An efficient algorithm to compute the minimum free energy of interacting nucleic acid strands

Ahmed Shalaby

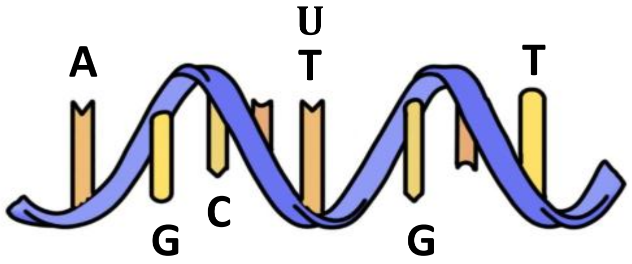
3rd year PhD

Supervisor: Damien Woods

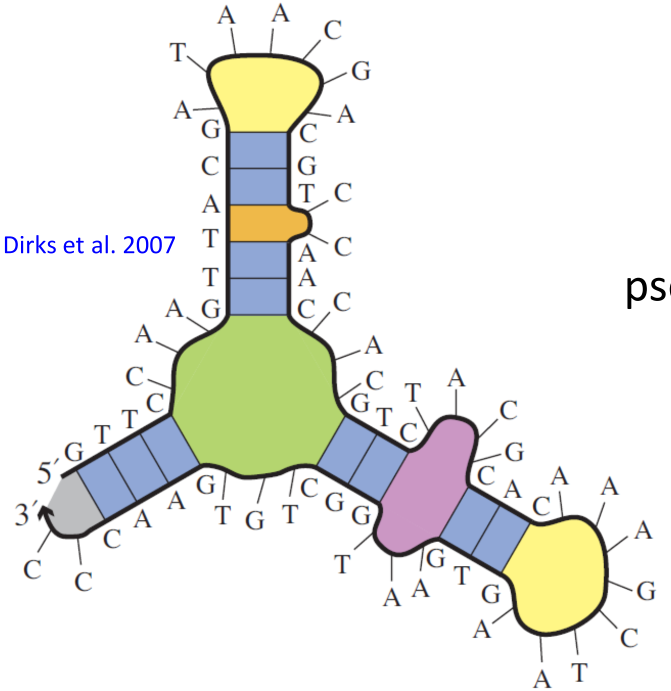
European Innovation Council

Funded by the European Union

Secondary structure



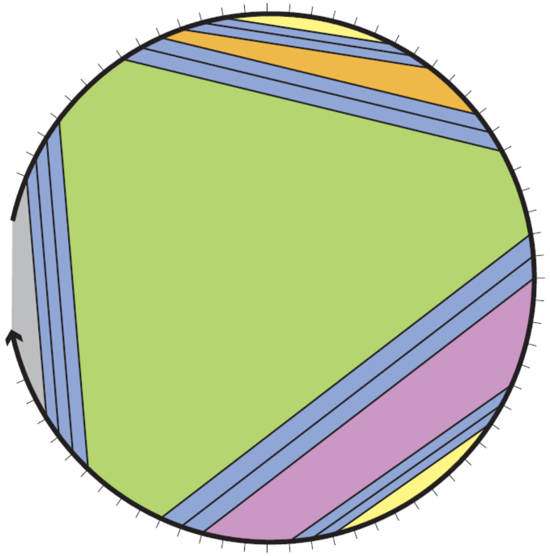
Single stranded **DNA/RNA**



Dirks et al. 2007

Secondary structure

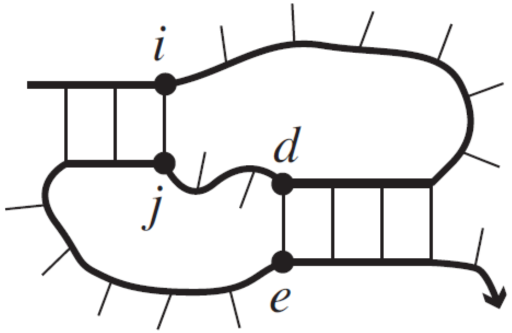
pseudoknot-free



Dirks et al. 2007

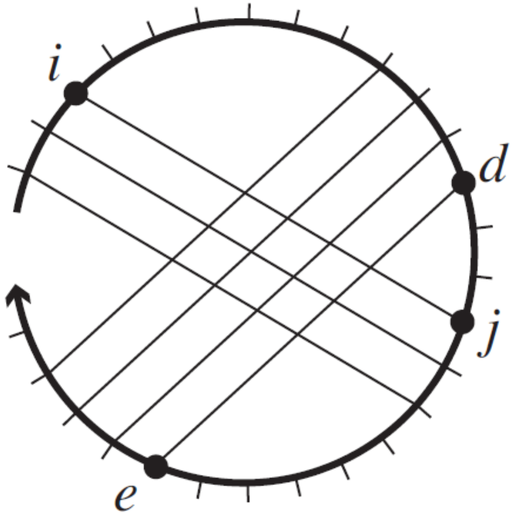
Polymer graph representation

NP – Complete

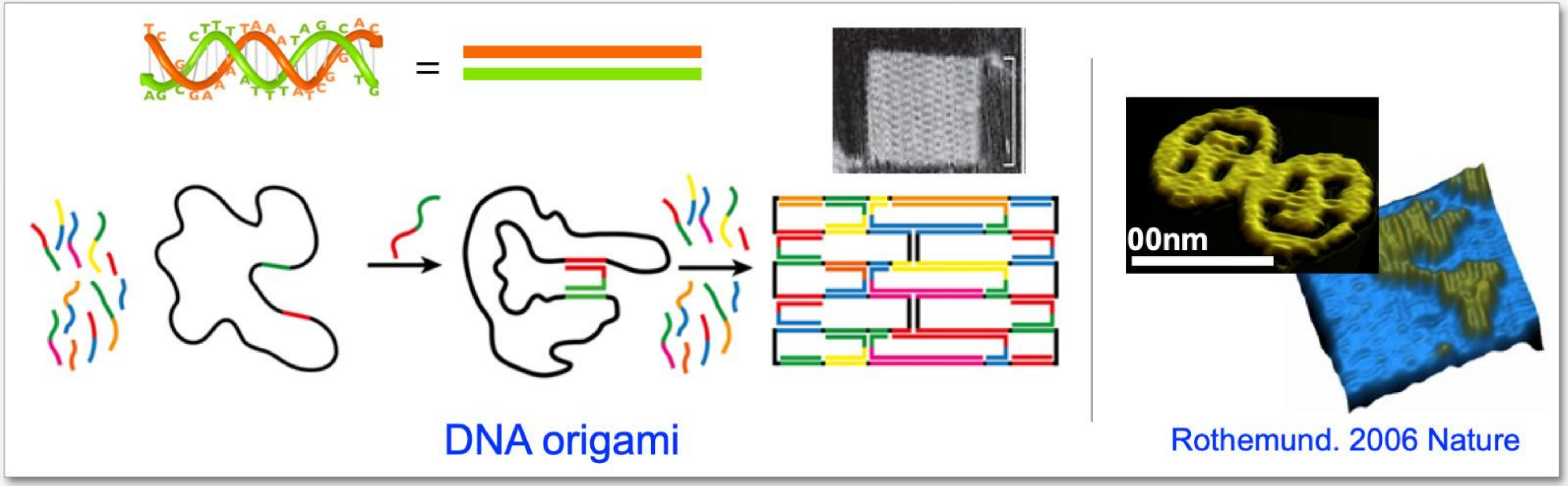


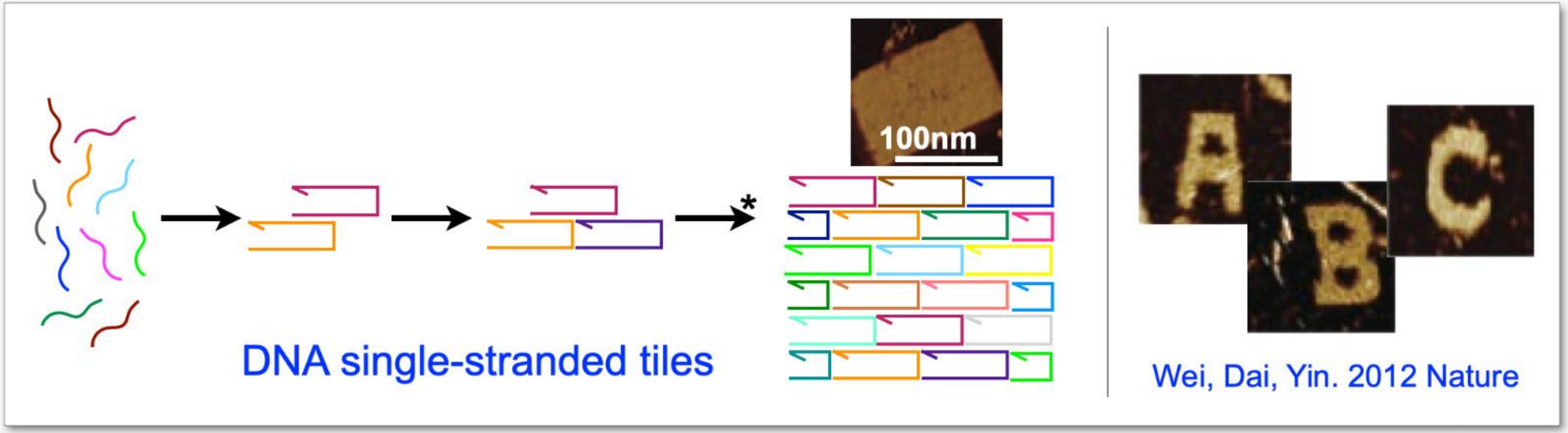
Dirks et al. 2007

pseudoknotted

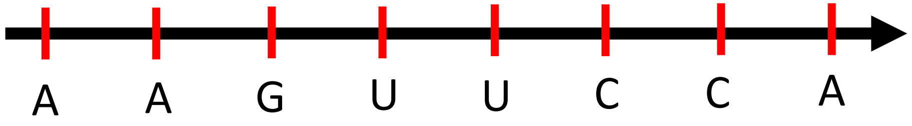


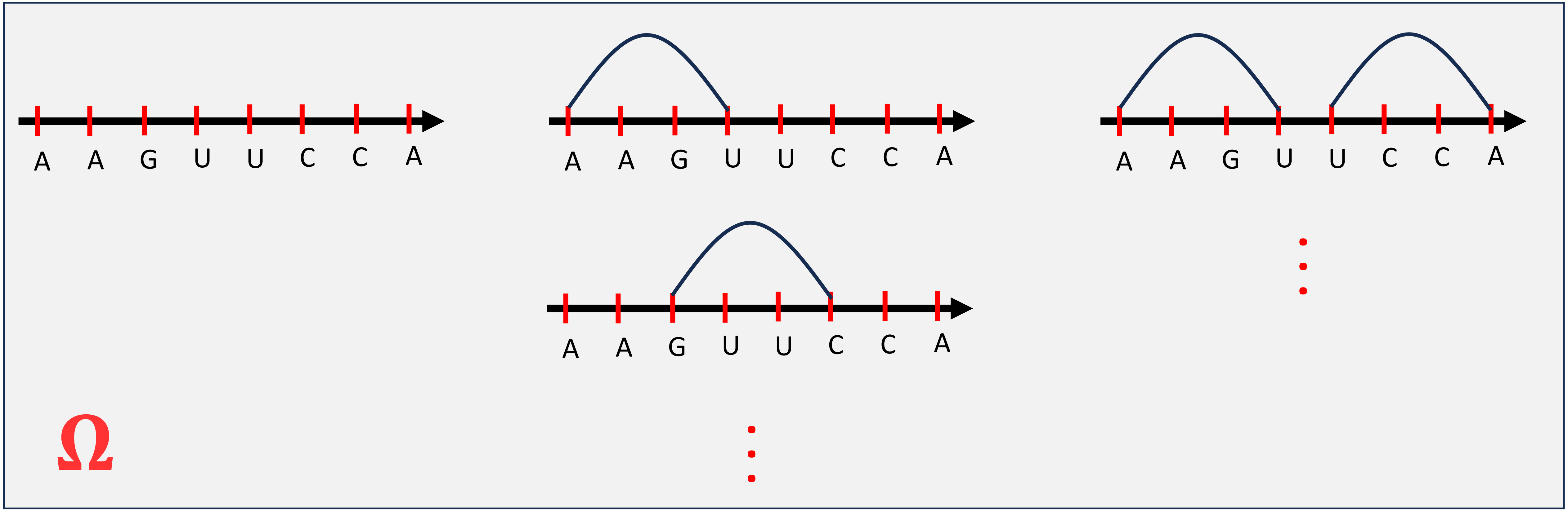
Dirks et al. 2007


00nm
DNA origami
Rothemund. 2006 Nature


100nm
DNA single-stranded tiles
Wei, Dai, Yin. 2012 Nature

Secondary structure

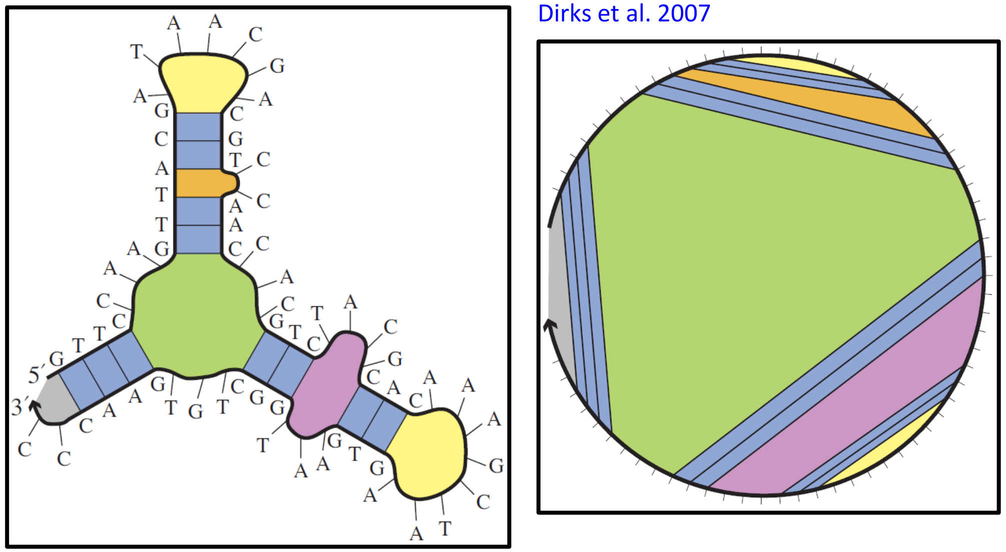




Which secondary structure is more favourable?

Energy models, Minimum Free Energy and Partition Function

Single stranded system



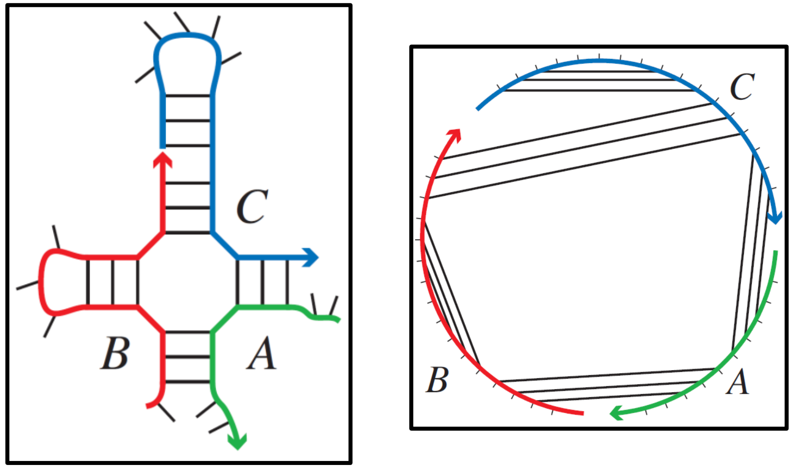
Dirks et al. 2007

$\boldsymbol{\Delta G(S)}$

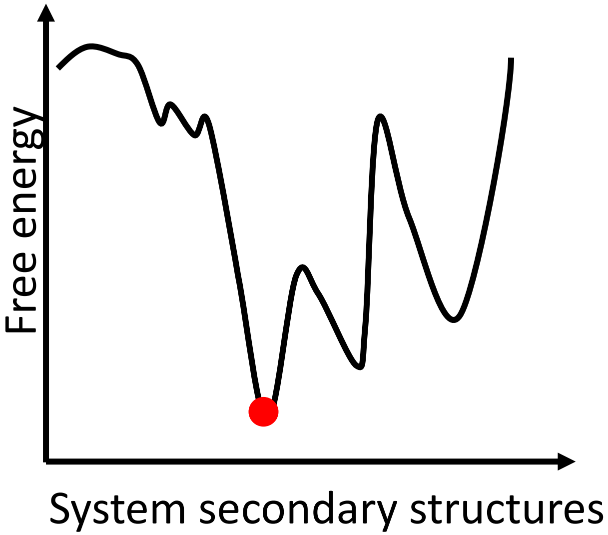
Energy model

Capture the free energy of secondary structure

Multi stranded system of c strands



Dirks et al. 2007



$$\text{MFE} = \min_{S \in \Omega} \Delta G(S)$$

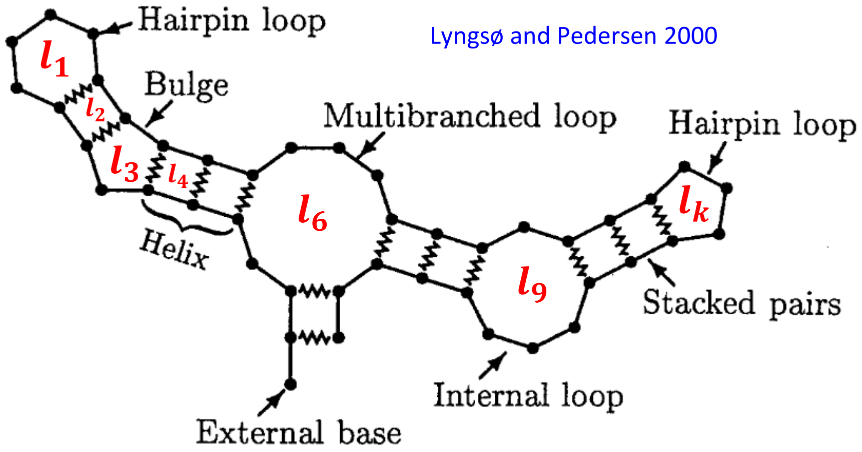
Minimum Free Energy

Boltzmann weighted sum

$$Q = \sum_{S \in \Omega} e^{-\Delta G(S)/k_{\text{B}}T} \Rightarrow \Pr[S] = \frac{e^{-\Delta G(S)/k_{\text{B}}T}}{Q}$$

Partition Function

Energy model: Nearest neighbour model

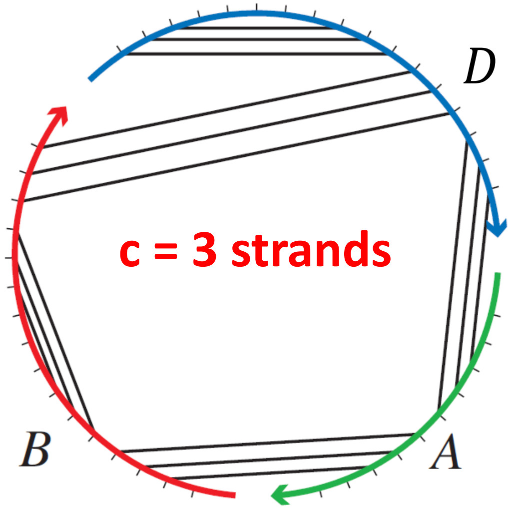


Lyngsø and Pedersen 2000

$$\boldsymbol{\Delta G(S) = \Delta G(l_1) + \Delta G(l_2) + \cdots + \Delta G(l_k)}$$

$$\Delta G(S) = \sum_{l \in S} \Delta G(l)$$

Dirks et al. 2007



unique strands

$$\Delta G(S) = \sum_{l \in S} \Delta G(l) + (c-1)\,\Delta G^{\text{assoc}}$$

$$\min_{S \in \Omega} \Delta G(S)$$

Ω: the set of all secondary structures

Energy model: Nearest neighbour model (allowing repeated strands)

Free energy − Loop energy + Entropic association cost + Symmetry penalty

$$\Delta G(S) = \sum_{l \in S} \Delta G(l) + (c-1) * \Delta G^{assoc} + k_B T * logR$$

c= 4 strands

Dirks et al. 2007

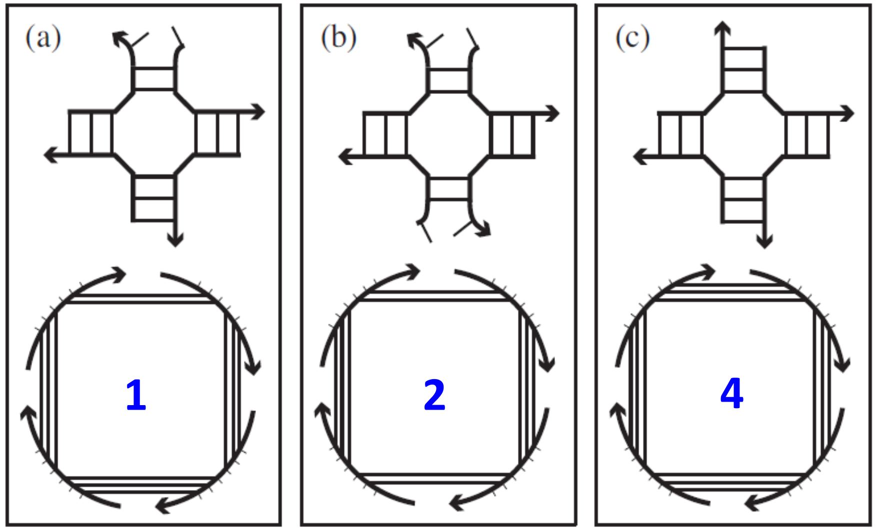


Fig. 2.2 *Sample secondary structures and polymer graphs for a complex of four indistinguishable strands.* (a) *1-fold (i.e., no) rotational symmetry.* (b) *2-fold rotational symmetry.* (c) *4-fold rotational symmetry.*

$$\min_{S \in \Omega} \Delta G(S)$$

Ω: the set of all connected structures

R: degree of rotational symmetry

Computational complexity of Minimum Free Energy algorithms

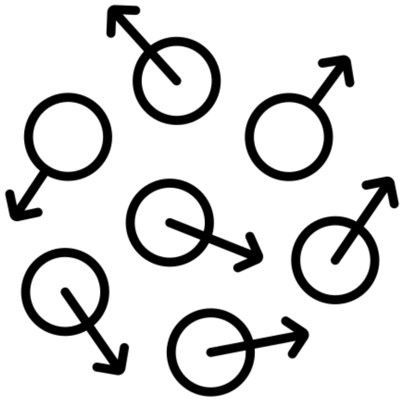
Input Type		MFE
Single Strand	Nussinov et al. 1980	$O(N^4)$
Multiple **unique** Strands, Bounded ($\leq$ c)	Dirkis er al. 2007	$O(N^4(c-1)!)$
Multiple Strands **allowing repeats**, Bounded ($\leq$ c)		?
Multiple Strands, Unbounded	Condon et al. 2021	$NP-$Complete

N bases, c strands

Open problem for $\approx$ 20 years

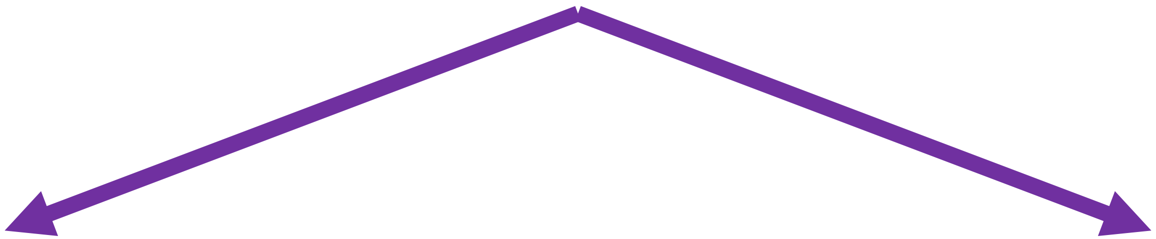
Why symmetry makes that difference?

Entropy



ΔG

Free energy

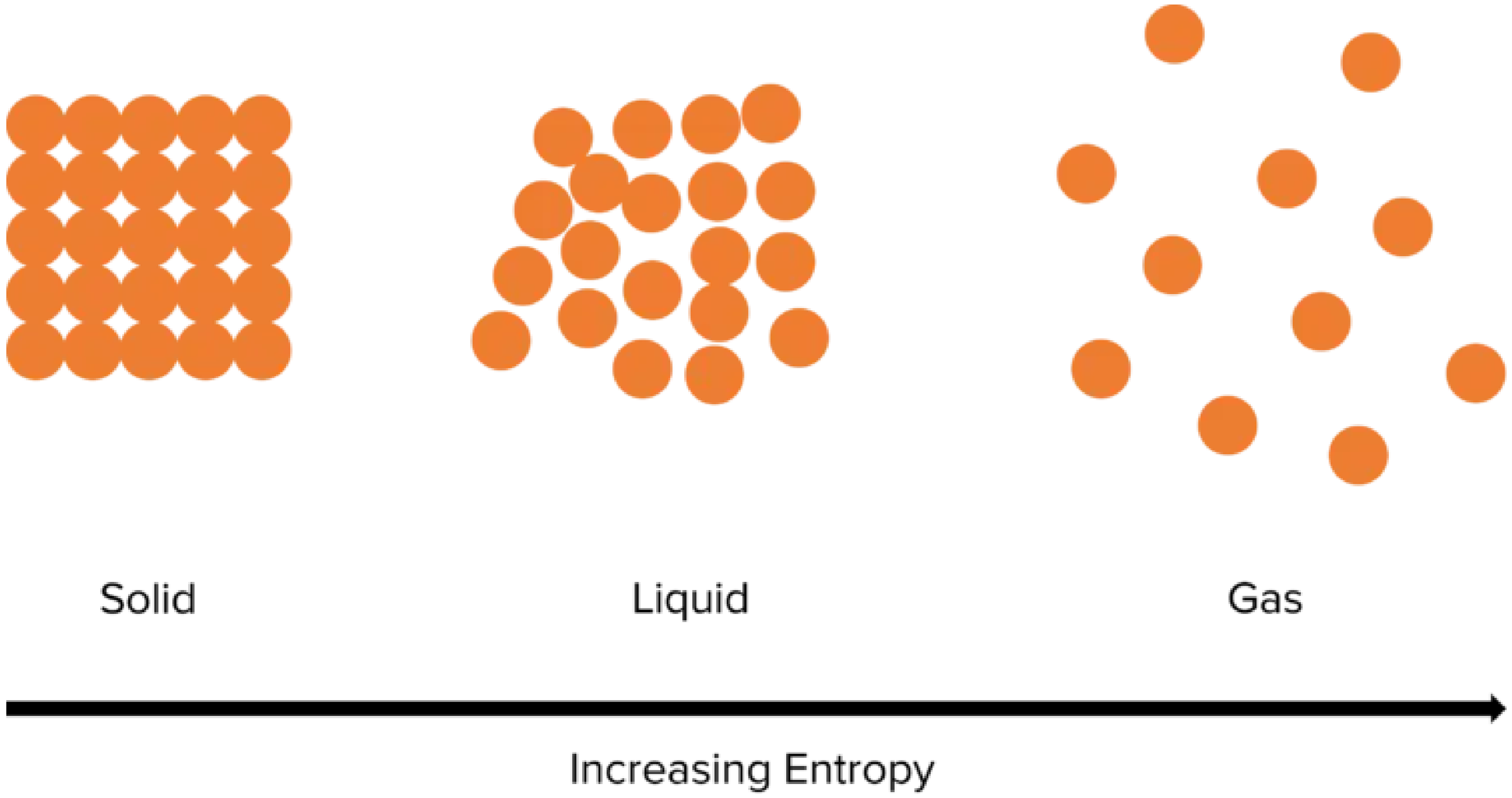


Enthalpy

H

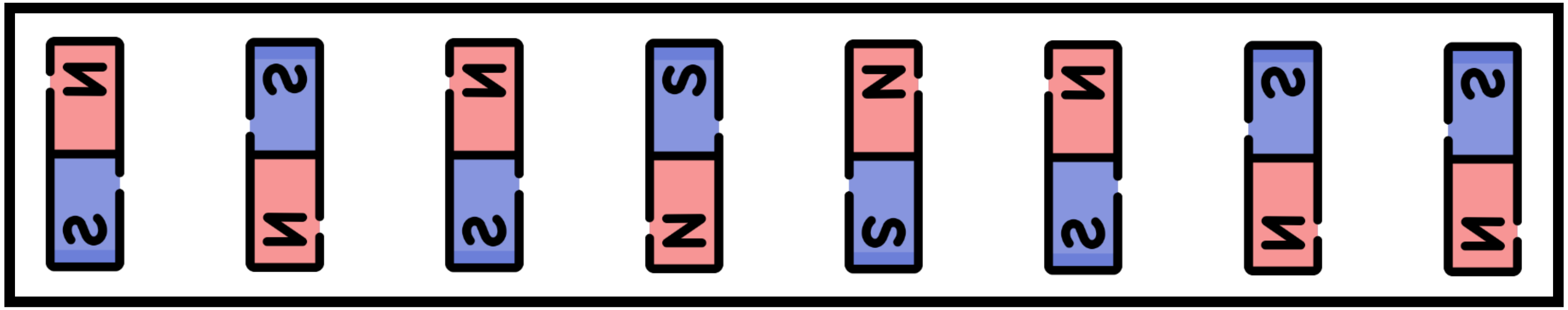
Entropy

S


Solid
Liquid
Gas
Increasing Entropy

$$S = k_B \log \Pi$$

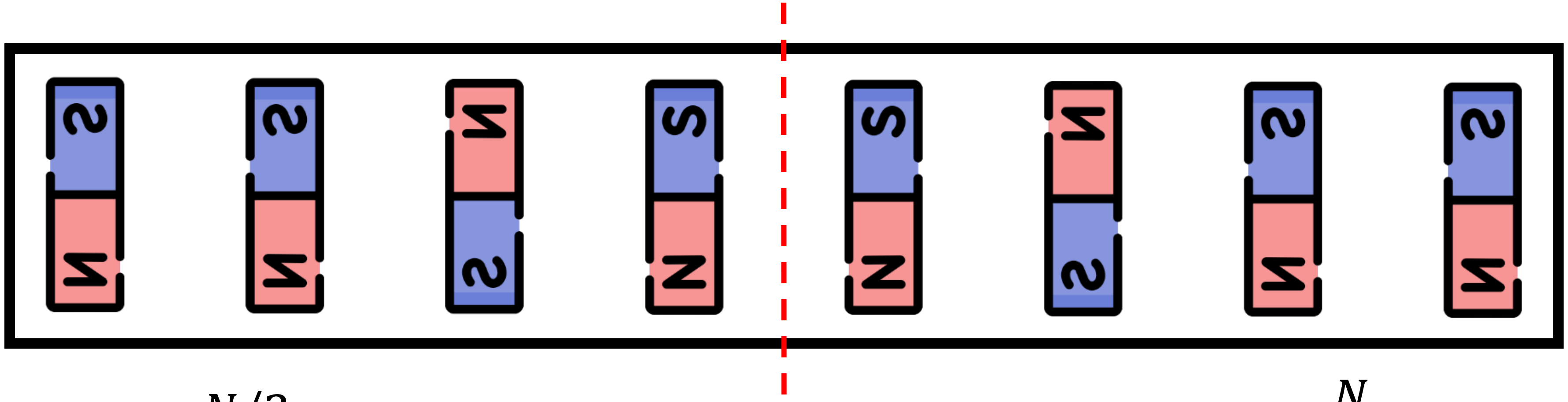
$\boldsymbol{N}$



The total number of states of the N magnets is $\Pi = 2^N$

$$\mathrm{S} = k_B N \log 2$$

$\boldsymbol{\frac{N}{2}}$



$\boldsymbol{\frac{N}{2}}$

$\Pi = 2^{N/2}$ $\longrightarrow$ $\mathrm{S} = k_B \frac{N}{2} \log 2$

Free energy — Loop energy — Entropic association cost — Symmetry penalty

$$\underbrace{\Delta G(S)}_{\text{Free energy}} = \underbrace{\sum_{l \in S} \Delta G(l)}_{\text{Loop energy}} + \underbrace{(c-1) * \Delta G^{\mathrm{assoc}}}_{\text{Entropic association cost}} + \underbrace{k_B T * \log R}_{\text{Symmetry penalty}}$$

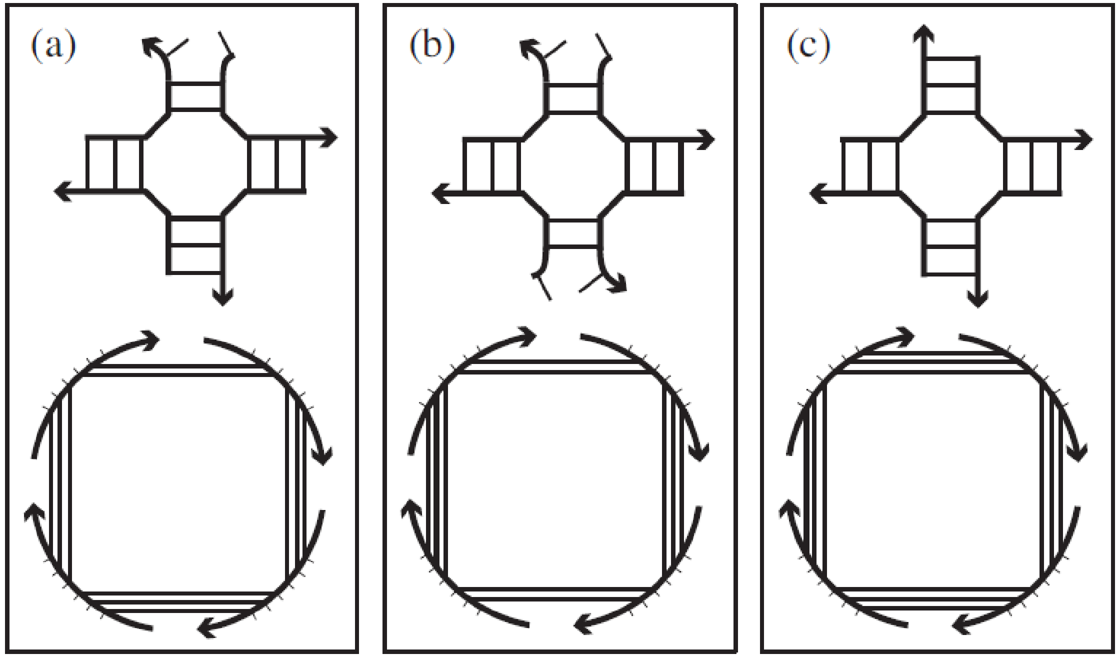


Fig. 2.2 *Sample secondary structures and polymer graphs for a complex of four indistinguishable strands.* (a) *1-fold (i.e., no) rotational symmetry.* (b) *2-fold rotational symmetry.* (c) *4-fold rotational symmetry.*

Why is this difficult?

Why symmetry makes it difficult?

Input Type	MFE
Single Strand	$O(N^4)$
Multiple **unique** Strands, Bounded ($\leq$ c)	$O(N^4(\mathrm{c}-1)!)$
Multiple Strands, Bounded ($\leq$ c)	?

} **Dynamic programming algorithms**

N bases, c strands

All of these are dynamic programming algorithms

Subproblems → **Big problem**

Level	Input Type	MFE
1	Single Strand (Maximum matching)	$O(N^3)$
2	Single Strand (Loop model)	$O(N^3)$
3	Multiple unique Strands, Bounded ($\leq c$)	$O\left(N^3(c-1)!\right)$
4	Multiple Strands, Bounded ($\leq c$)	?

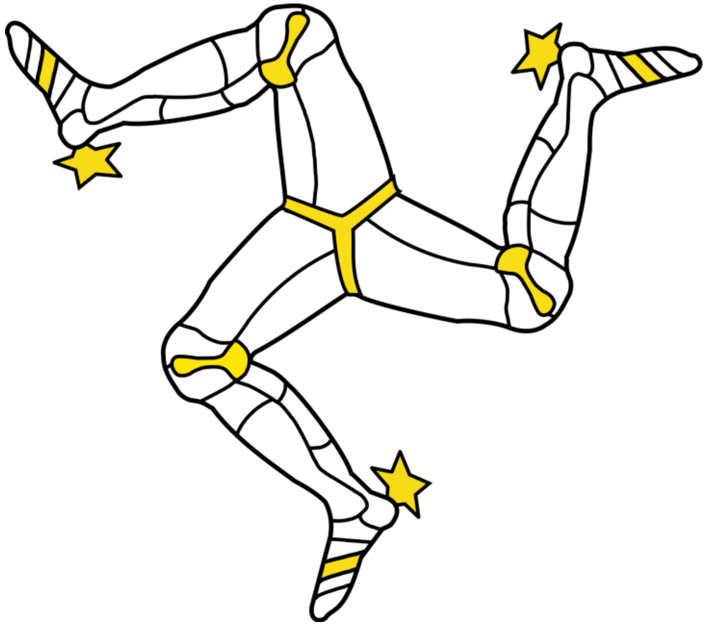
N bases, c strands

All of these are dynamic programming algorithms

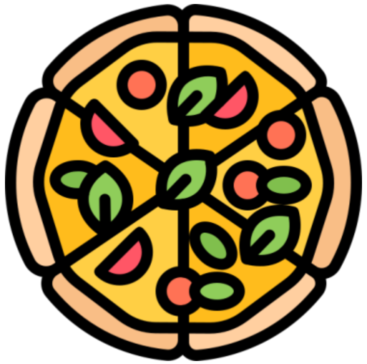
Subprblems → Big problem

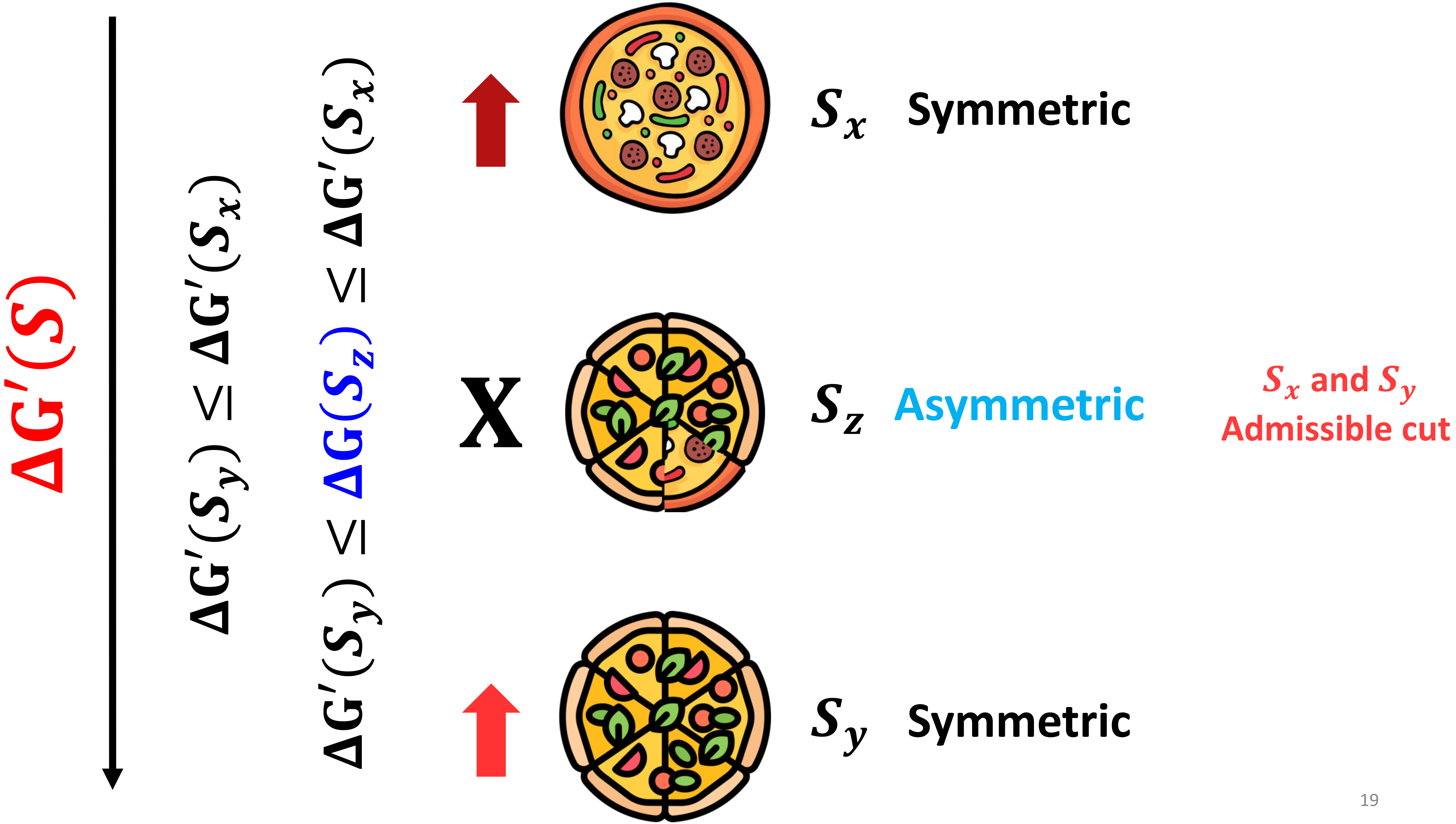
↓

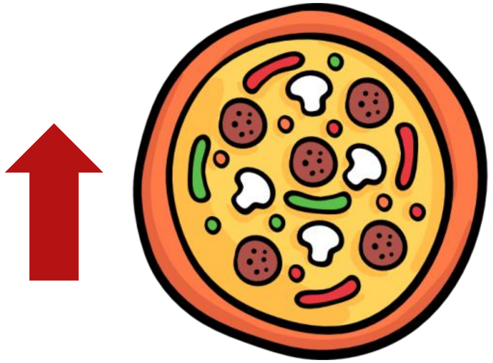
Local point of view → ? → Global property



Our solution




$\Delta G'(S)$
$\Delta G'(S_y) \leq \Delta G'(S_x)$
$\Delta G'(S_y) \leq \Delta G(S_z) \leq \Delta G'(S_x)$
X
S_x Symmetric
S_z Asymmetric
S_y Symmetric
S_x and S_y
Admissible cut



S_x

Symmetric

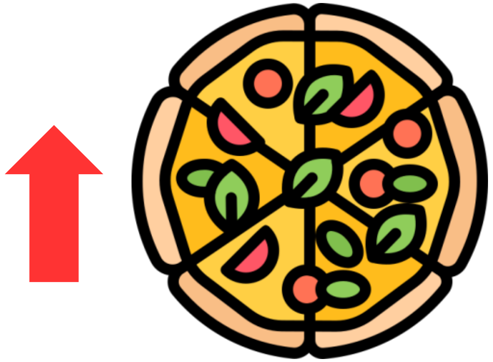
X 

S_z

Asymmetric

S_x and S_y

Admissible cut



S_y

Symmetric

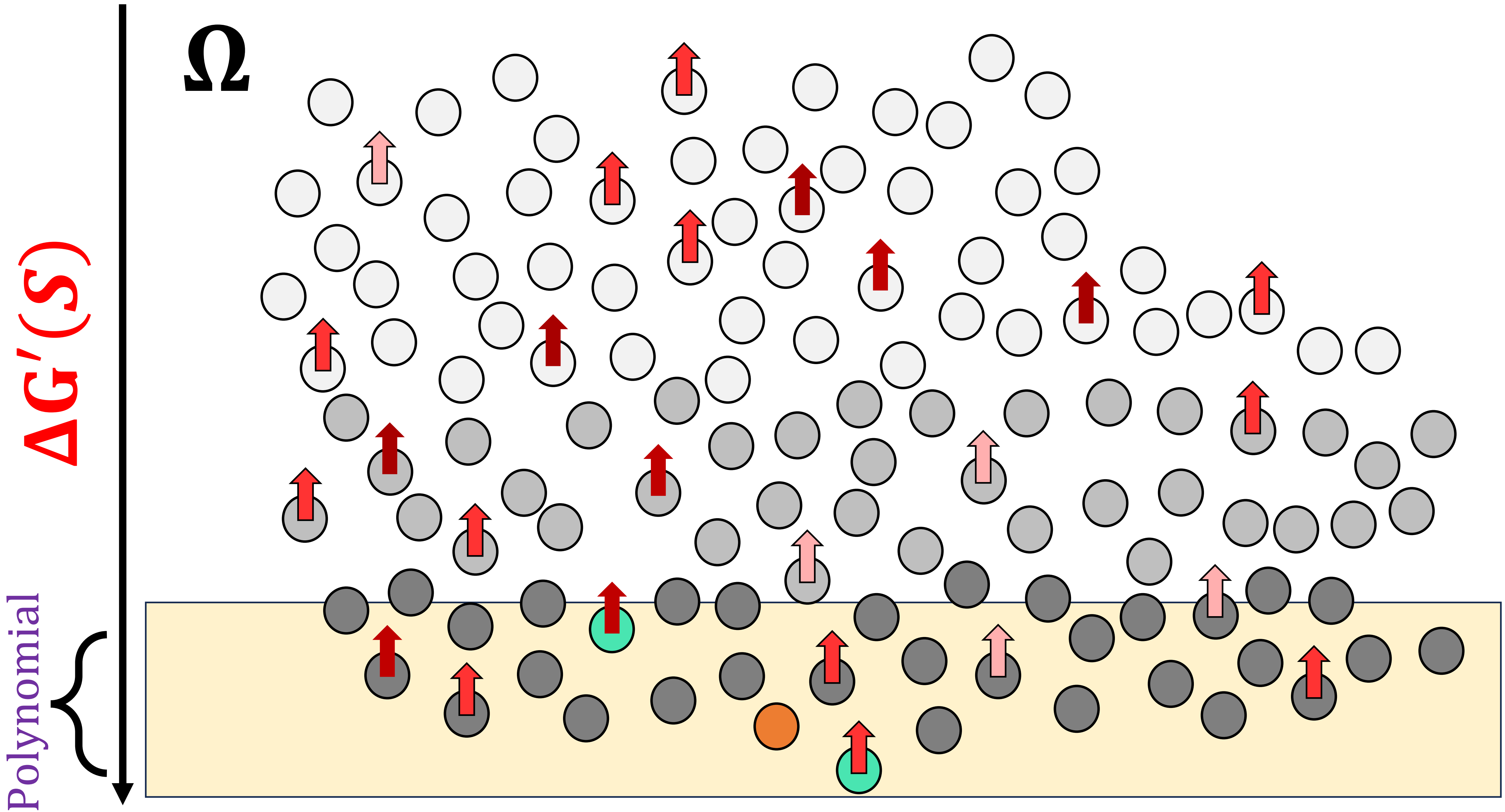
Upper bound

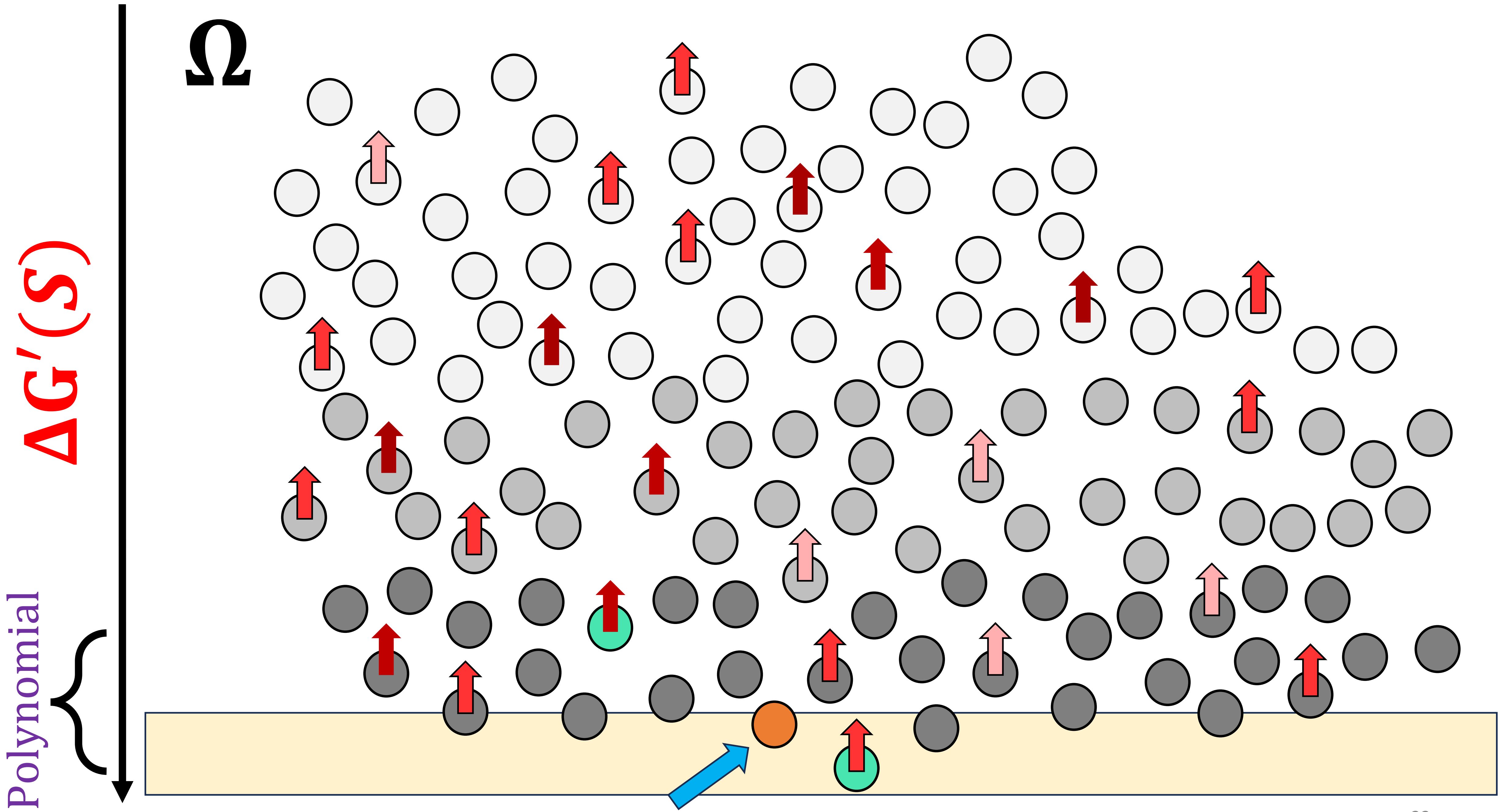
$$\boxed{\frac{N-c}{v(\pi)}\left(\sigma\left(v\left(\pi\right)\right)-v\left(\pi\right)\right)}$$

$$+$$

$$\boxed{N^2/16}$$

Adjusting the backtracking algorithm to go through energy levels sequentially starting from the MFE level.


Ω
ΔG′(S)
Polynomial



Ω
ΔG′(S)
Polynomial

Computational complexity of Minimum Free Energy algorithms

Input Type	MFE
Single Strand	$O(N^4)$
Multiple **unique** Strands, Bounded ($\leq$ c)	$O(N^4(c-1)!)$
Multiple Strands, Bounded ($\leq$ c)	$\mathbf{O(N^4(c-1)!)}$
Multiple Strands, Unbounded	NP − Complete

N bases, c strands

Thanks

Hamilton Institute

Maynooth University
National University of Ireland Maynooth

dna.hamilton.ie/shalaby

We're hiring!

sfi Science Foundation Ireland

erc

European Innovation Council

Funded by the European Union