cover space but are random
- differ in two individuals

## predicting immune systems:

- use dynamics to anticipate frequencies
- memory useful in sparse environments

## optimal immunity:

- known immunity from evolutionary constraints
- depends on environment statistic

T Mora, AM Walczak, W Bialek, CG Callan, PNAS (2010)
A Murugan, T Mora, AM Walczak, CG Callan, PNAS (2012)
Y Elhanati, A Murugan, CG Callan, T Mora, AM Walczak, PNAS (2014)
A Mayer, V Balasubramanian, T Mora, AM Walczak, PNAS (2015)
Y Elhanati, Z Sethna, Q Marcou, CG Callan, T Mora, AM Walczak, Phil. Trans. B (2015)
J. Desponds, T. Mora, AM Walczak, PNAS (2015)
A Mayer, T Mora, O Rivoire, AM Walczak, PNAS (2016)
Y Elhanati, Q Marcou, T Mora, AM Walczak, Bioinformatics (2016)
RM Adams, JB Kinney, T Mora, AM Walczak, eLife (2017)
M. Pogorelyy et al, PloS CB (2017)
T Mora, AM Walczak, qbio/bioarxiv (2016)
Z Sethna, Y Elhanati, CG Callan, T Mora, AM Walczak, PNAS (2017)
M Laessig, V Mustonenen, AM Walczak Nature Ecology & Evolution (2017)
Q. Marcou. T. Mora, AM Walczak, qbio/bioarxiv (2017)
M. Pogorelyy et al, qbio/bioarxiv (2017)