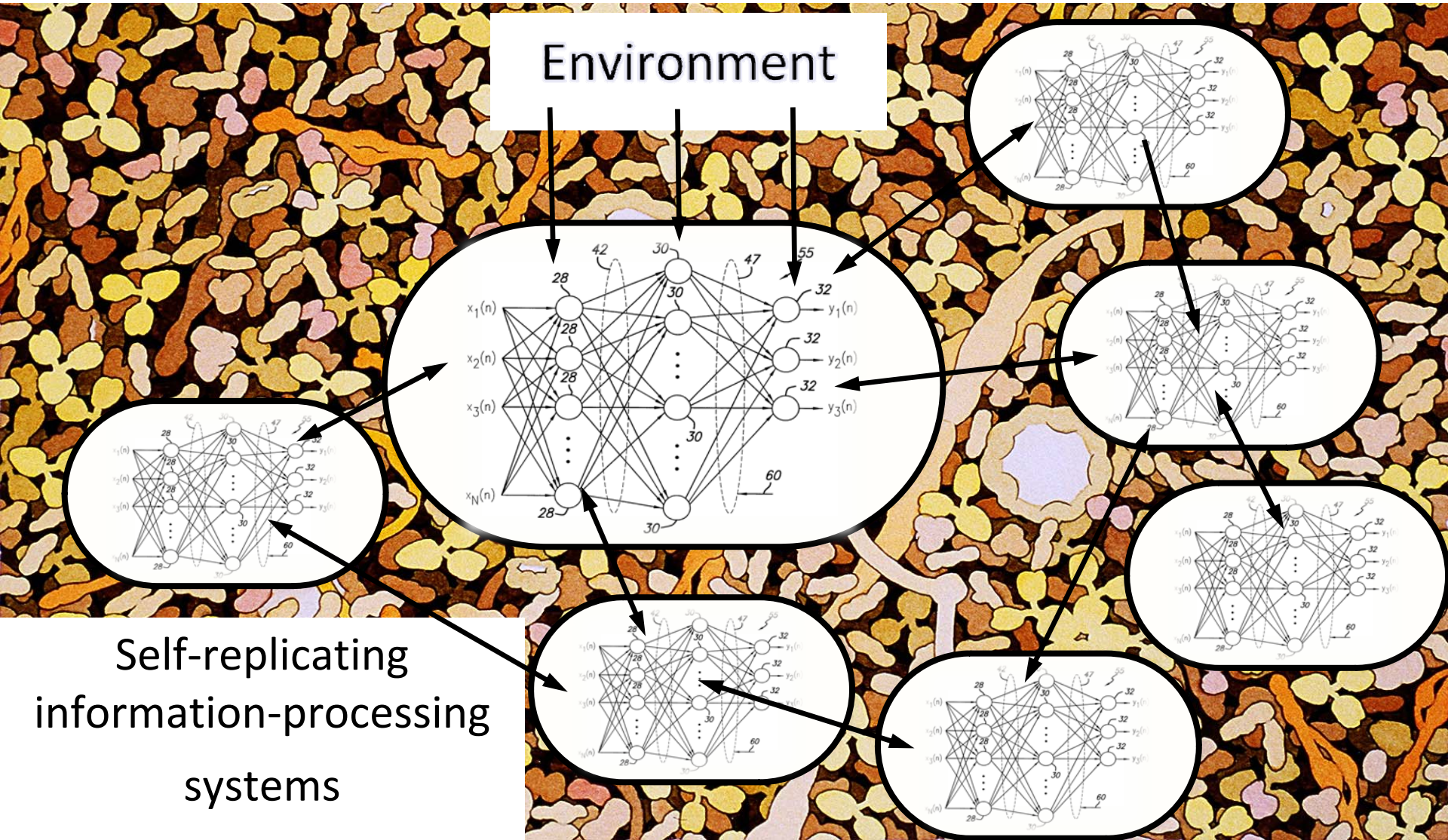


# **Casting Polymer Nets To Optimize Molecular Codes**

The physical language of molecules

Mathematical Biology Forum

# Biological information is carried by molecules



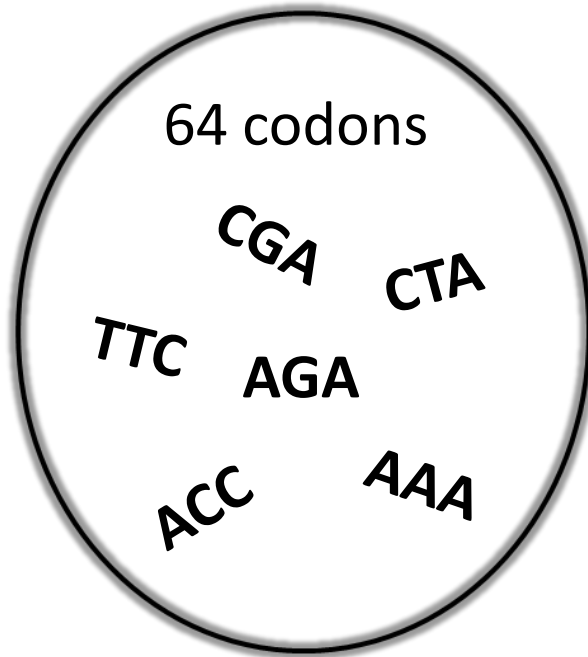
# Outline – The physical language of molecules

- Molecular code = **Mapping** symbols to meanings.
- Codes imply **partition** of symbol space.
- Code **fitness** = Quality + Cost of the partition.
- Code fitness = free energy of **polymer networks**.
- Polymer networks  $\leftrightarrow$  Spin networks.
- Evolution of networks  $\rightarrow$  coding transition where code emerges.

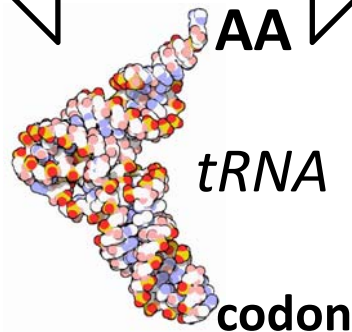
# Coding theory: Molecular Codes as Maps or Information Channels

- Molecular code = **map** relating two sets of molecules.
- Relation by **molecular recognition**.

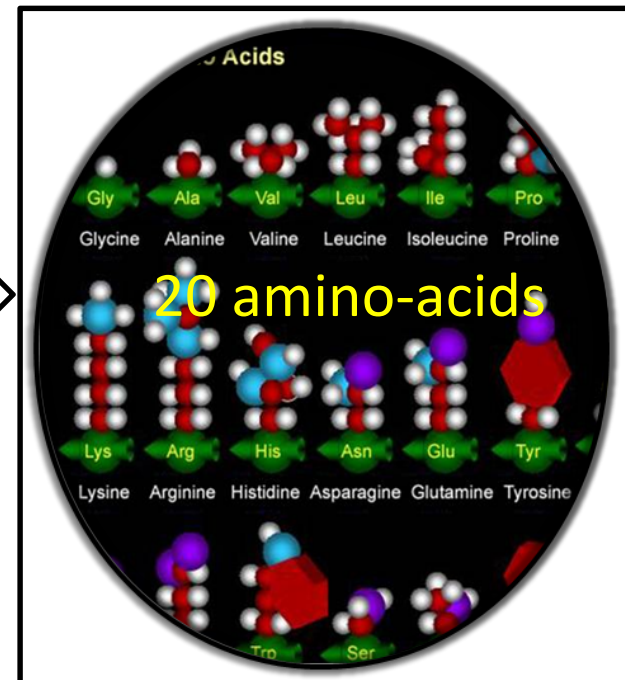
## *Symbols*



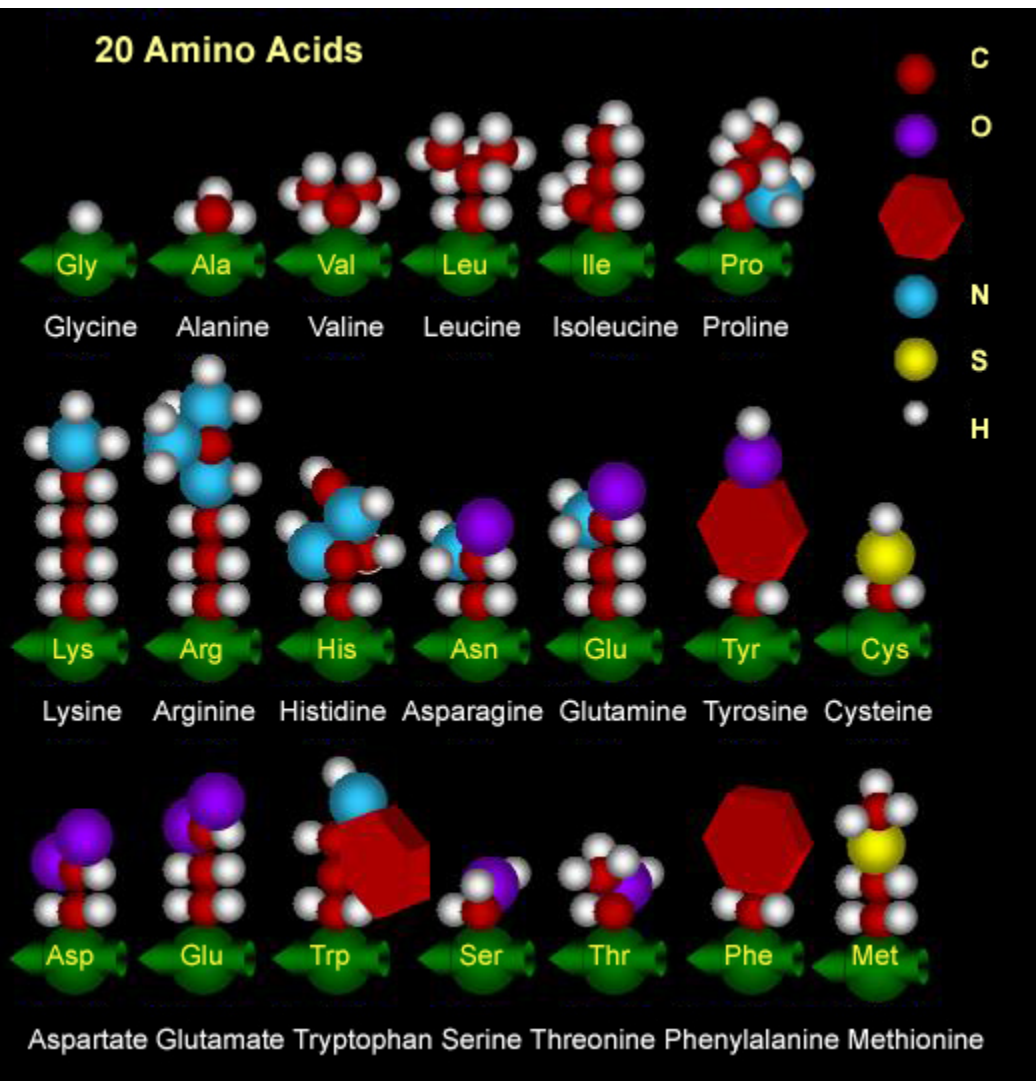
**Genetic Code**



## *Meanings*



# Diversity of molecular meanings is essential



- Diversity of amino acids allows proteins to perform a wide variety of functions efficiently.

# Semantic challenge of molecular coding is 3-fold

- Minimize **Error-Load**:

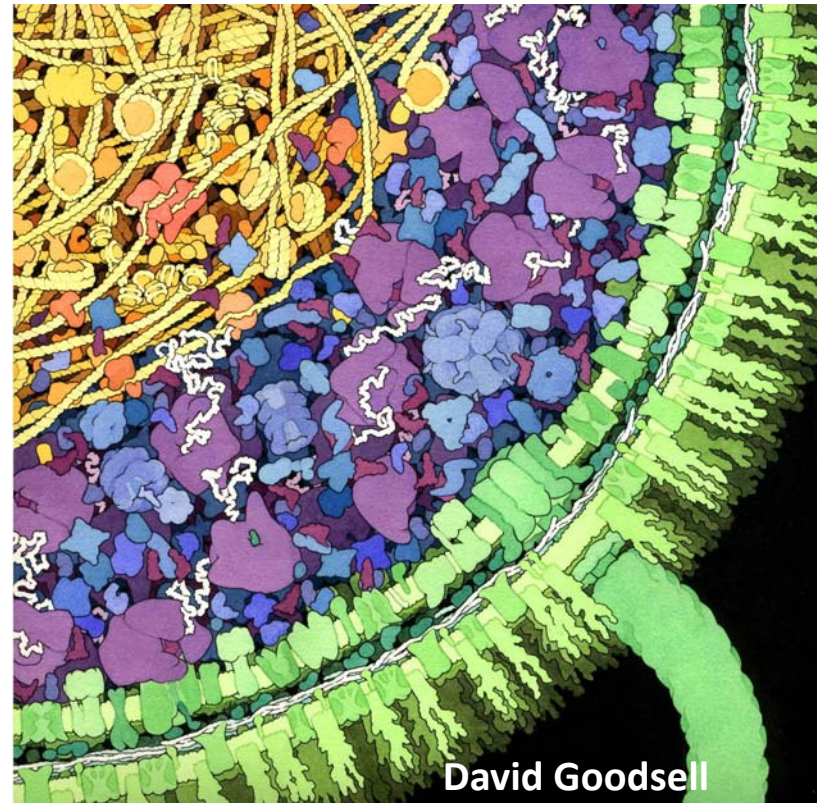
Information transfer via molecular recognition in a noisy, crowded milieu: Fluctuations, competing lookalikes...

- Minimize **Cost**:

How to construct codes at minimal cost of resources?

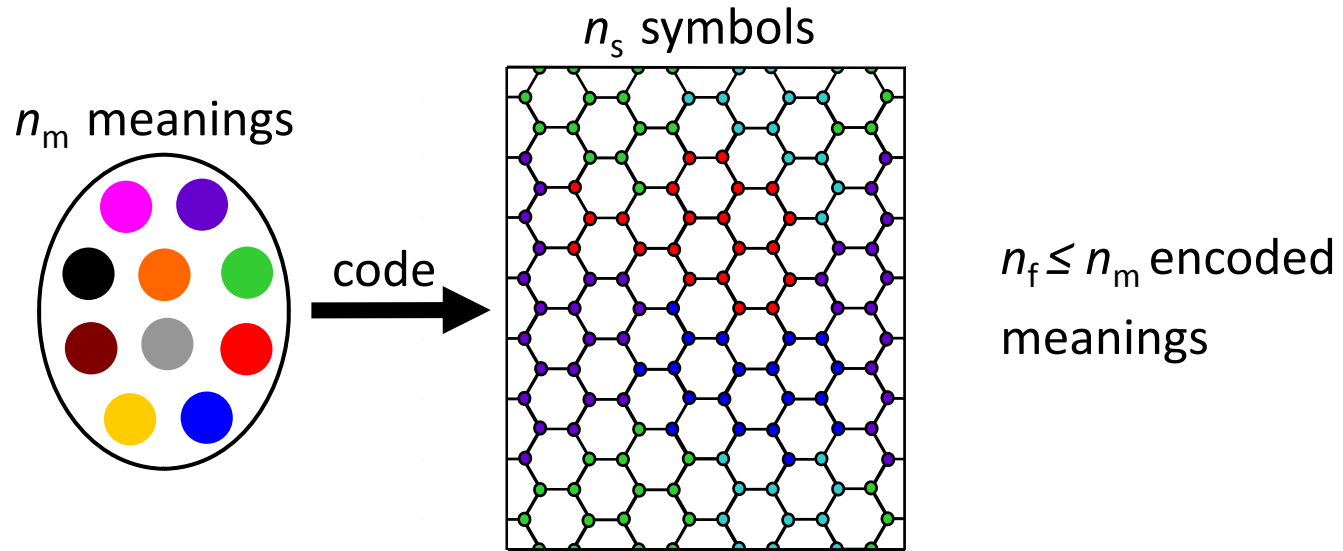
- Maximize **Diversity**:

Enough meanings for efficiency.



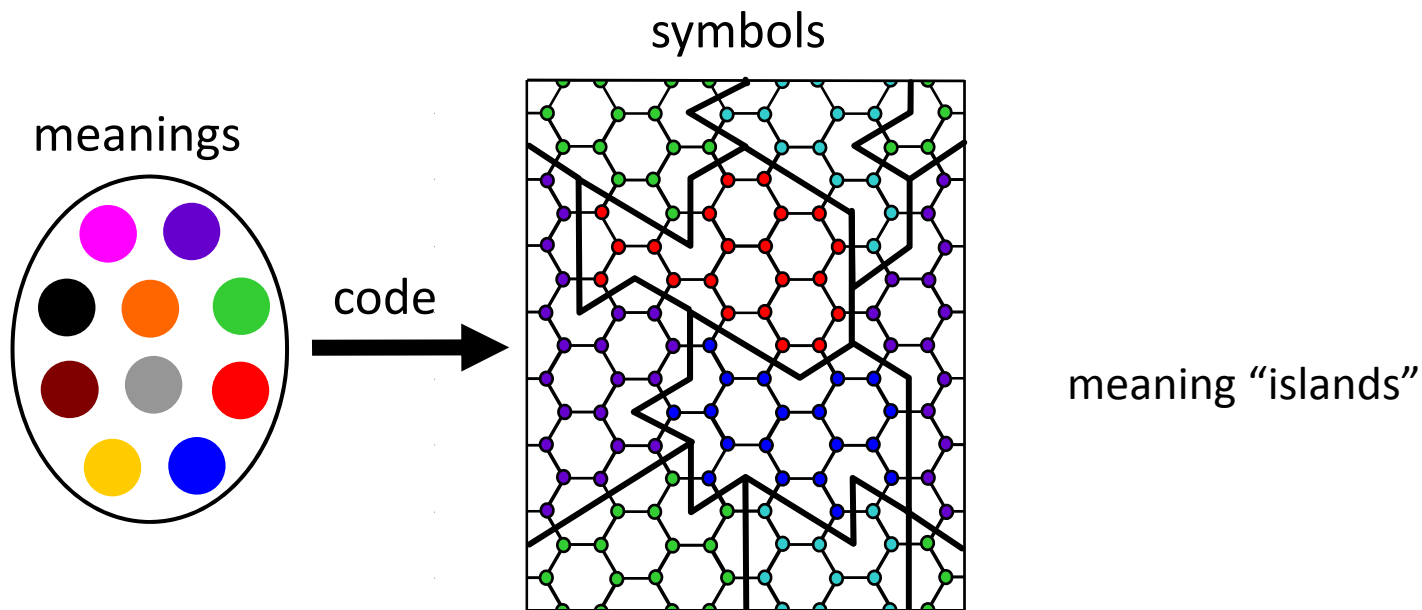
David Goodsell

# Molecular code is a noisy communication channel



- Molecular reading is inherently noisy.
- Symbol space is a **graph**: vertices = symbols,  
edges between symbols confused by misreading.
- **Code** = coloring of symbol space.
- Code emerges when number of encoded meanings  $n_f > 1$ .

# Molecular code is a partition of symbol space



- Coloring partitions symbol space into “islands” of meanings.
- Islands separated by self-avoiding random walks = polymers net.
- To be seen:

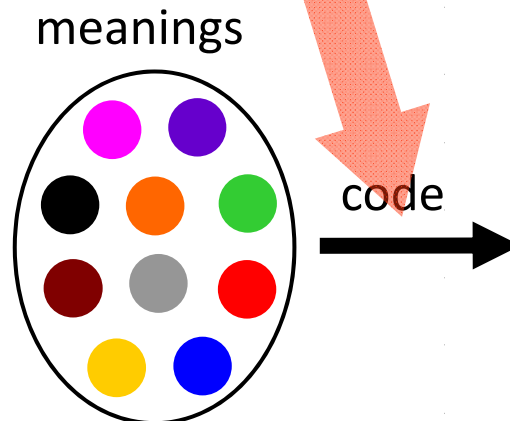
**code fitness = Free energy of polymer net.**

# Partition determines fitness of noisy codes

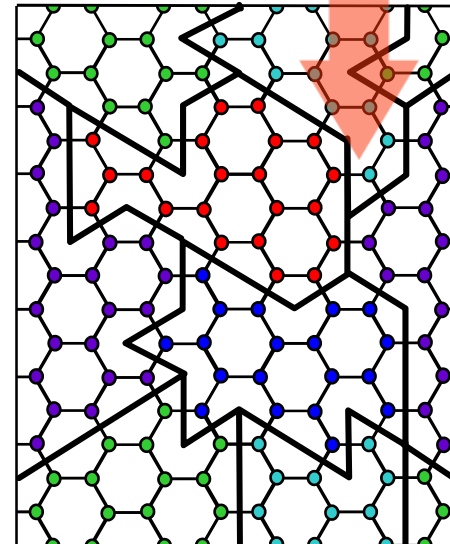
- $\text{Fitness} = \text{Error-load} + \text{Diversity} + \text{Cost}$ .
- Two sources for **noise**:

Noisy mapping (Cost)

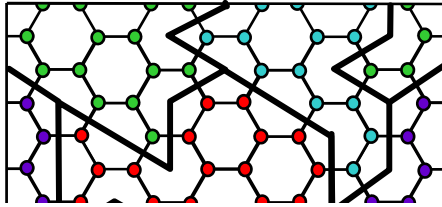
Misreading = confusion of  
read symbols (Error-load)



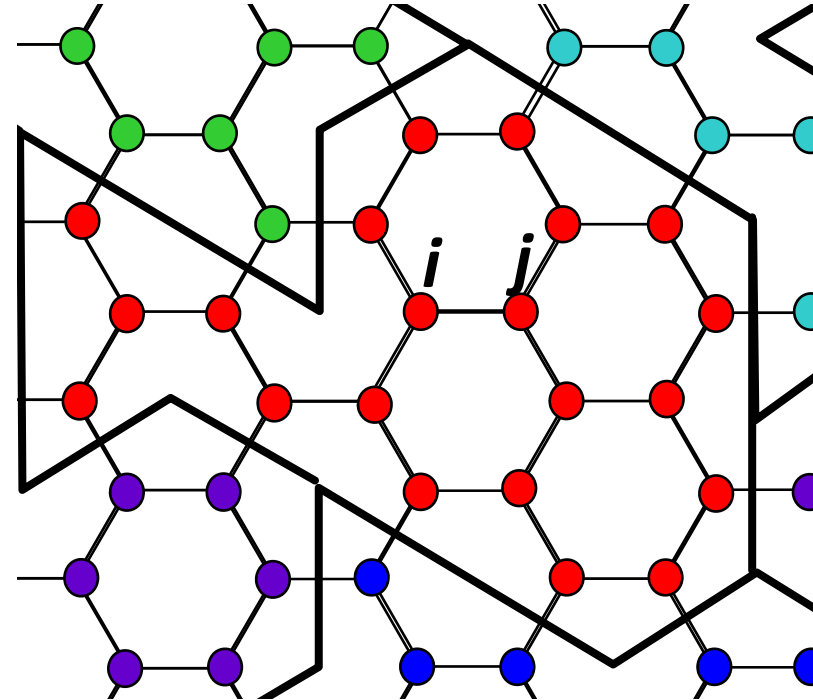
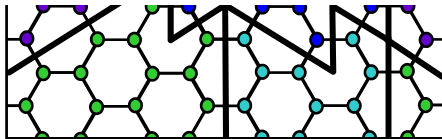
symbols



# Boundary network determines error-load



$$\text{Error-load} = \sum_{\text{edges}} r_{ij} E_{ij} = \sum_{\text{boundary}} r_{ij}$$

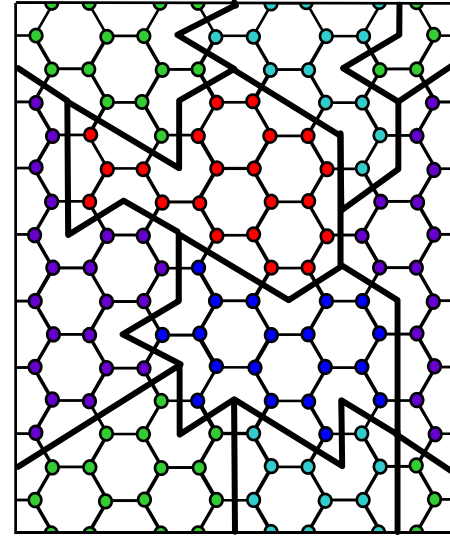


- Boundary network specified by  $E_{ij}$  (= 1 if  $i, j$  code different meanings).
- Error-load = average chance to cross the boundaries by misreading.
- Probability of misreading  $i$  as  $j$  is given by the matrix  $r_{ij}$ .

# Diversity counteracts error-load

Diversity = # islands =  $n_f$

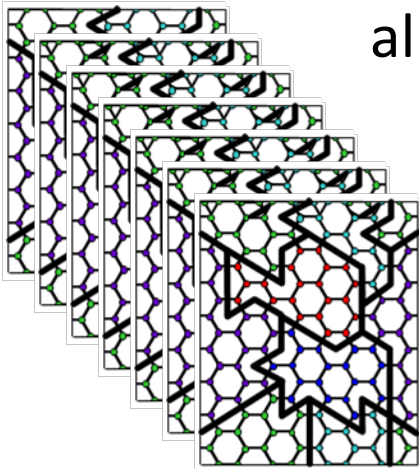
$$H_E = \sum_{edges} r_{ij} E_{ij} - w_D n_f.$$



- Approximated diversity = number of encoded meanings,  $n_f$
- $\max(\text{Diversity}) \rightarrow \text{more islands}$ ,  $\min(\text{Error-load}) \rightarrow \text{less islands}$ .
- Quality  $H_E$  combines interplay of these two conflicting needs.
- $w_D$  measures relative significance error-load/diversity.

# Molecular codes cost chemical specificity

- Molecular codes are implemented by recognition interactions.
- Diverse meanings requires **specificity** = high binding energies  $E_b$ .
- Cost  $\sim$  average binding site size  $\sim$  average binding energy  $\langle E_b \rangle$ .
- Binding probability  $\sim$  Boltzmann:  $P_b \sim \exp(E_b/T)$ .
- Specificity cost is measured by entropy of mapping  $-S_c = \langle \ln P_b \rangle$ .
- Cost is entropy of ensemble of all possible mappings:



all possible partitions and colorings of every partition.

# Code's fitness combines quality and cost

$$\text{Fitness} = \text{Quality} + w_C \times \text{Cost}$$

$$F_C = \langle H_E \rangle - w_C \times S_C$$

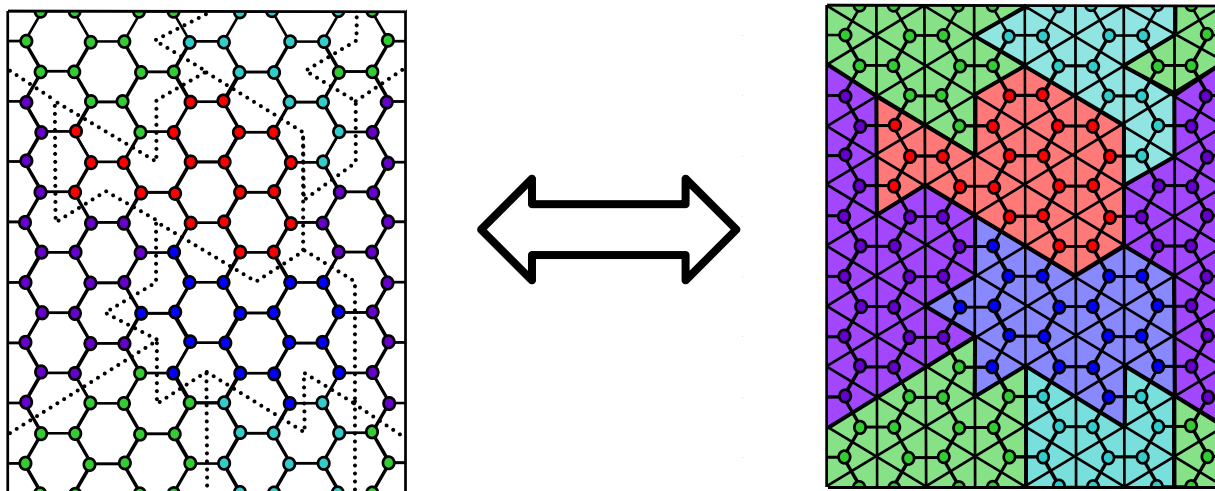
- Parameter  $w_C$  measures significance cost/quality.
- High  $w_C \rightarrow$  fuzzy codes, low  $w_C \rightarrow$  sharp codes.
- Organism complexity/environment richness  $\rightarrow$  low  $w_C$ .
- Thermodynamic analogy:

Quality  $H \leftrightarrow$  Energy, Cost  $\leftrightarrow$  entropy,  $w_C \leftrightarrow$  temperature,

- Code fitness is **free energy**,  $F_C$ :

$$Z_C = e^{-F_C/w_C} = \sum_{\text{mappings}} e^{-H_E/w_C}$$

# Code fitness is free energy of polymer network



- Code fitness = free energy of “polymers” on the dual graph.
- Monomers pass along edges  $E_{ij} = 1$  and vertices  $V_i = 1$ .
- # encoded meanings  $n_f$ , # vertices  $n_v$ , # edges  $n_e$  related by

$$n_f = C + n_e - n_v = C + \sum_{\text{edges}} E_{ij} - \sum_{\text{vertices}} V_i$$

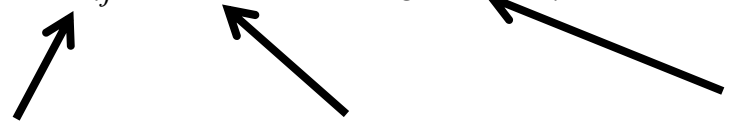
# Fitness is free energy of polymer network

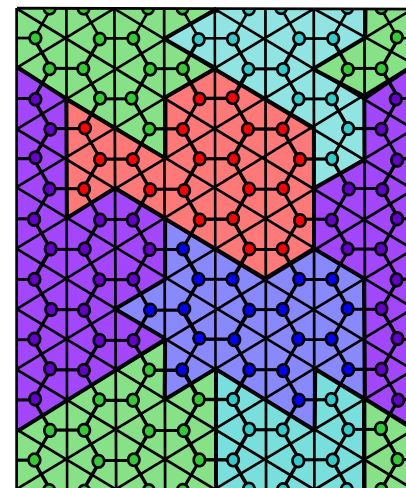
- From the polymer description

$$Z_C = e^{-F_C/w_C} = \sum_{\text{networks}} \left( \prod_{\text{edges}} e^{-\beta E_{ij}/w_C} \prod_{\text{vertices}} e^{-\alpha V_i/w_C} \right)$$

- $\alpha, \beta$  are excitation energies of vertex and edge:

$$\beta = (r_{ij} - w_D) - w_C \ln n_m, \quad \alpha = w_D + w_C \ln n_m$$


  
 misreading      diversity      ln(#meanings)



- Summation is hard (vertices and edges are not independent).

# Spin networks enable calculation of fitness

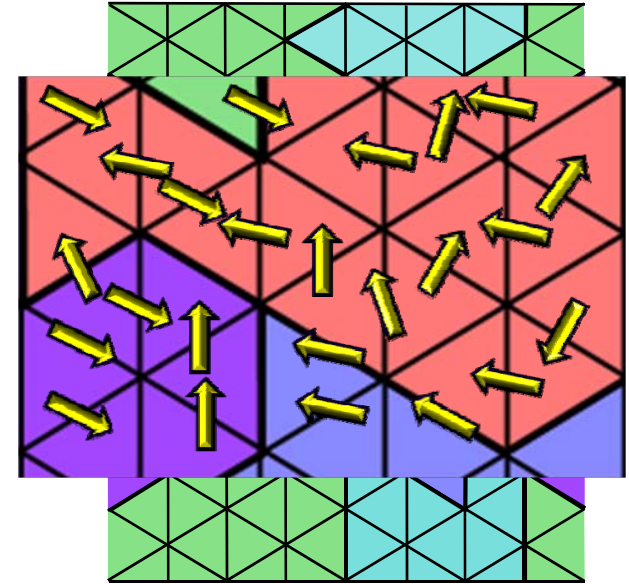
- Mathematical tool:  $n = 0$  formalism

[de Gennes, networks: Zilman & Safran]

- Assign magnetic spin  $S_{ij}$  to each edge.
- Spin interaction  $H_S$  and spin free energy,

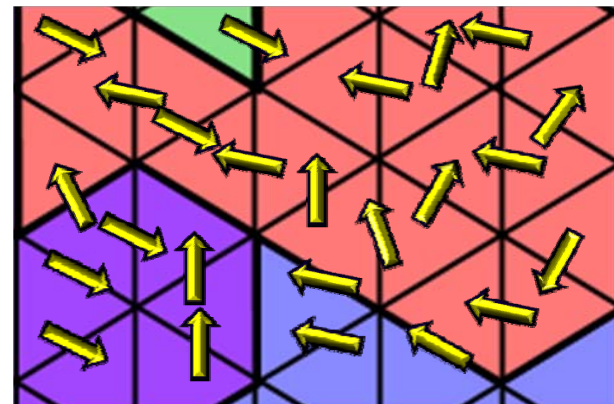
$$Z_S = \exp(-F_S) = \langle \exp(-H_S) \rangle.$$

- When  $n \rightarrow 0$ : The only non-vanishing contributions to  $F_S$  correspond to a configuration of the polymer network,
- As a result:  $F_S = F_C = \text{code fitness}$ .
- Spin free energy  $F_S$  is relatively easy to calculate (mean-field).



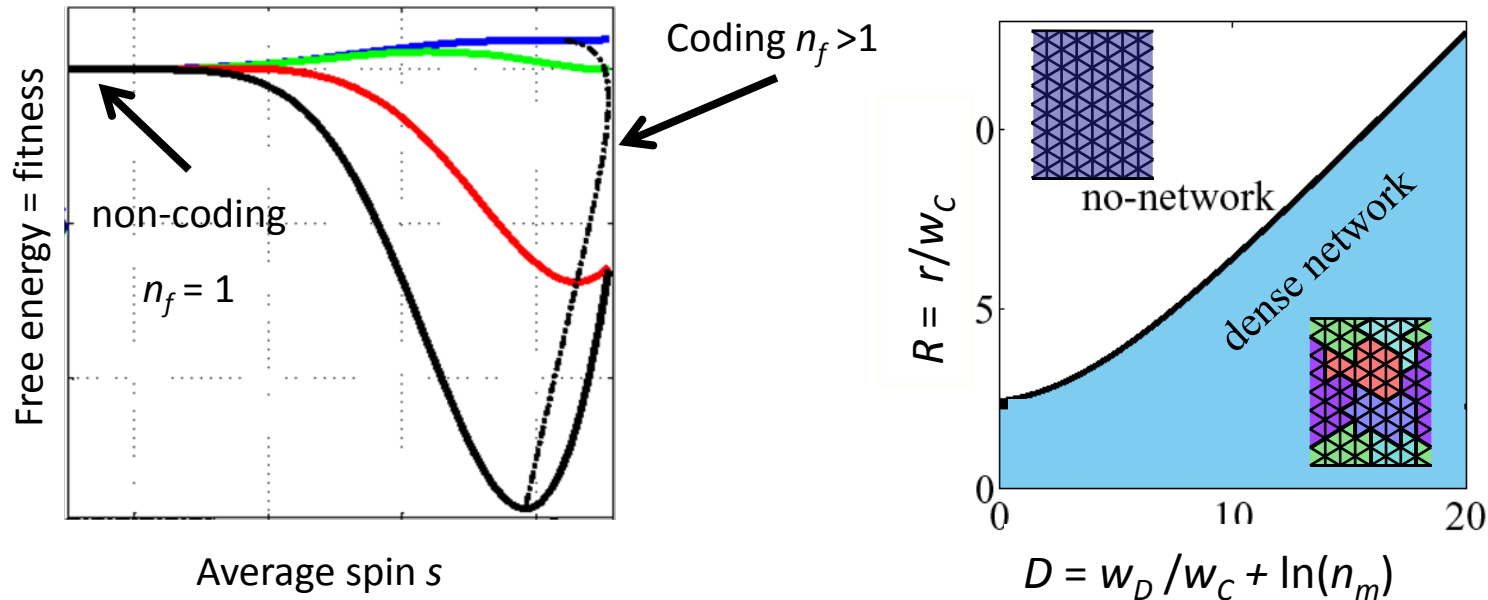
(further details: PNAS 2008)

# Code emerges at a phase transition



- Using equivalence to spin networks we trace the evolution of the coding system.
- Control parameters,  $w_D$ ,  $w_C$ ,  $r_{ij}$ ,  $n_m$ 
  - combined into normalized diversity  $D = w_D / w_C + \ln(n_m)$ ,
  - normalized misreading  $R_{ij} = r_{ij} / w_C$ .
- Find average mapping as a function of  $D$ ,  $R_{ij}$ .
- A high-dimensional problem (dimension = number of edges).
- Symmetry reduces the dimension of code space (for regular graphs code-space is 1D).

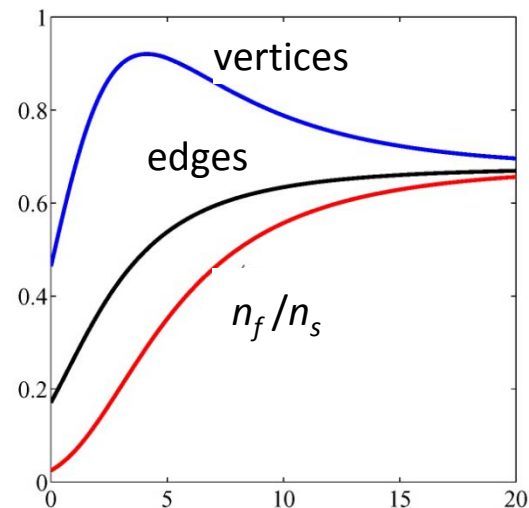
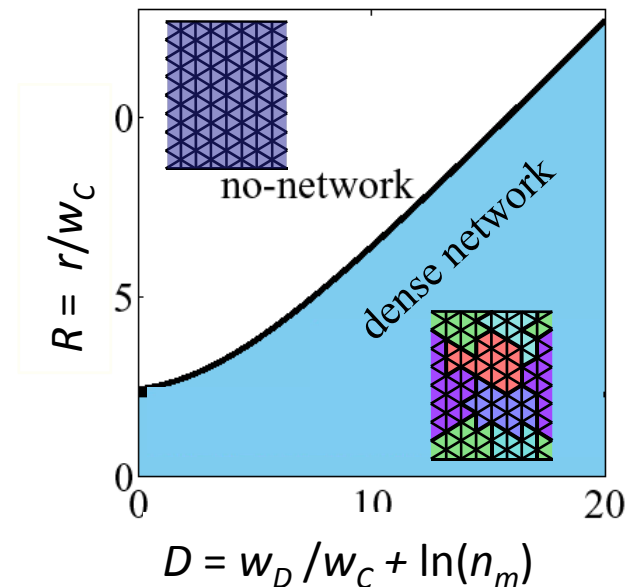
# Code emerges at a phase transition



- Non-coding state  $s = 0$ ,  $n_f = 1$ . Code conveys no information.
- Coding state  $n_f > 1$  emerges. Code transmits  $\log_2 n_f$  bits/symbol.
- Abrupt jump (1<sup>st</sup>-order transition) non-coding  $\rightarrow$  coding.
- Emergent network is dense and highly connected.

# Dense coding network emerges

- At the transition many meanings appear together.
- Meanings per symbol  $n_f/n_s > 0$ .
- Pathways towards transition:
  - Increasing  $n_m$ ,
  - Increasing diversity  $w_D$ ,
  - Decreasing misreading  $r$ .



# Survival of the fittest code

- Population of “organisms” that compete and evolve according to code fitness,  $F_C$ .
- Population dynamics:

Min(error-load)

Min (cost)

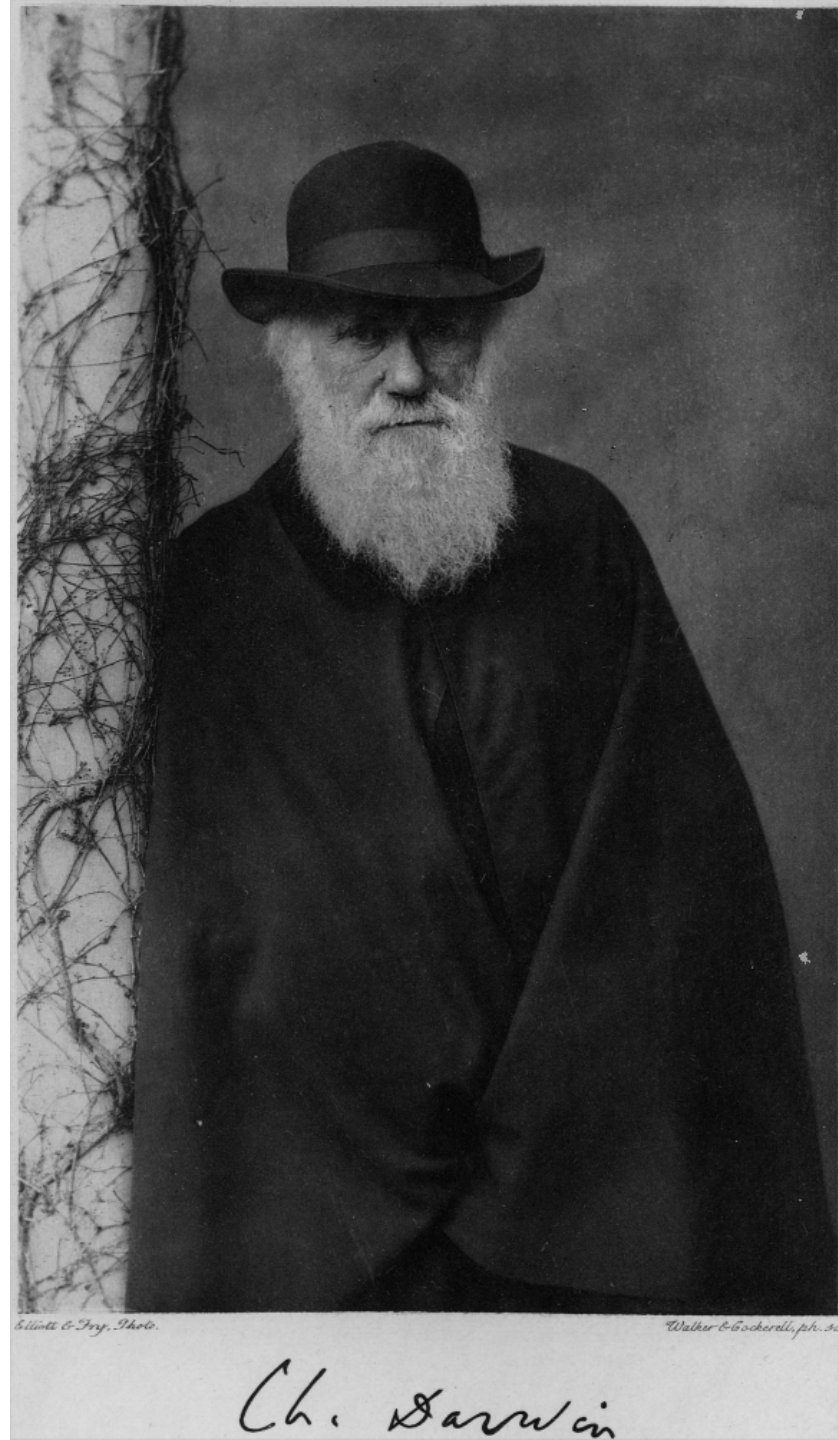
Max(diversity)

+

Mutation

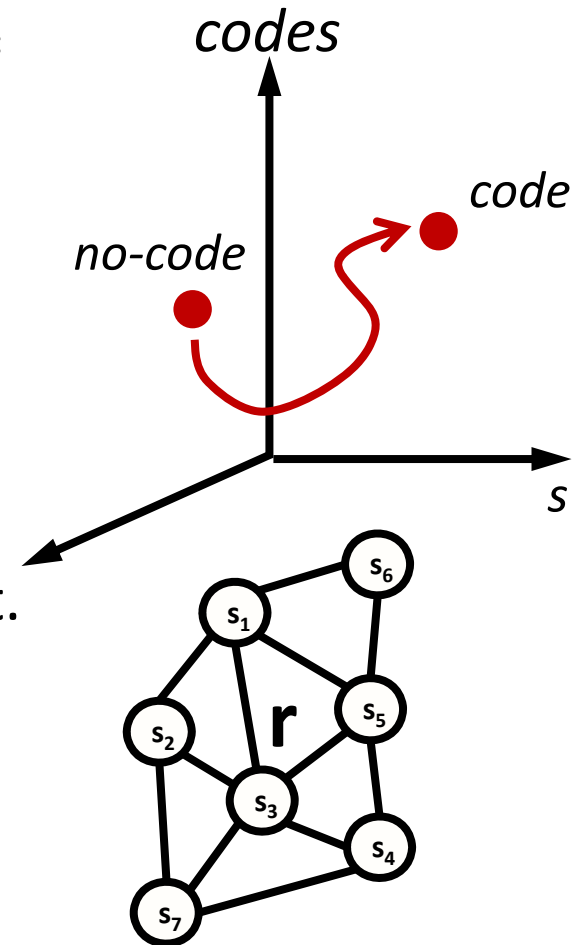
Selection

Random drift

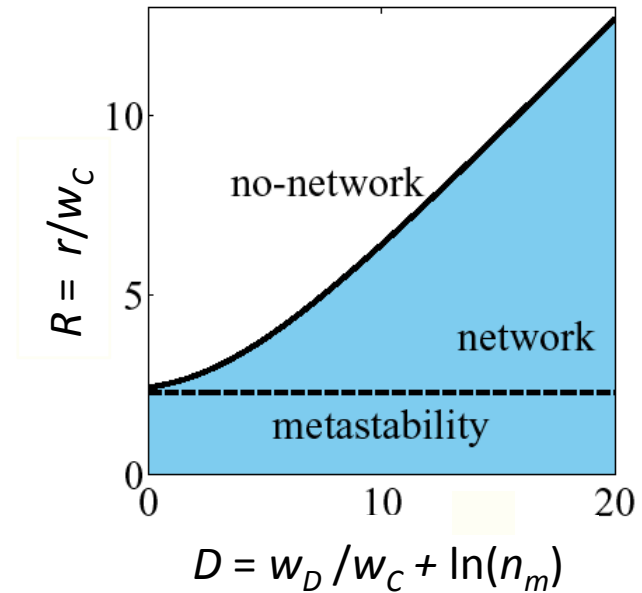
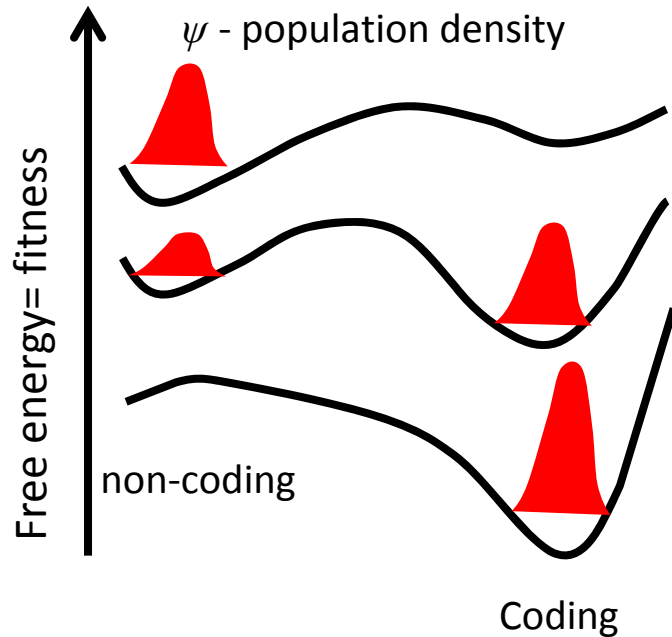


# Evolution in code fitness landscape

- The population evolves in the space of all possible codes.
- Dimension of code fitness landscape =  
# misreading edges = # spins.
- Population of organisms compete by  
the code fitness (density  $\Psi(s)$ ).
- Evolutionary dynamics effects:  
Local trapping, mutations, genetic drift.



# Trapping in metastable non-coding states



- Non-coding state may remain locally stable.
- Metastability exists as long as curvature is positive.
- Metastable network is dilute.

# Mutations smear population in code space

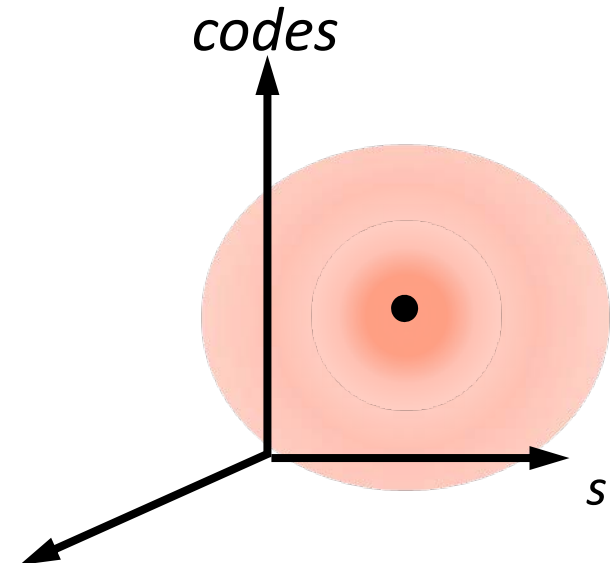
- Mutations drive organisms to diffuse to suboptimal codes.
- Reaction-diffusion dynamics

$$\frac{\partial \Psi(s)}{\partial t} = -F_C(s)\Psi(s) + \mu \nabla^2 \Psi(s)$$

( $\mu$  - mutation rate,  $\psi$  - population density)

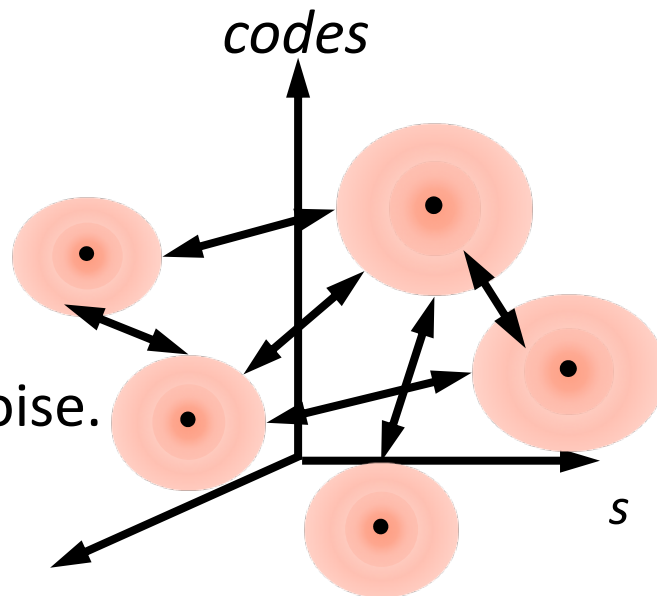
- Reach Gaussian steady-state:

$$\Psi \sim \exp\left(-\mu^{-1/2}(s - s_0)^2\right)$$



# Genetic drift: diffusion between optimal codes

- Genetic drift = reproduction fluctuations = Noise.
- The population migrates between many possible optima.
- At steady state  $P(F) \sim \exp(-F/T)$  [Sella and Hirsh]  
with “temperature”  $\sim 1/(\text{population size})$ .
- Small populations are “hotter”.
- Populations effects add  
external noise to internal coding noise.



# Summary – The physical language of molecules

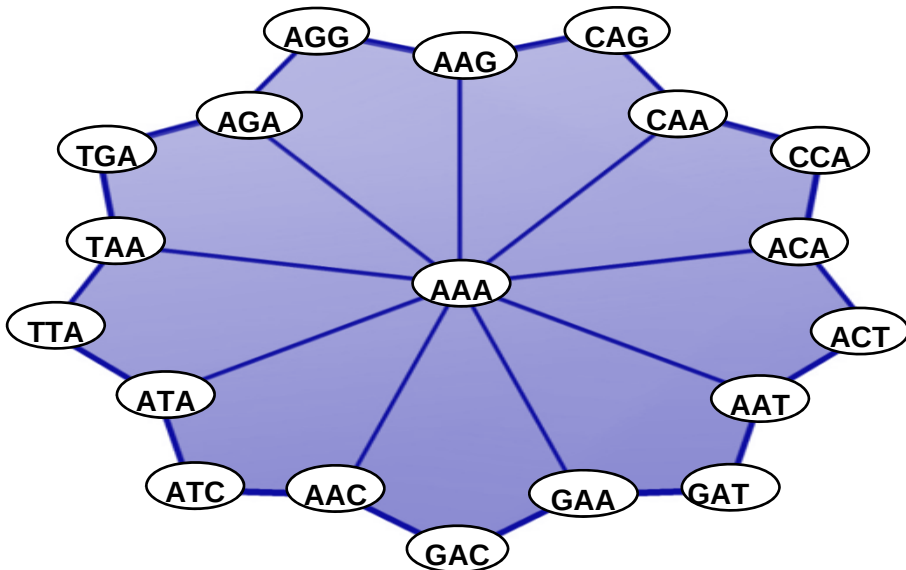
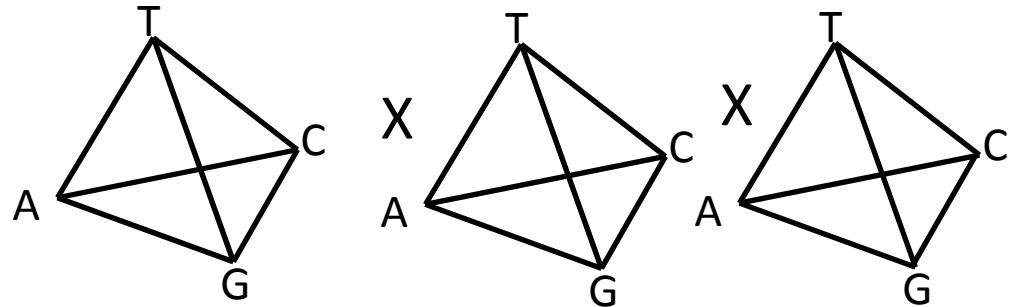
- Molecular codes **map** symbols to meanings by recognition.
- Code **fitness** = error-load + cost + diversity of mapping.
- Fitness = free energy of **polymer nets** (= spin nets).
- Evolutionary dynamics: Code emerges at a transition.
- Simple physical description of evolving molecular information channels. A potential tool to study noisy biological codes (Transcription regulation network, genetic code, operons...)

**Additional slides...**

# The probable errors define the graph and the topology of the genetic code

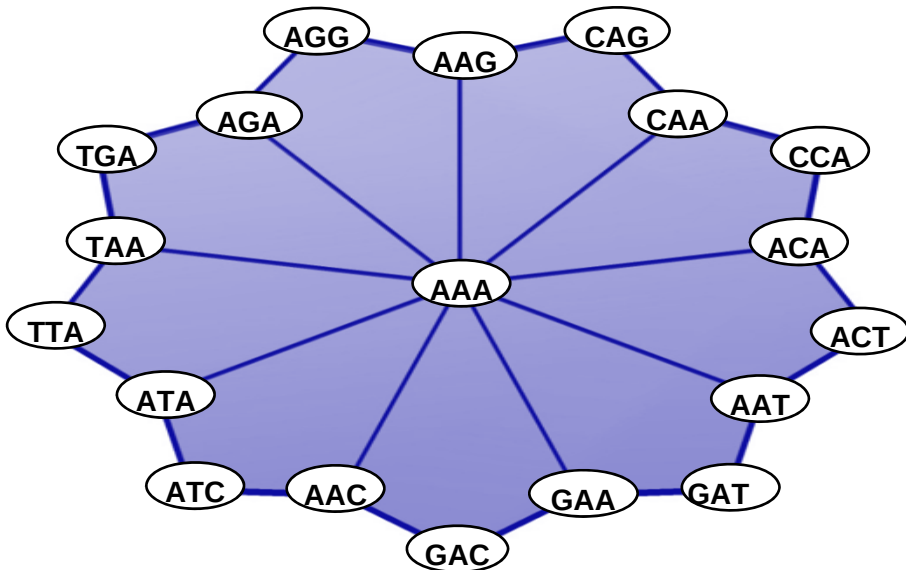
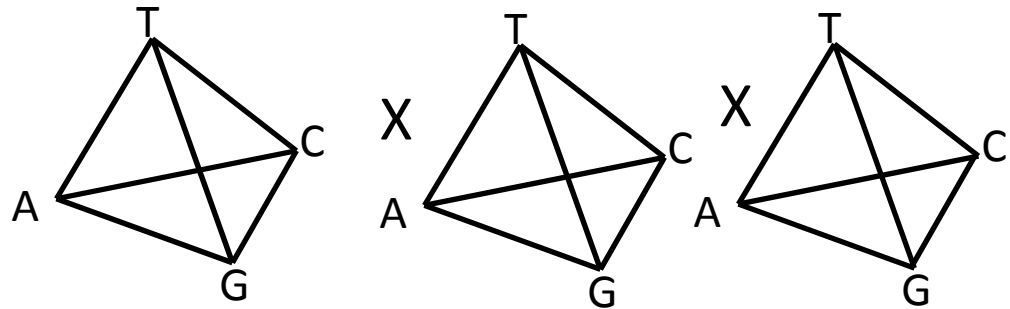
- Symbol (codon) Graph = codon vertices + one-letter difference edges ( Hamming = 1 )

$$K_4 \times K_4 \times K_4$$



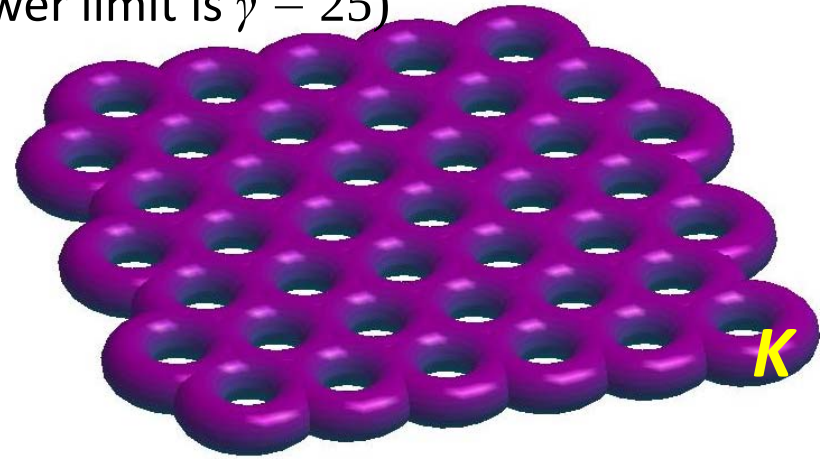
# The surface of the code graph is holey

$$K_4 \times K_4 \times K_4$$



*Holey graph:  $\gamma = 41$*

(lower limit is  $\gamma = 25$ )



# Coloring number is the upper limit for the number of smooth islands

- What is the minimal number of colors required for a map so that no two adjacent countries have the same color?
- Coloring number is a topological invariant and a function of the genus,  $chr(\gamma) = \left\lfloor \frac{1}{2} \left( 7 + \sqrt{1 + 48\gamma} \right) \right\rfloor$ .

# of meanings =  $chr(\gamma)$

|    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|
| 4  | 7  | 8  | 9  | 10 | 11 | 12 | 12 | 13 | 13 |
| 14 | 15 | 15 | 16 | 16 | 16 | 17 | 17 | 18 | 18 |
| 19 | 19 | 19 | 20 | 20 | 20 | 21 | 21 | 21 | 22 |
| 22 | 22 | 23 | 23 | 23 | 24 | 24 | 24 | 24 | 25 |
| 25 | 25 | 25 | 26 | 26 | 26 | 27 | 27 | 27 | 27 |

# Other molecular codes:

## Transcription regulatory network:

- Controls gene expression via binding proteins to DNA.
- Mapping between proteins and DNA is.
- Number of proteins is limited by the coloring number.

## Logic design of operons:

- Logic gates made of binding proteins are smooth.

# A sketch for an “experiment”: 2X2 coding system

- 2 binding sites (symbols).
- 2 transcription factors (meanings) after duplication.
- Control gain by environment  $A(t)$ .
- Coding transition when 2<sup>nd</sup> factor becomes advantageous.

