Protein-Protein Docking
Basics and New Applications

Oliver Kohlbacher
Hans-Peter Lenhof

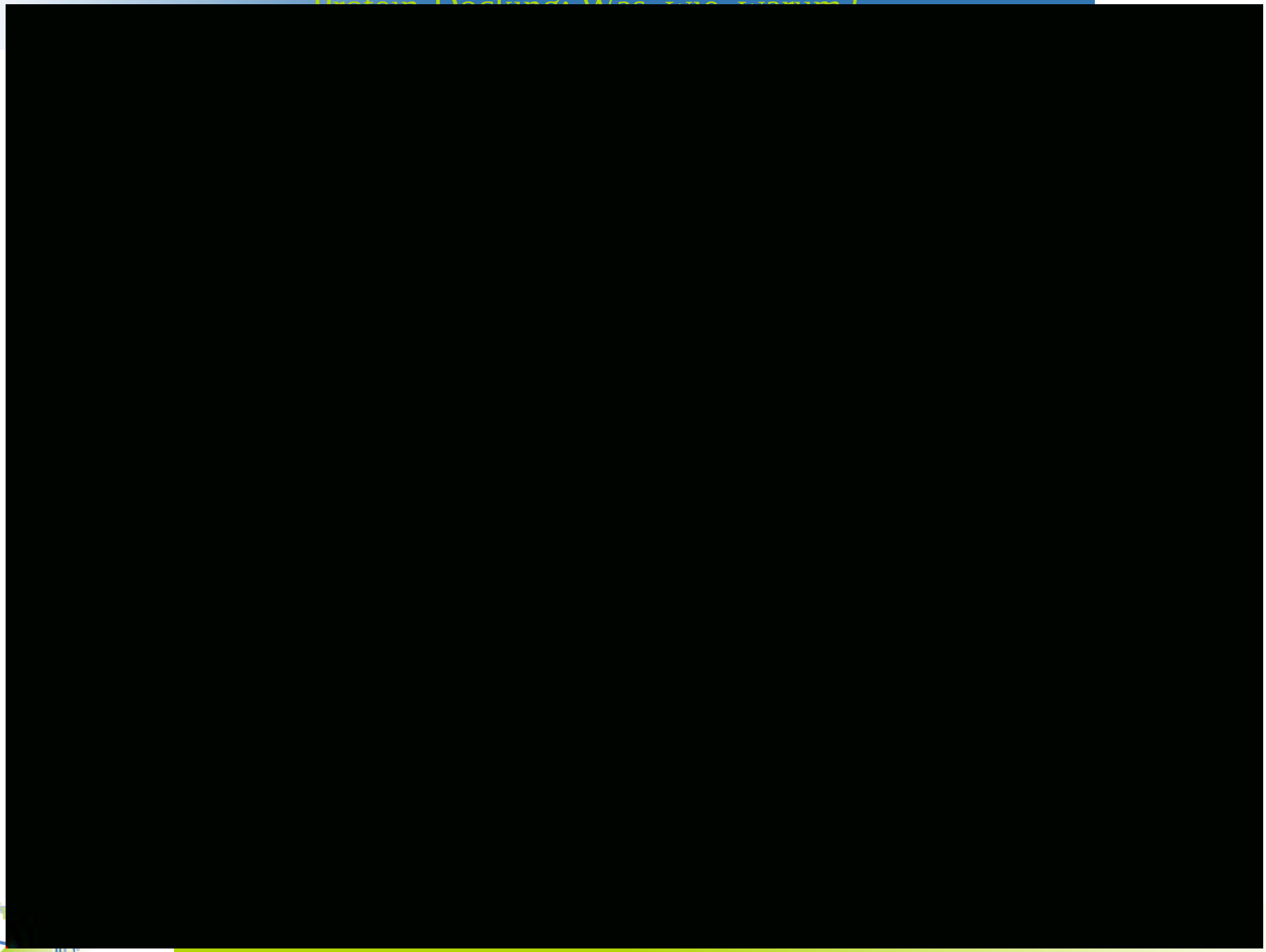
Center for Bioinformatics, Saarbrücken, Germany

Celera Genomics, Rockville, MD, USA

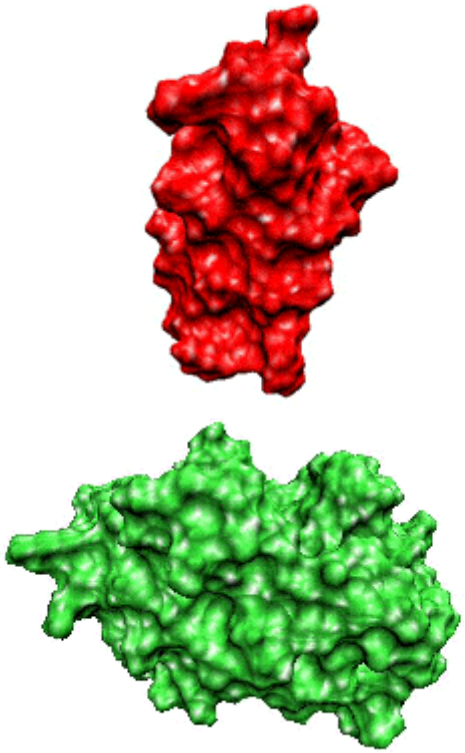
Overview

- Introduction
- Basic Algorithms + Scoring Functions
- Integration of Protein Flexibility
  - Flexibility in Proteins
  - Algorithms for Semi-Flexible Docking
- Integration of Experimental Data
  - NMR Spectroscopy
  - NMR-Based Protein Docking
- Outlook + Summary

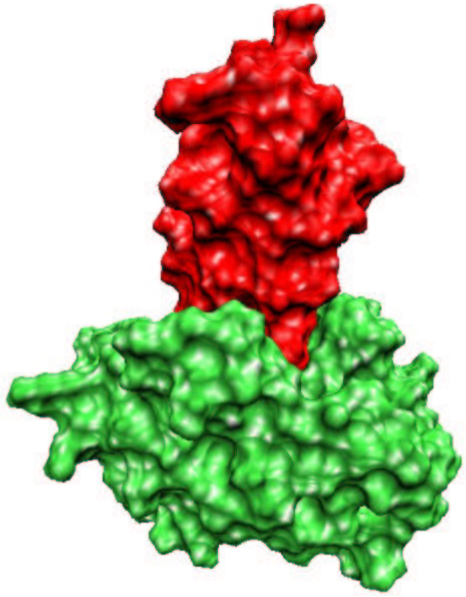
Protein Docking: Was, wie, warum?



Protein-Protein Docking

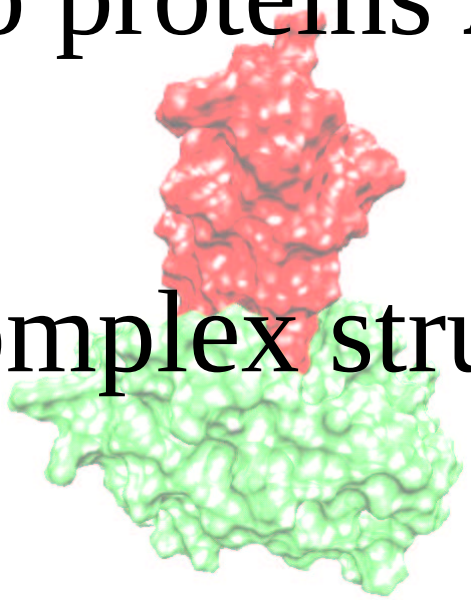


Protein-Protein Docking



Protein-Protein Docking

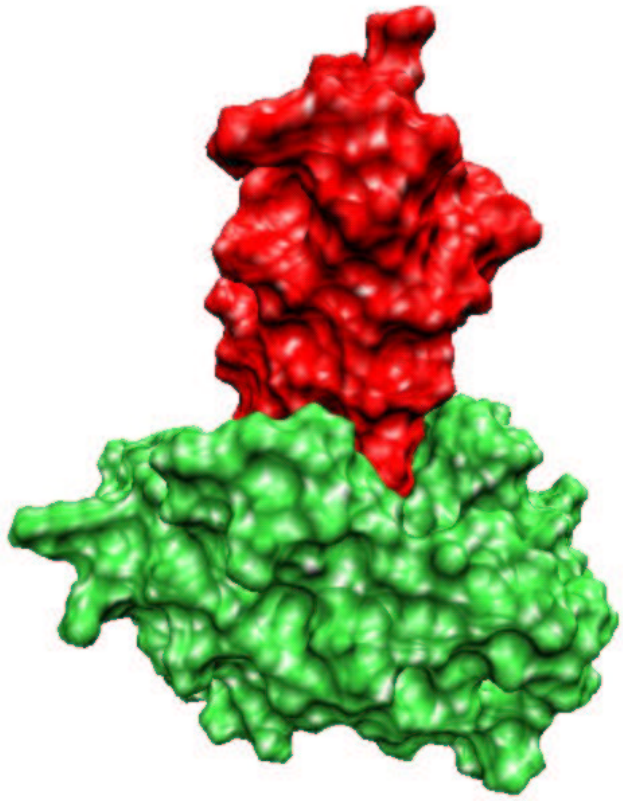
- Given two proteins **A** and **B**



- Predict complex structure **AB**

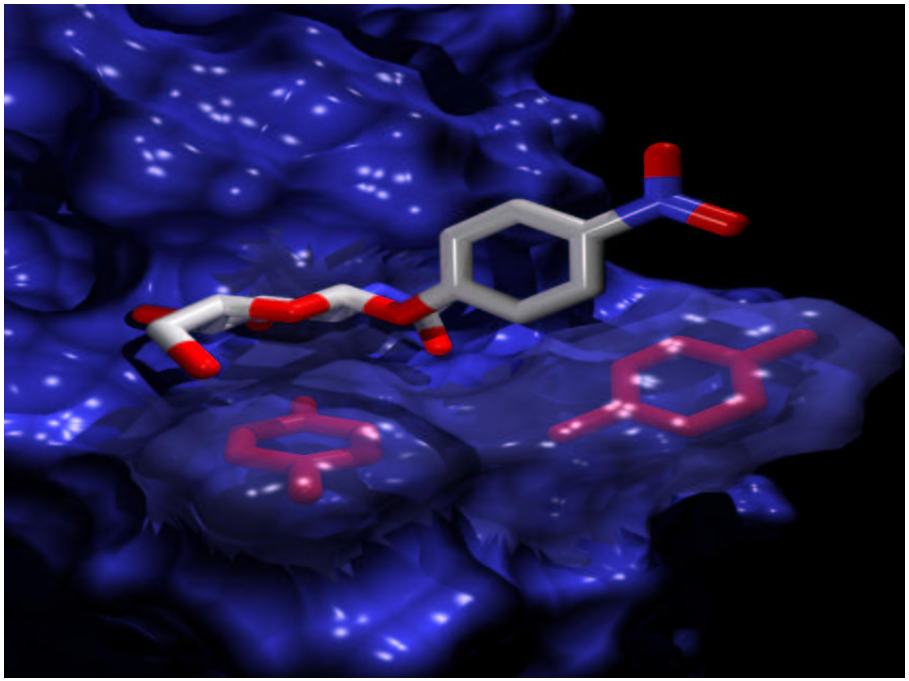
Protein-Protein Docking

- Uniform chemistry
- Understanding protein interactions
- Speed-up structure elucidation
- Predict protein interactions



Protein-Ligand Docking

- Small, flexible Ligand
- Ligands are chemically diverse
- Crucial for Drug Design
- Virtual screening

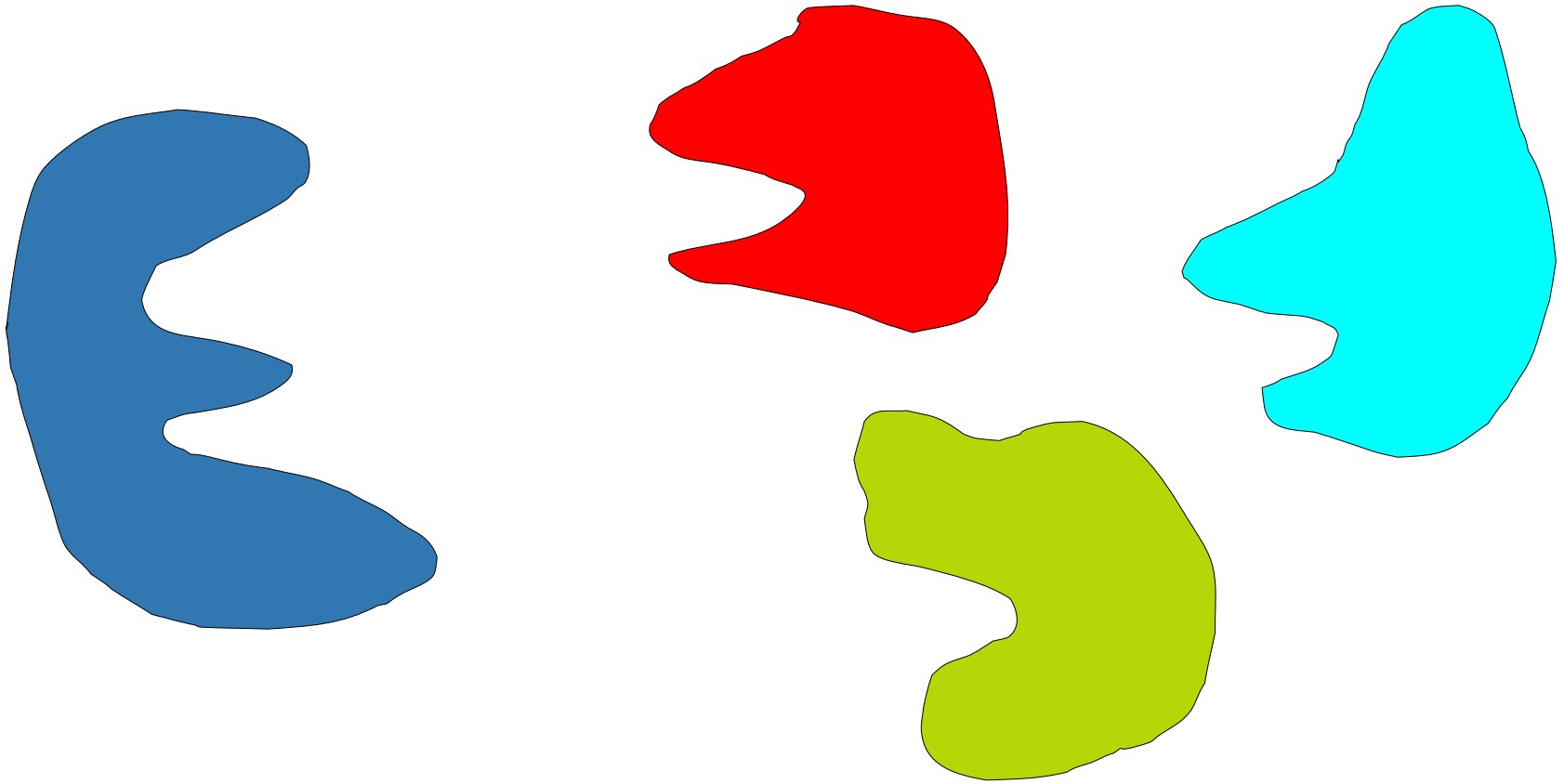


Lock-and-Key Principle

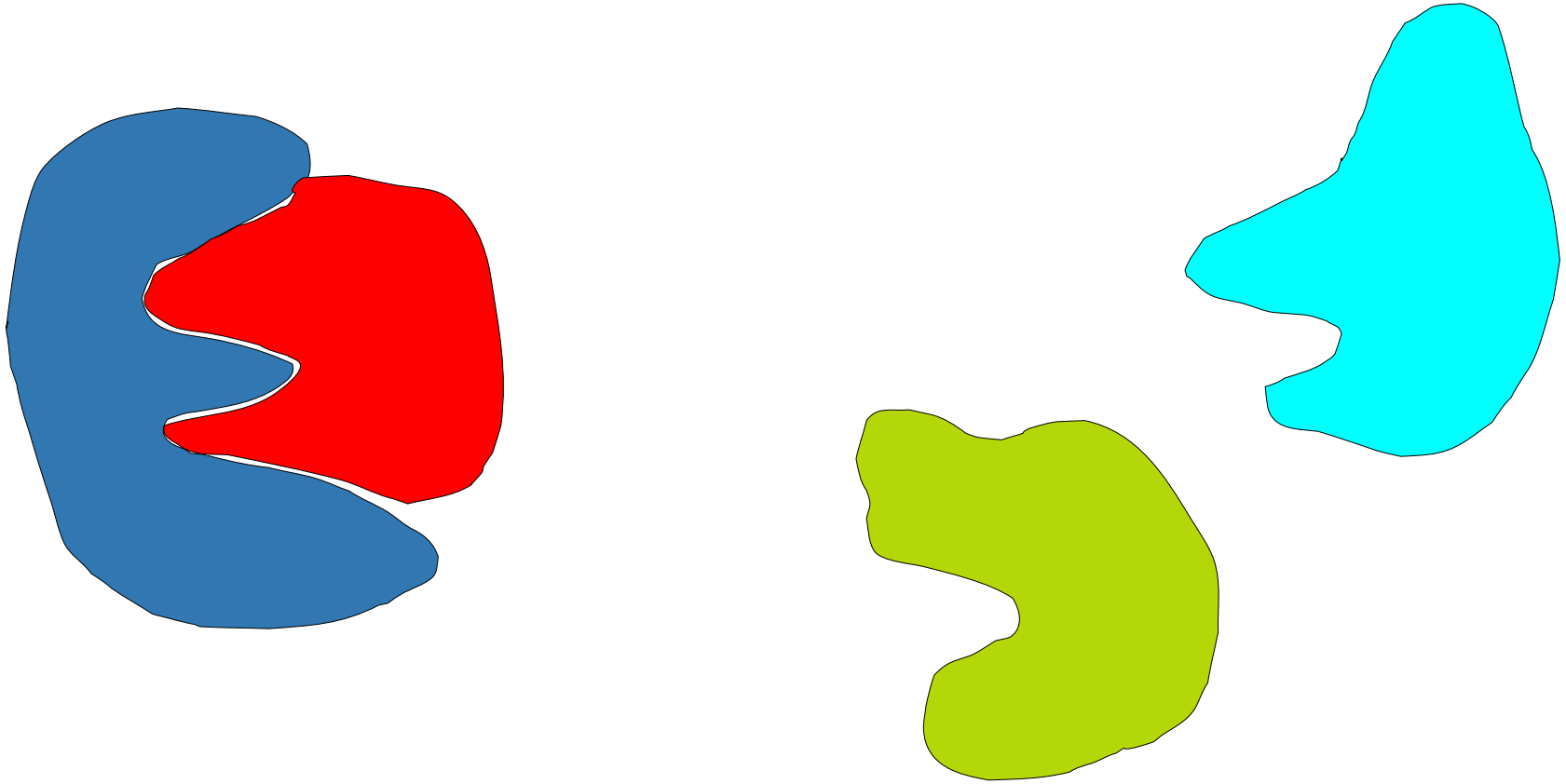
Emil Fischer 1894

" To use an image, I would say that enzyme and glycoside have to fit into each other like a lock and a key, in order to exert a chemical effect on each other."

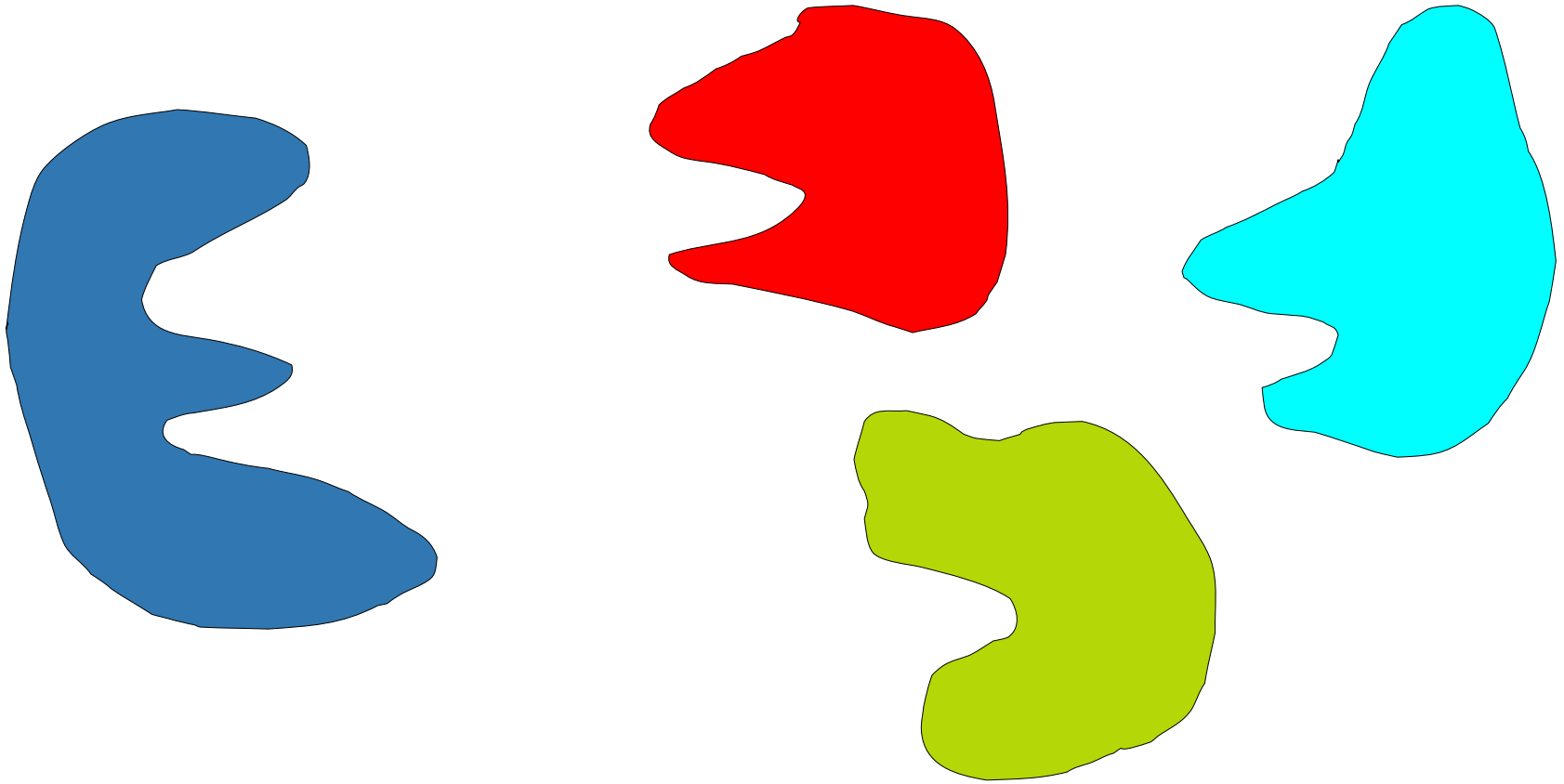
Lock-and-Key Principle



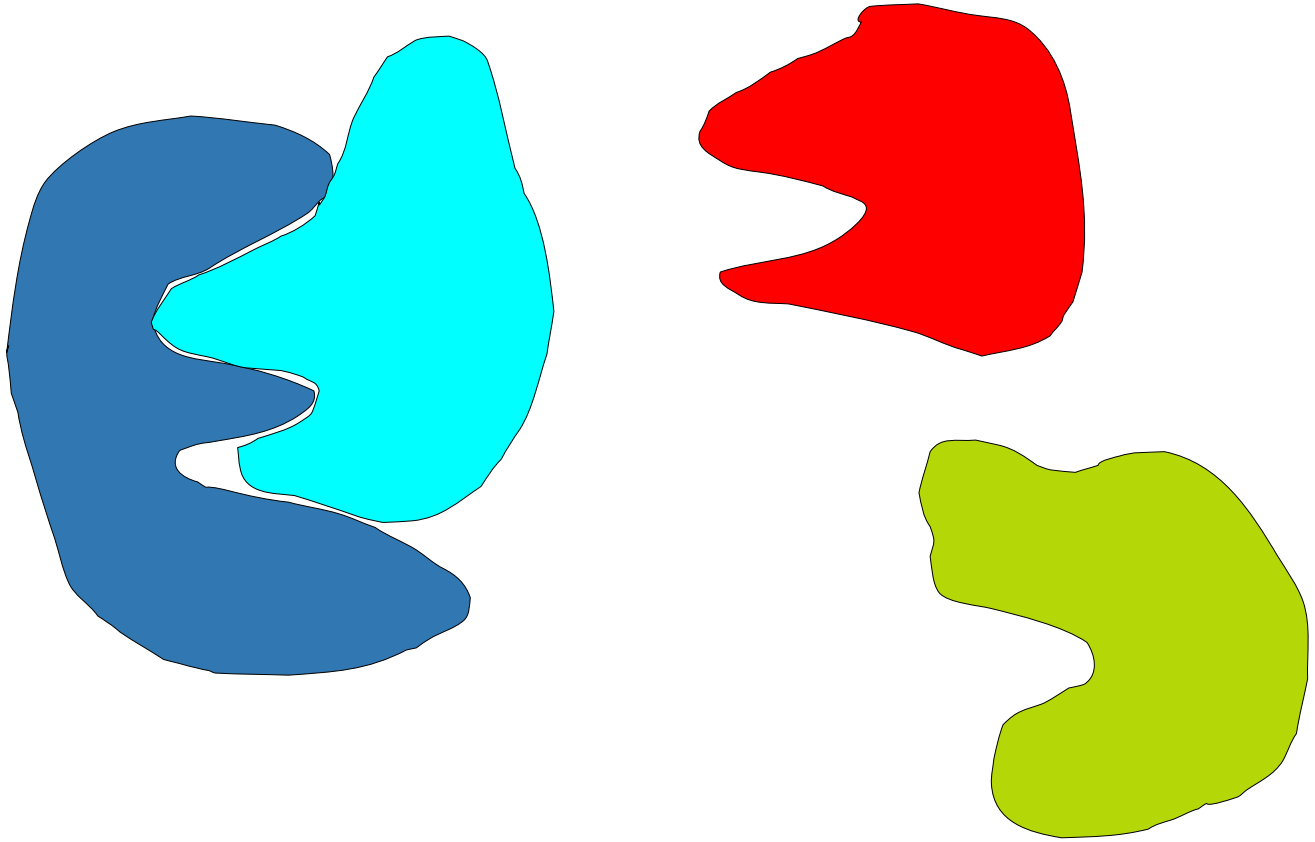
Lock-and-Key Principle



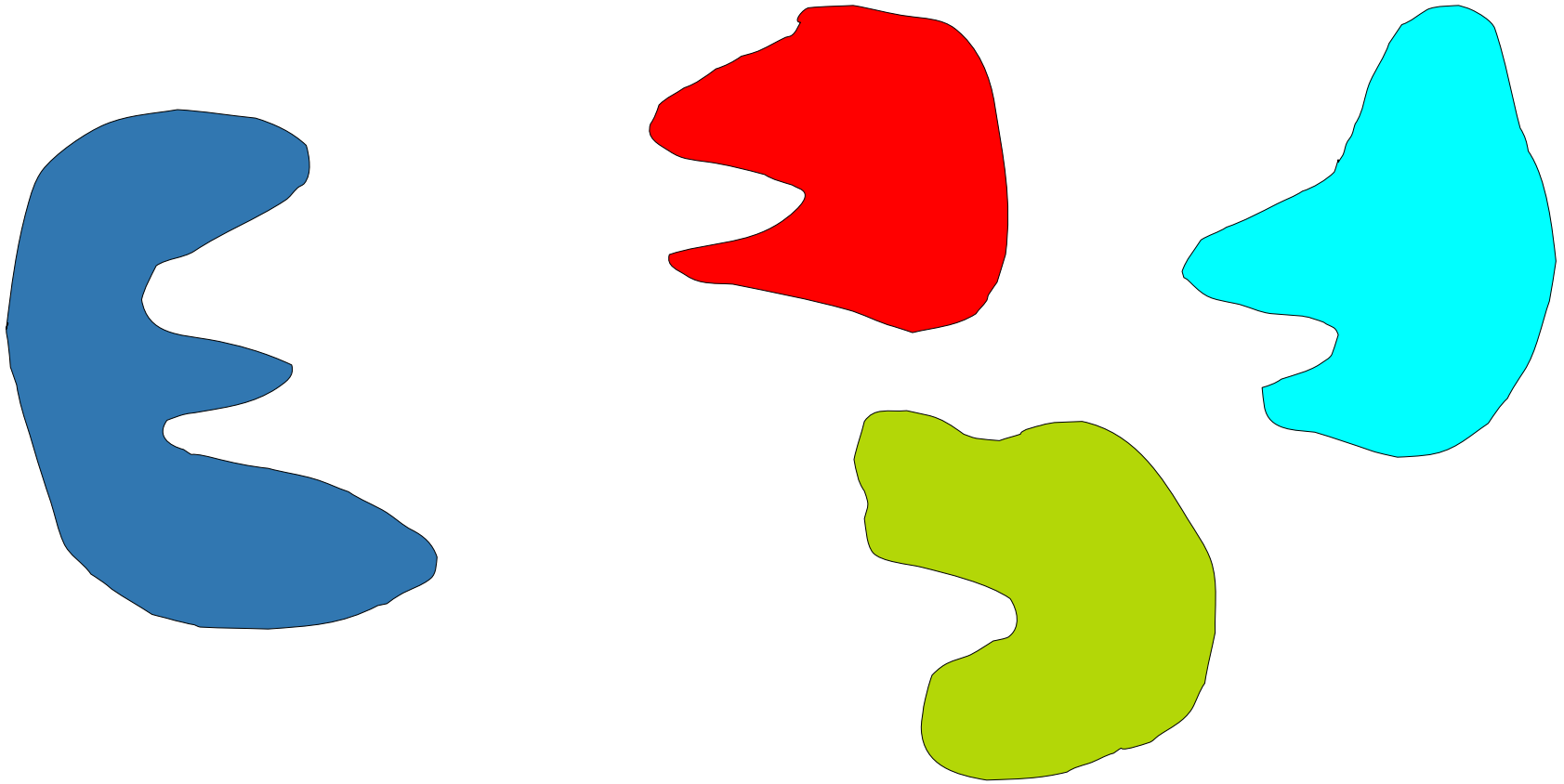
Lock-and-Key Principle



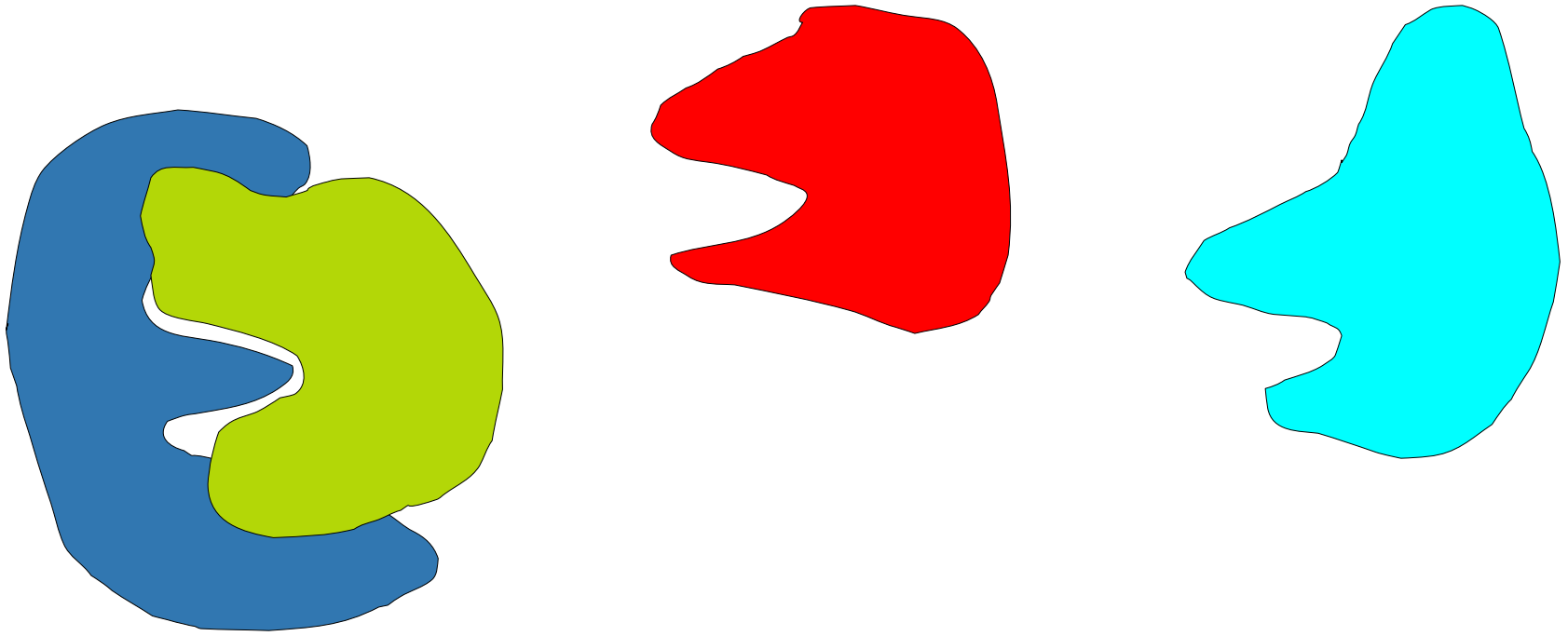
Lock-and-Key Principle



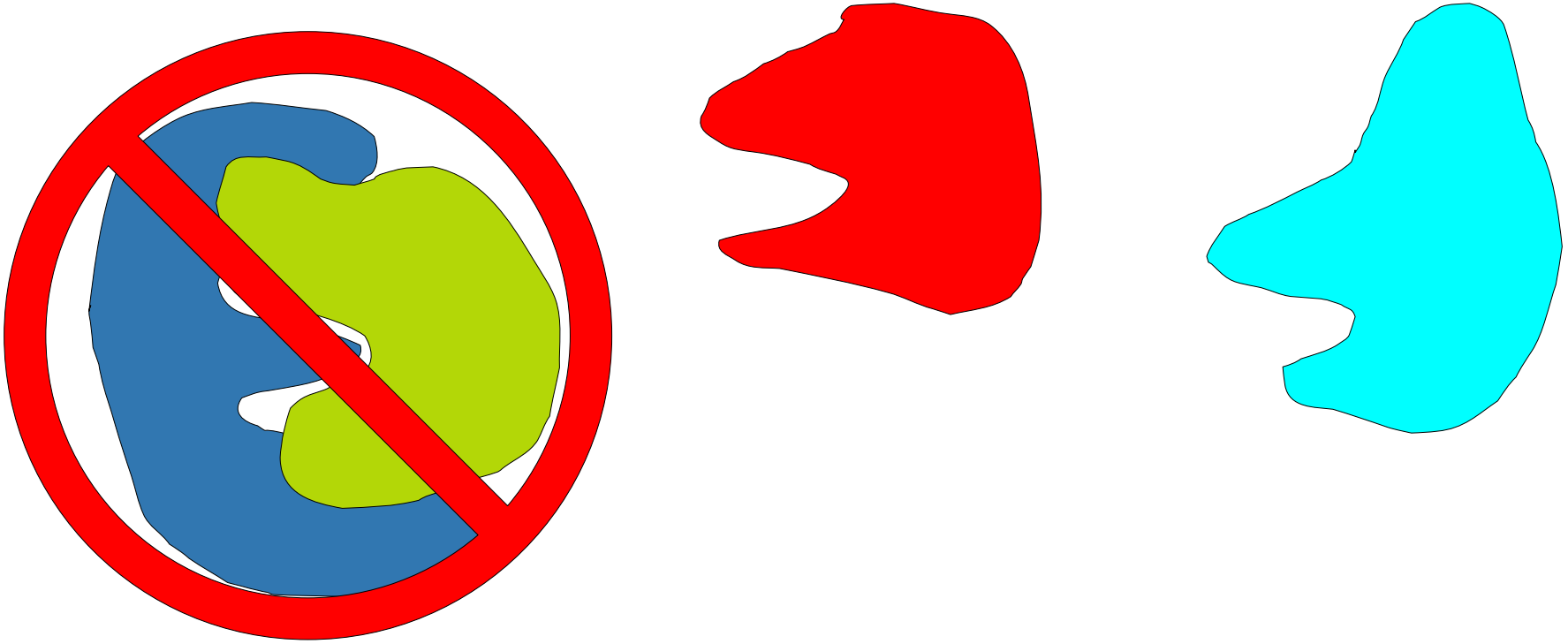
Lock-and-Key Principle



Lock-and-Key Principle

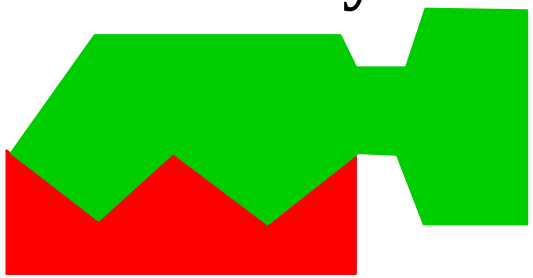


Lock-and-Key Principle

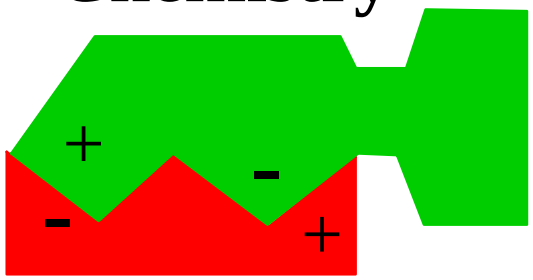


Lock-and-Key Principle

Geometry

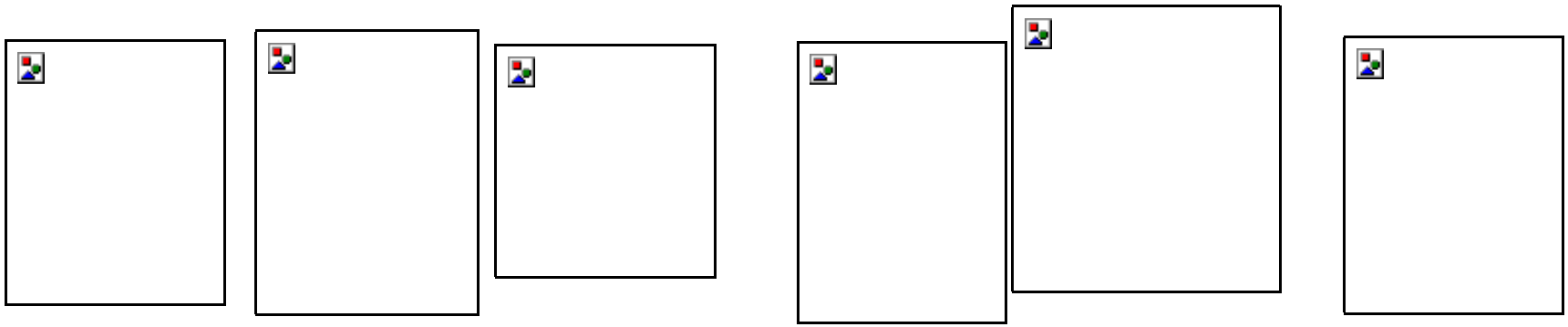


Chemistry



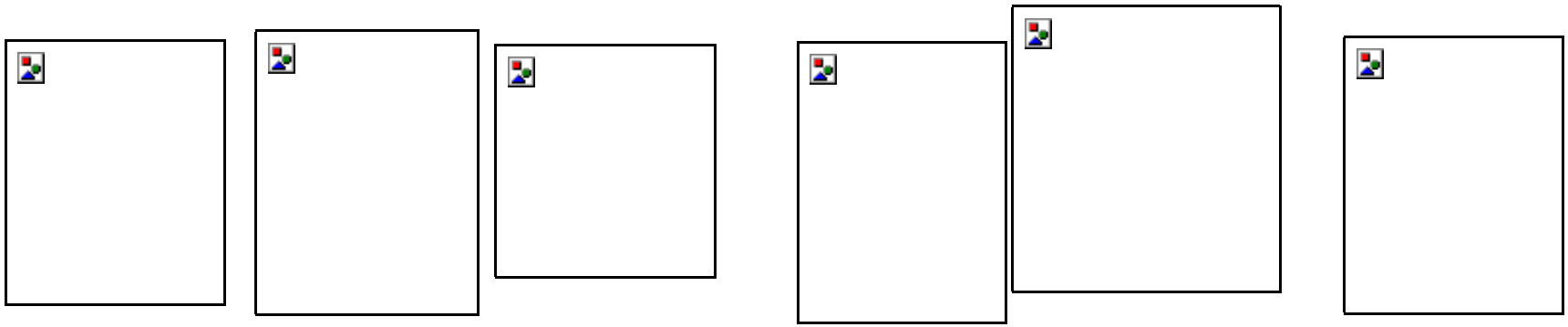
Protein Docking: How?

- Structure generation



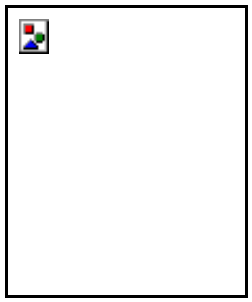
Protein Docking: How?

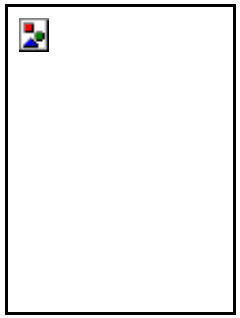
- Structure generation
- Filtering

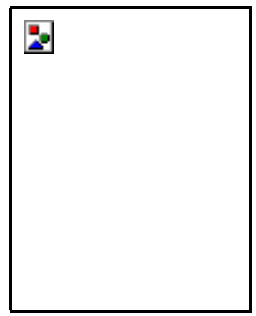


Protein Docking: How?

- Structure generation
- Filtering

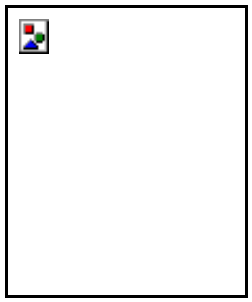


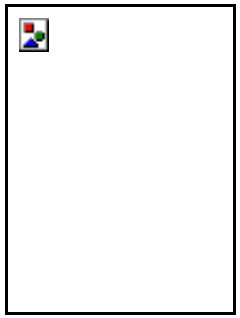


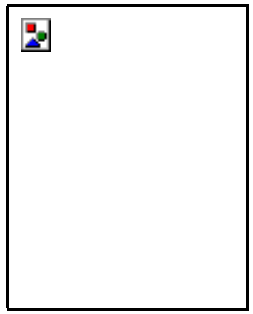


Protein Docking: How?

- Structure generation
- Filtering
- Evaluation

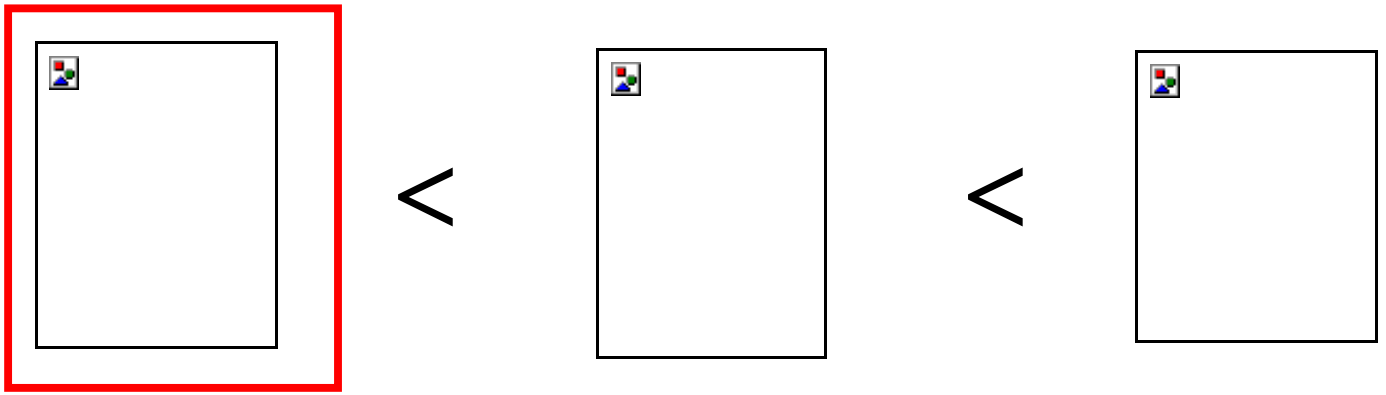






Protein Docking: How?

- Structure generation
- Filtering
- Evaluation



Overview of Docking Techniques

Connolly Surface	
Connolly	1986
Bacon, Moult	1992
Fischer et al.	1995
Lin et al.	1994
Norel et al.	1995
Sandak et al.	1995
Campbell et al.	1996
Ackermann et al.	1995

Fuzzy Logic	
Extner, Brickmann	1997

Cube Representation	
Jiang, Kim	1991

Graph Representation	
Shoichet et al.	1992
Shoichet, Kuntz	1996
Kasinos et al.	1992

Correlation	
Katchalski-Katzir et al.	1992
Vakser, Aflalo	1994
Vakser	1996
Gabb et al.	1997
Meyer et al.	1996

Monte Carlo Approach	
Cherfils et al.	1991
Totrov, Abagyan	1994

Slices Representation	
Walls, Sternberg	1992
Helmer-Citterich et al.	1994
Ausiello et al.	1997

Genetic Algorithms	
Levine et al.	1997

Database	
Ester et al.	1995

Overview of Docking Techniques

Connolly Surface	
Connolly	1986
Bacon, Moult	1992
Fischer et al.	1995
Lin et al.	1994
Norel et al.	1995
Sandak et al.	1995
Campbell et al.	1996
Ackermann et al.	1995

Fuzzy Logic	
Extner, Brickmann	1997

Cube Representation	
Jiang, Kim	1991

Graph Representation	
Shoichet et al.	1992
Shoichet, Kuntz	1996
Kasinos et al.	1992

Correlation	
Katchalski-Katzir et al.	1992
Vakser, Aflalo	1994
Vakser	1996
Gabb et al.	1997
Meyer et al.	1996

Monte Carlo Approach	
Cherfils et al.	1991
Totrov, Abagyan	1994

Slices Representation	
Walls, Sternberg	1992
Helmer-Citterich et al.	1994
Ausiello et al.	1997

Genetic Algorithms	
Levine et al.	1997

Database	
Ester et al.	1995

What are the differences?

- Method for determining tentative structures
  - Based on surface complementarity
  - Based on triangles
  - ….
- Scoring functions employed
  - Contact surface area
  - Electrostatics
  - Hydrogen bonds
  - ….

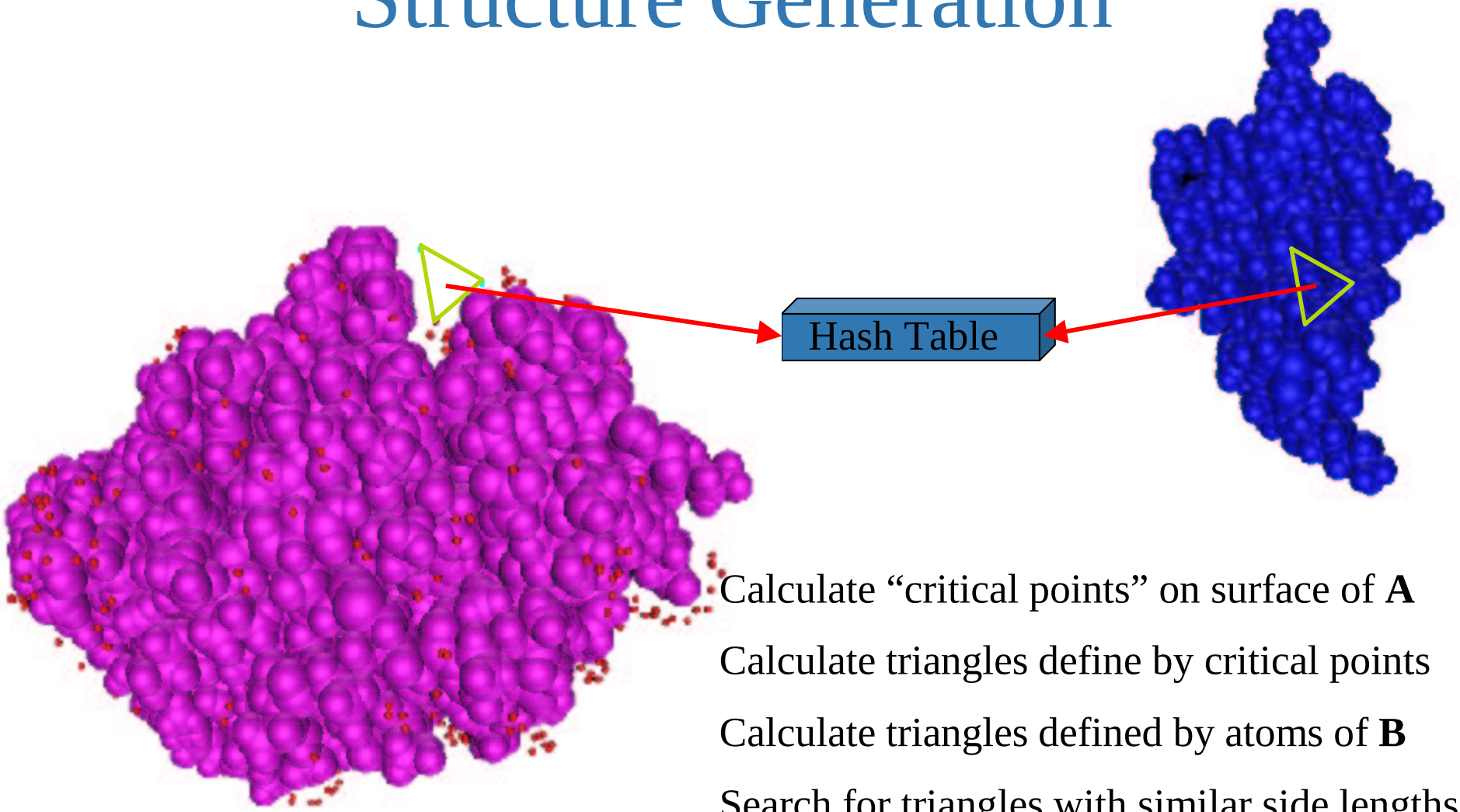
Structure Generation

- Assumption: Proteins are rigid bodies!
- Three-dimensional “puzzle”
- Six degrees of freedom
  - Rotation
  - Translation
- Identify rigid transformations bringing **B** in contact with **A**
- Discretization
  - Of proteins, protein surface
  - Rotational/translational space

Structure Generation

- Katchalski-Katzir et al., Proc. Natl. Acad. Sci. USA, 1992
  - Grid-based correlation of **A** and **B**
- Lenhof, RECOMB 97
  - Mapping of triangles

Structure Generation



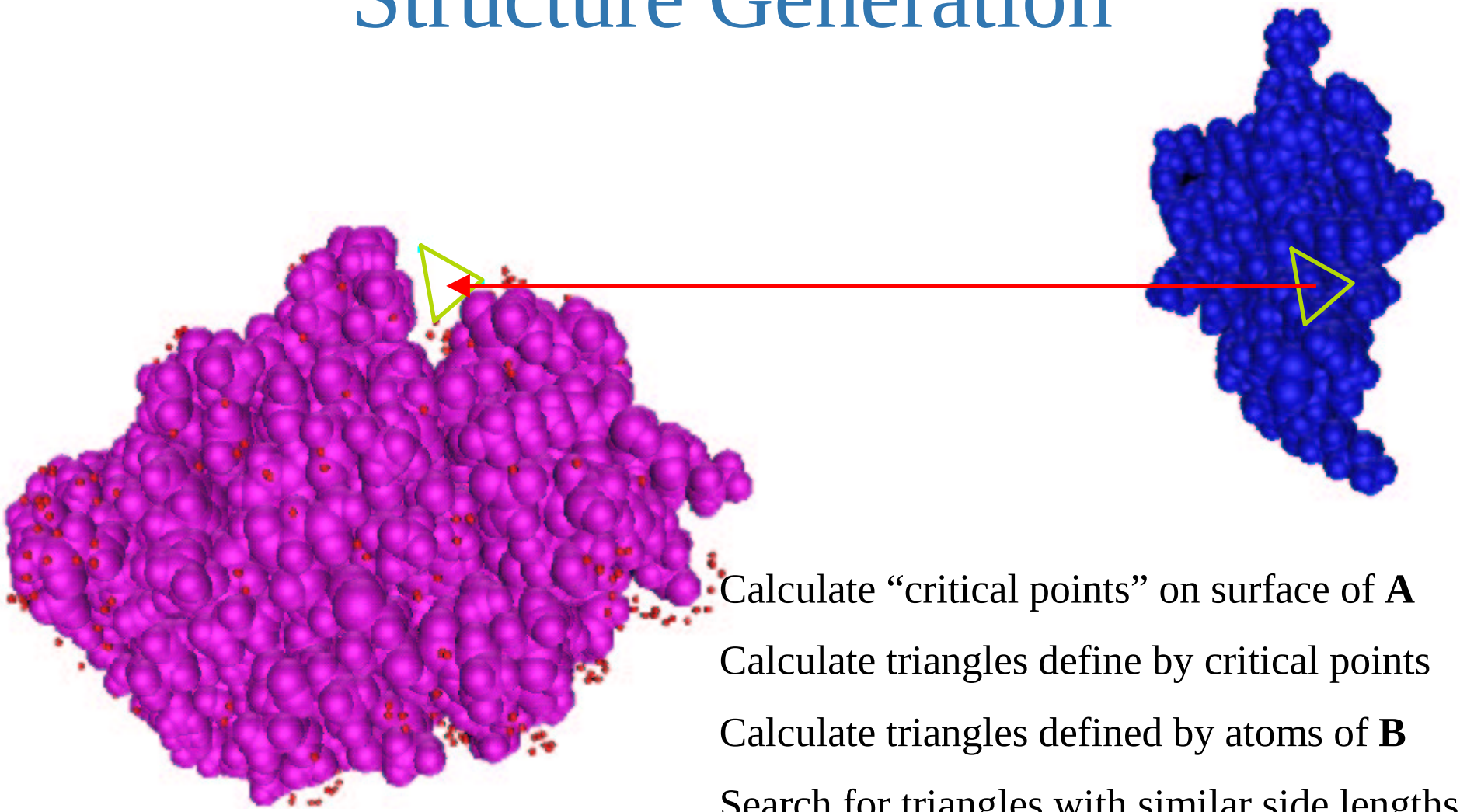
Calculate “critical points” on surface of **A**

Calculate triangles define by critical points

Calculate triangles defined by atoms of **B**

Search for triangles with similar side lengths

Structure Generation



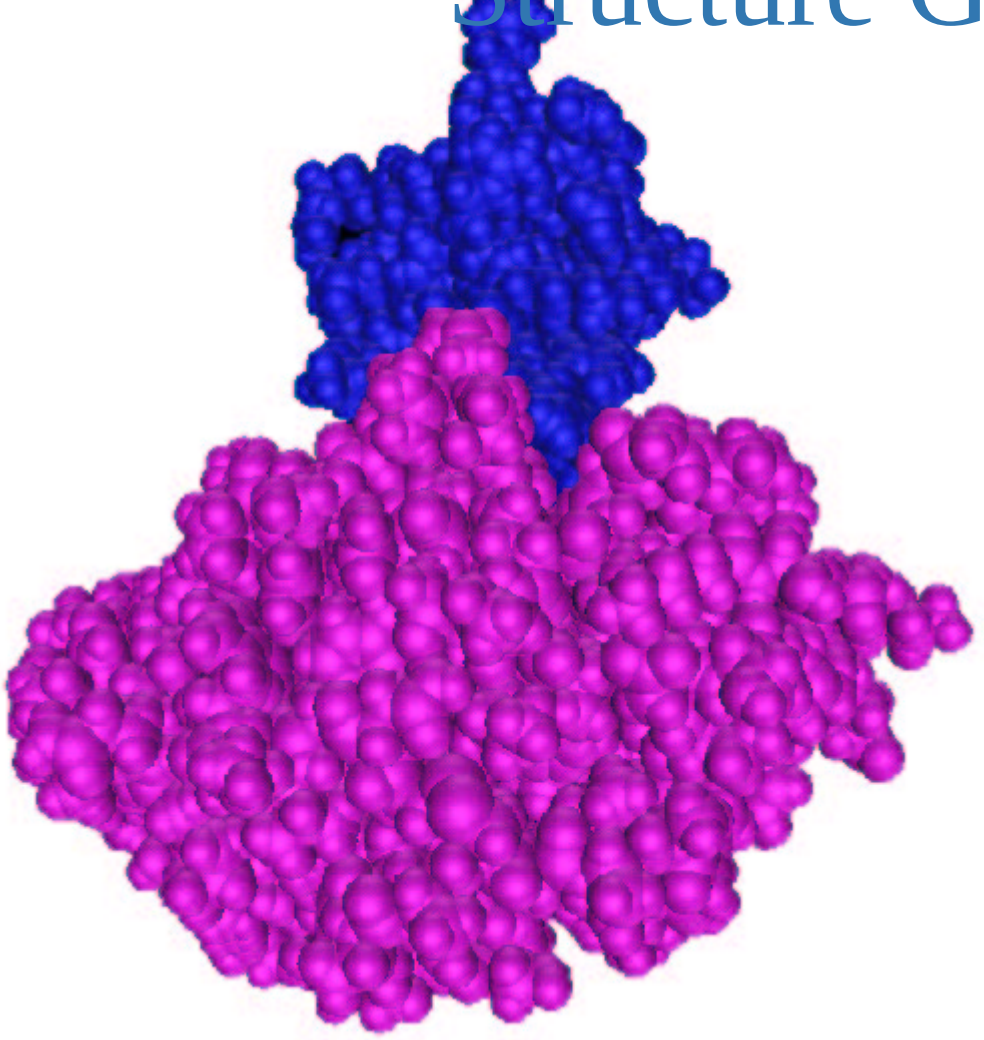
Calculate “critical points” on surface of **A**

Calculate triangles define by critical points

Calculate triangles defined by atoms of **B**

Search for triangles with similar side lengths

Structure Generation



Scoring Functions

- What distinguishes the true complex structure from “false positives”?
- Physical chemistry:

  Complex structure with the lowest binding free energy is the one observed in nature.
- *Caveat:* relies on sufficiently complete sampling of conformation space

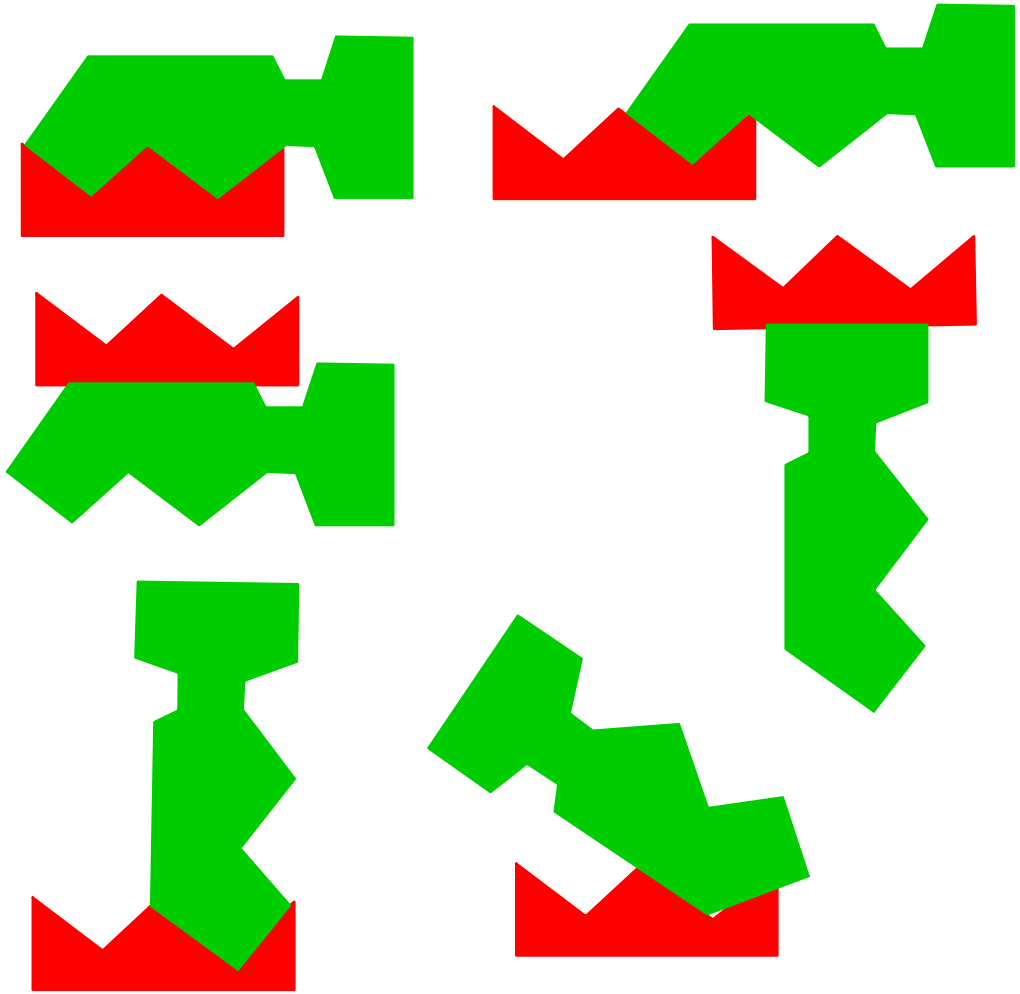
Prediction of Free Energy

- Also hardest part in related problems:
  - Ligand docking
  - Protein structure prediction
- Many methods neglect
  - entropic contributions
  - solvent effects

Scoring Functions

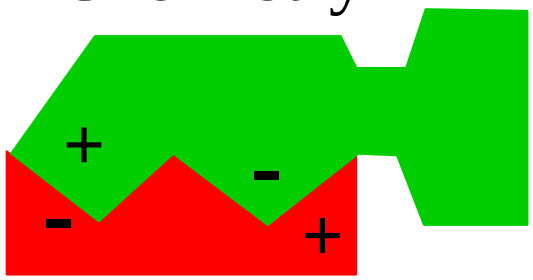
- Simplest functions: Geometry
  - Key-and-lock principle
  - Large contact areas are favorable
  - Overlaps between **A** and **B** are unfavorable
- More sophisticated: “Chemistry”
  - Models based on physicochemistry
  - Compromise between complexity and accuracy

Scoring Function

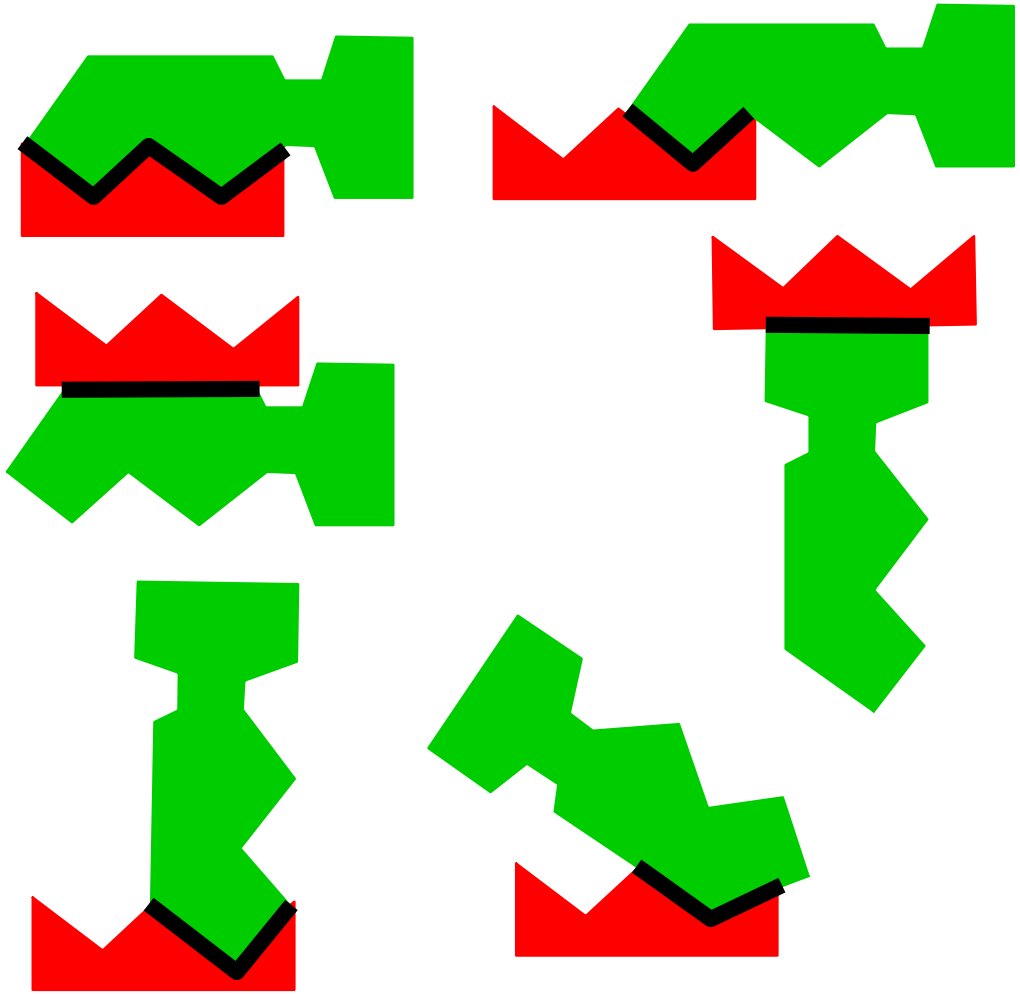


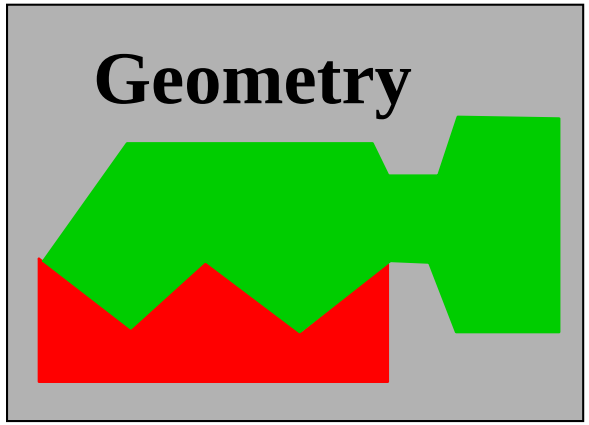
Geometry

Chemistry

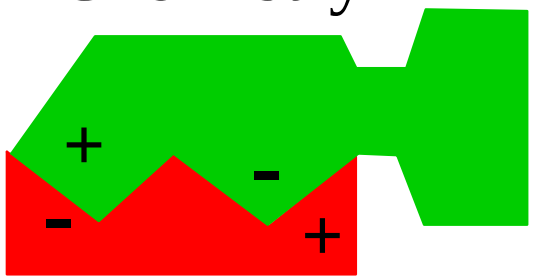


Scoring Function

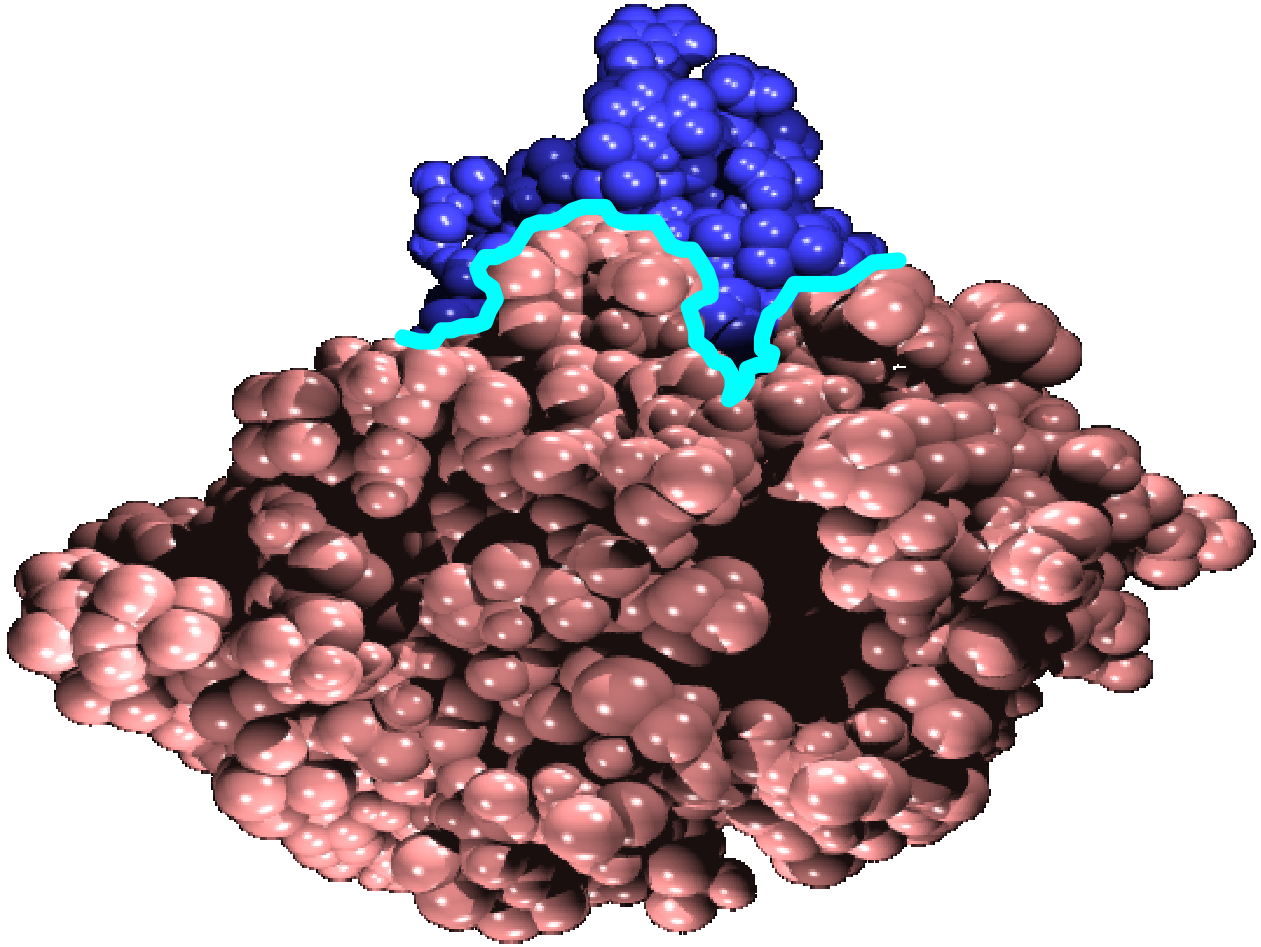




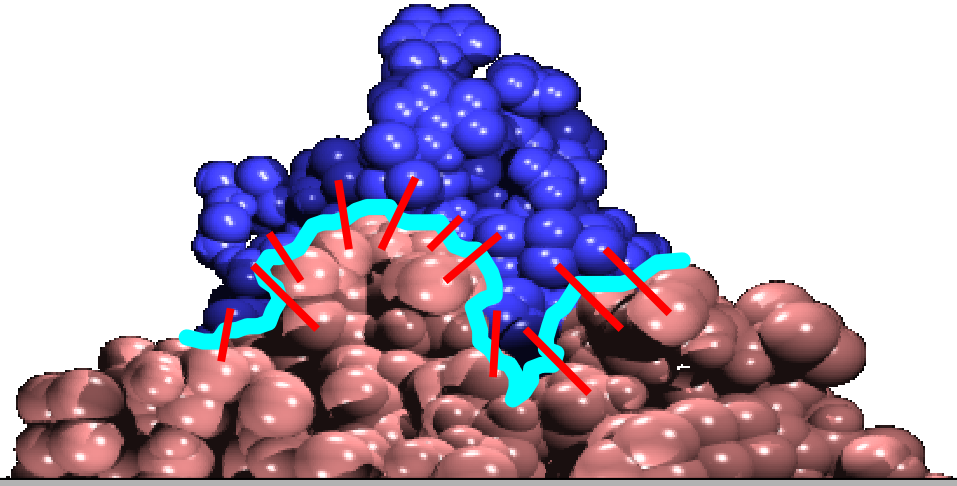
Chemistry



Geometric Scoring Function



Geometric Scoring Function



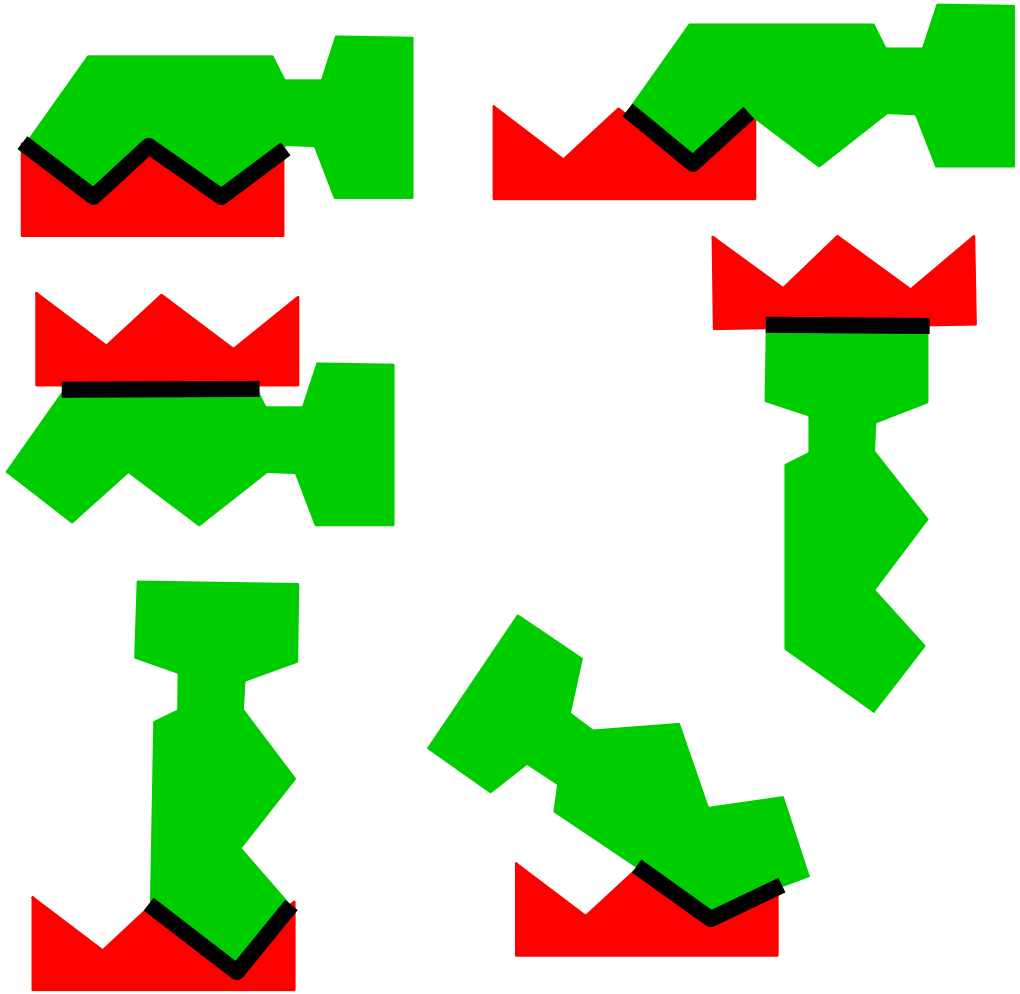
Determine distance between (a, b) ($a \in \mathbf{A}$, $b \in \mathbf{B}$)

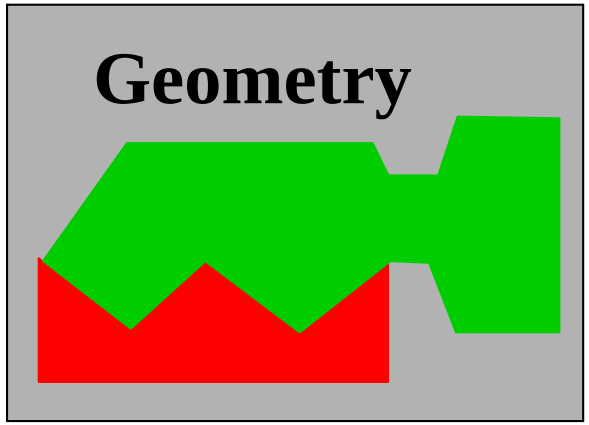
$$contact(\mathbf{AB}) = |\{(a,b) \mid 2.75 < dist(a,b) < 4.0\}|$$

$$overlap(\mathbf{AB}) = |\{(a,b) \mid dist(a,b) < 2.75\}|$$

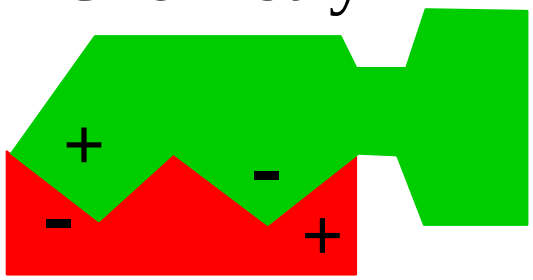
$$geom\_score(\mathbf{AB}) = k_1 \cdot contact(\mathbf{AB}) - k_2 \cdot overlap(\mathbf{AB})$$

Scoring Function

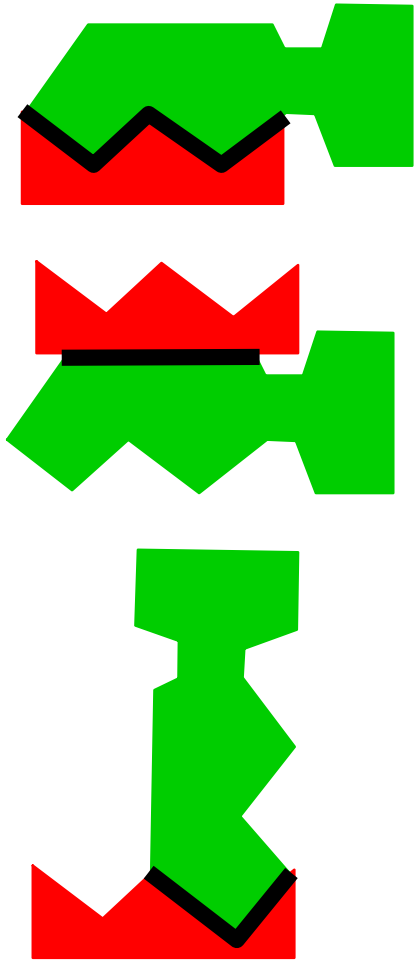


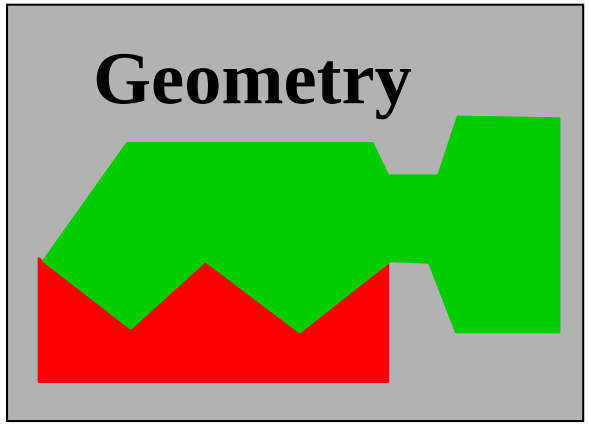


Chemistry

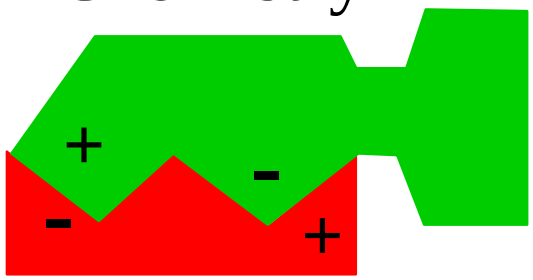


Scoring Function

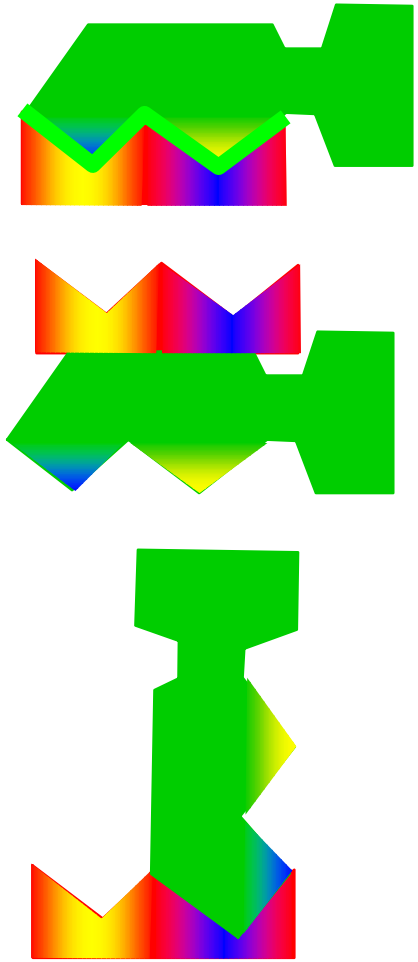




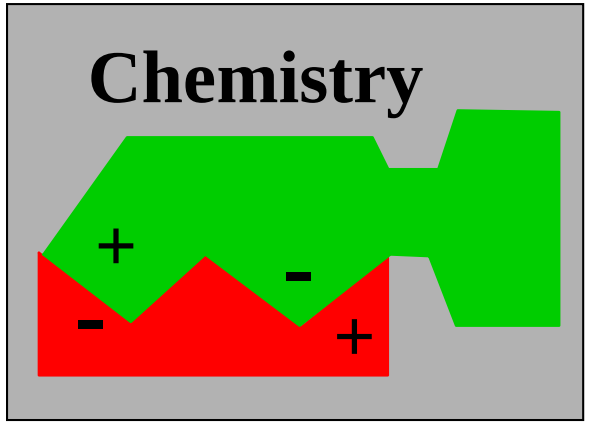
Chemistry



Scoring Function

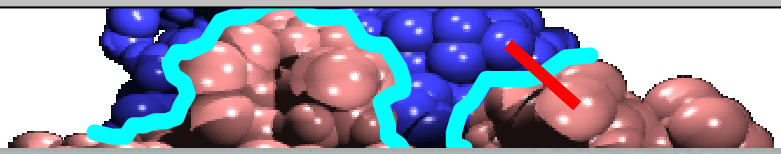


Geometry



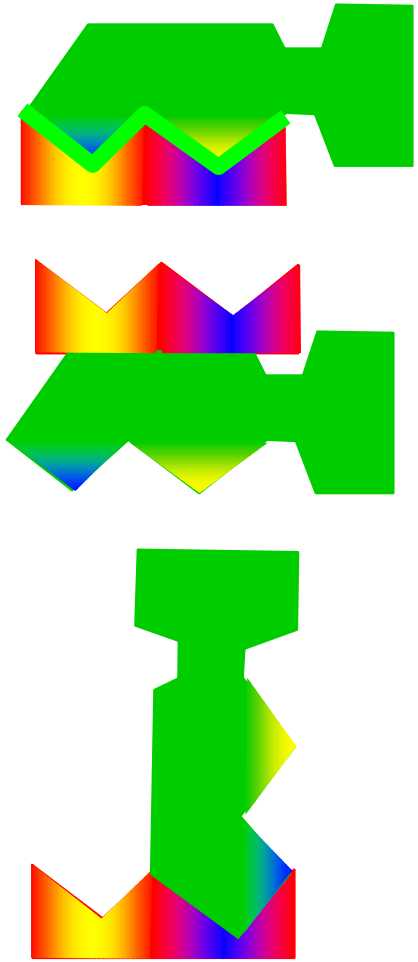
Chemical Scoring

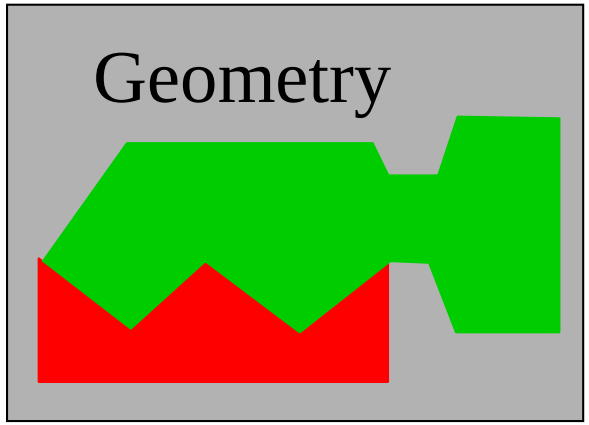
$$chem\_score(\boldsymbol{AB}) = \sum_{(a,b)\in\, contact} M_{type(a),type(b)}$$

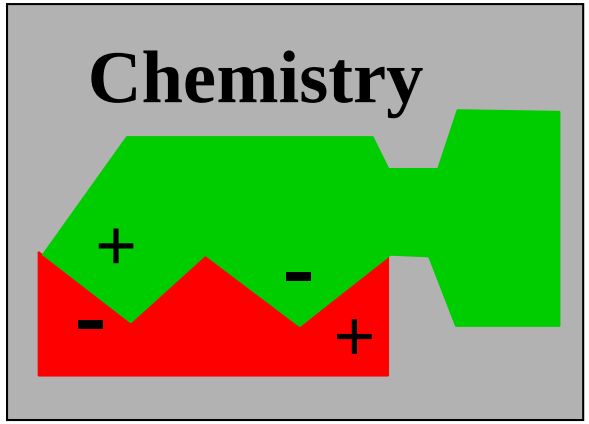


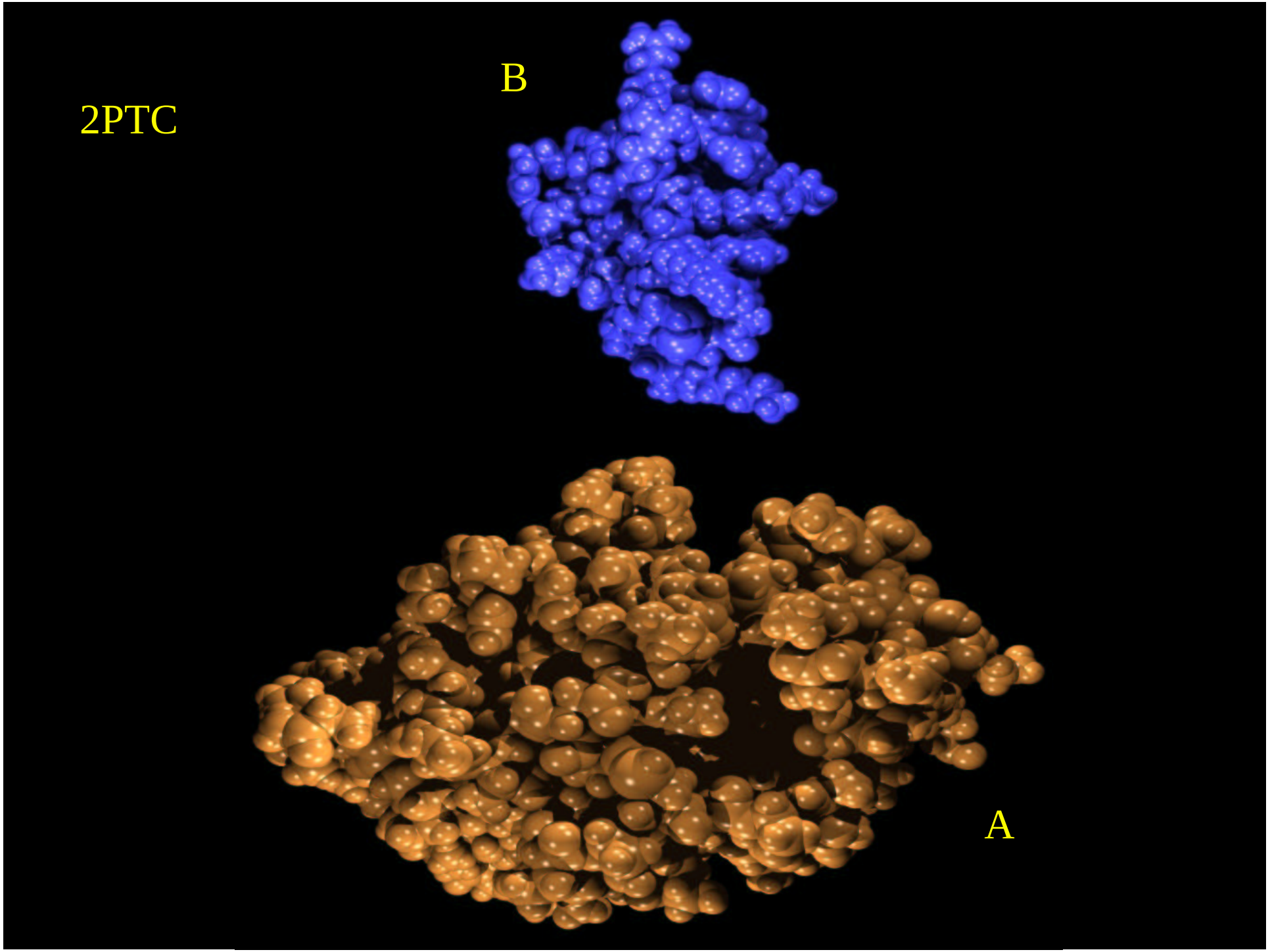
247 contacts		ARG									
		N	CA	C	O	CB	CG	CD	NE	CZ	NH1	NH2
ASP	N	0	0	0	1	0	0	0	0	0	2	1
	CA	0	0	0	1	0	0	0	0	0	0	4
	C	0	0	0	1	0	0	0	1	0	1	4
	O	0	0	0	0	0	0	2	5	6	2	7
	CB	0	0	0	0	0	1	1	1	1	4	6
	CG	1	0	0	0	0	3	6	10	7	9	16
	OD1	0	1	0	2	5	10	9	10	18	8	16
	OD2	2	1	0	0	3	2	7	9	13	12	15

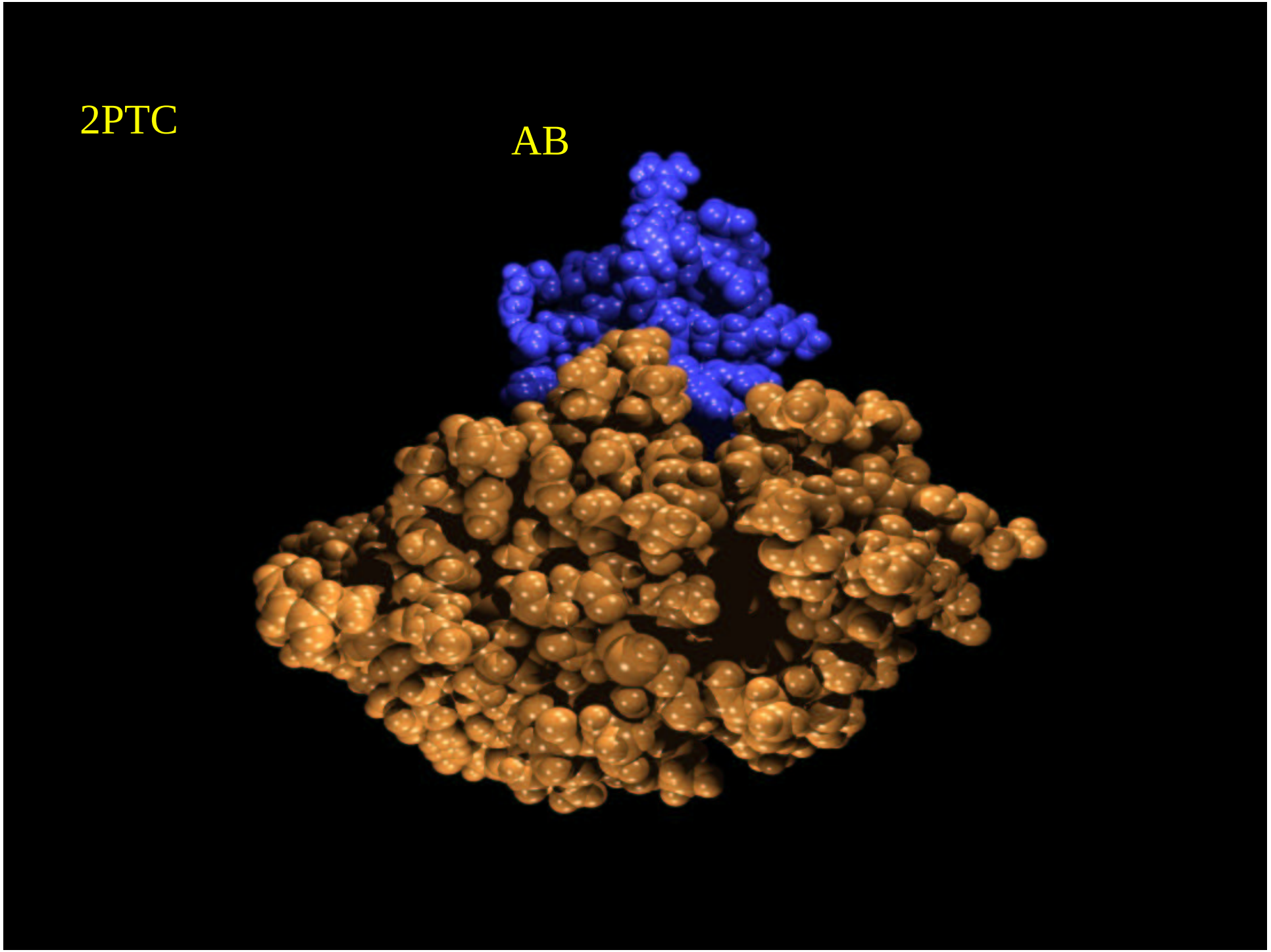
Scoring Function

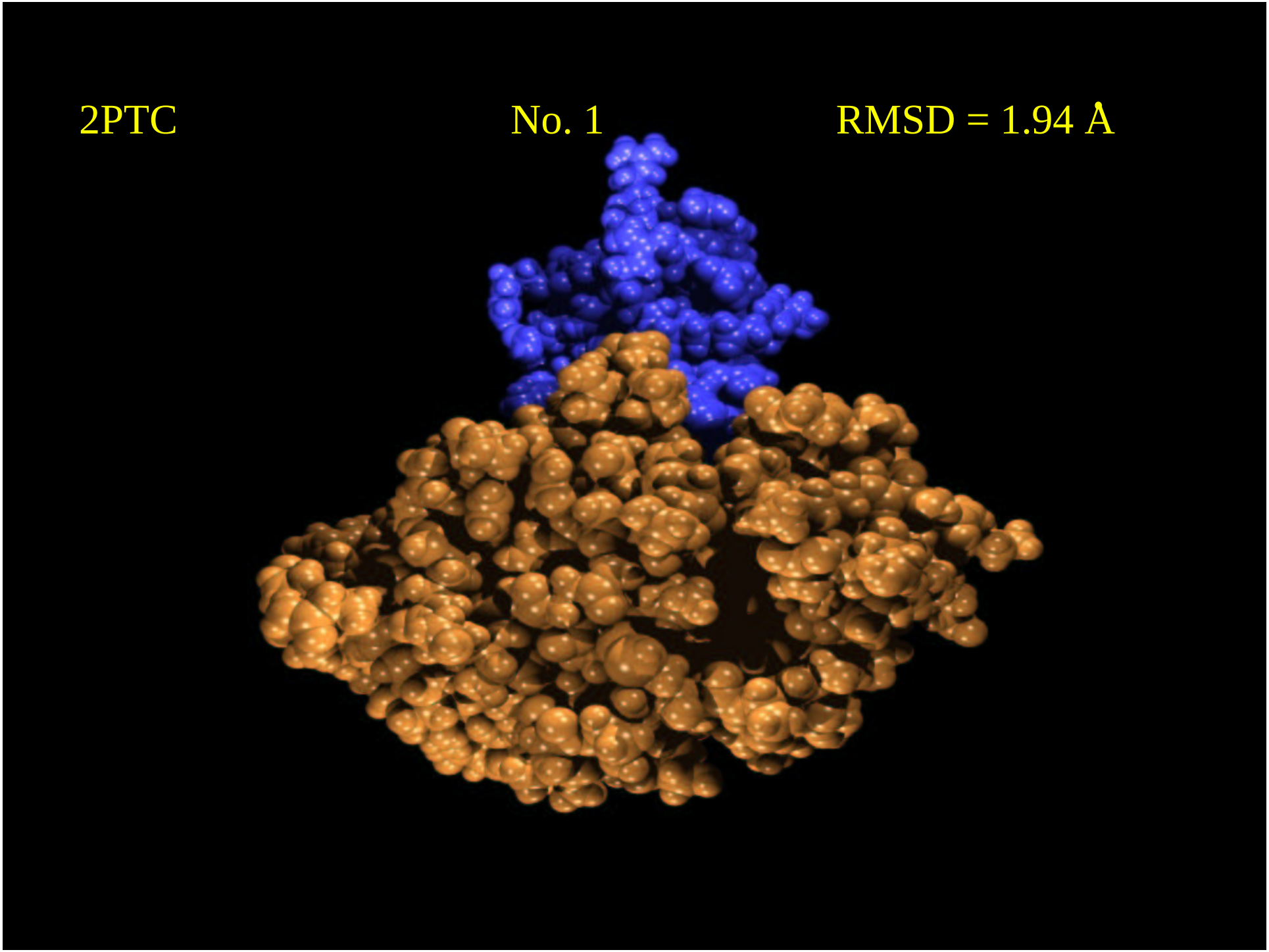




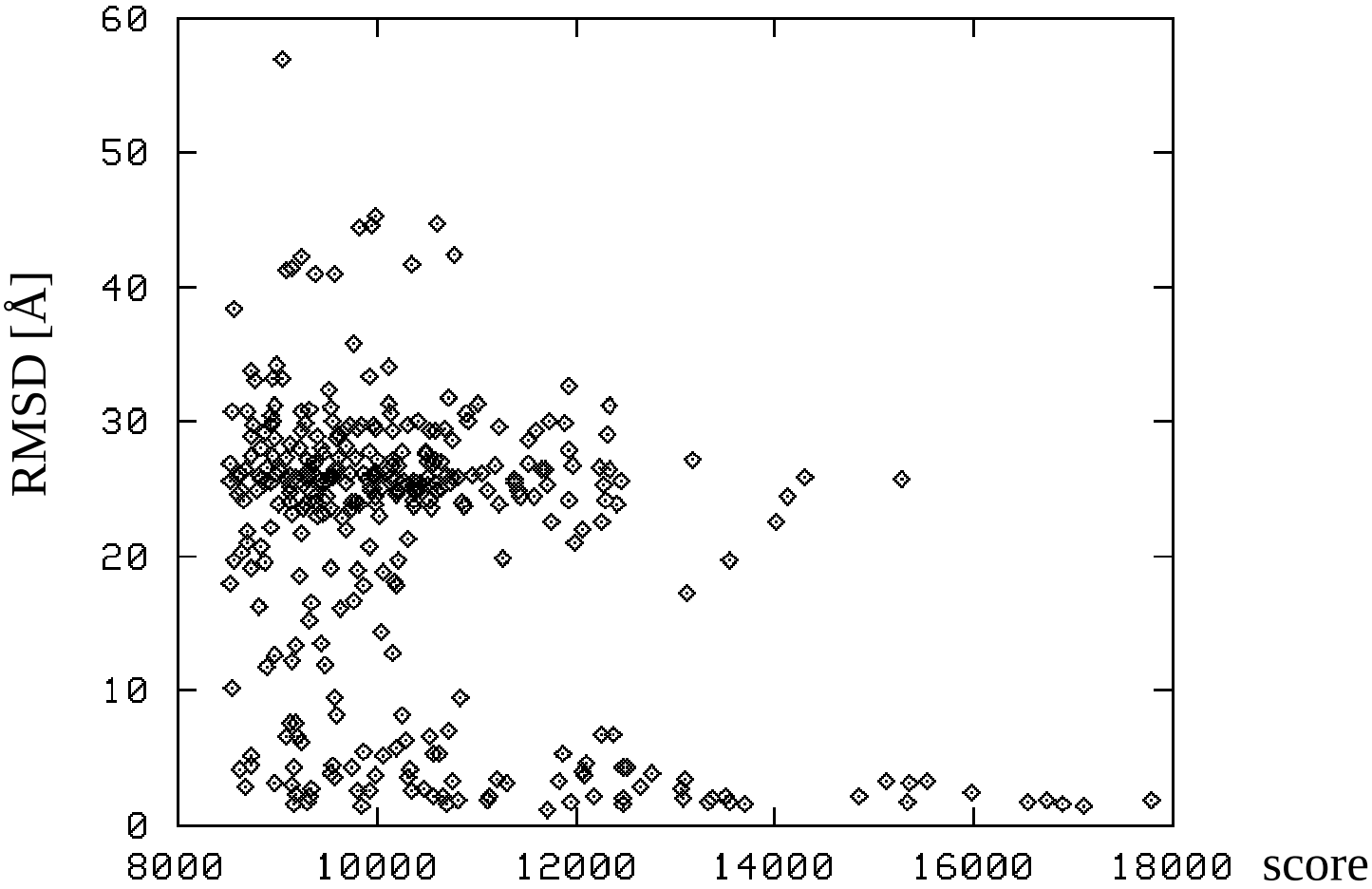



2PTC
B
A


2PTC
AB

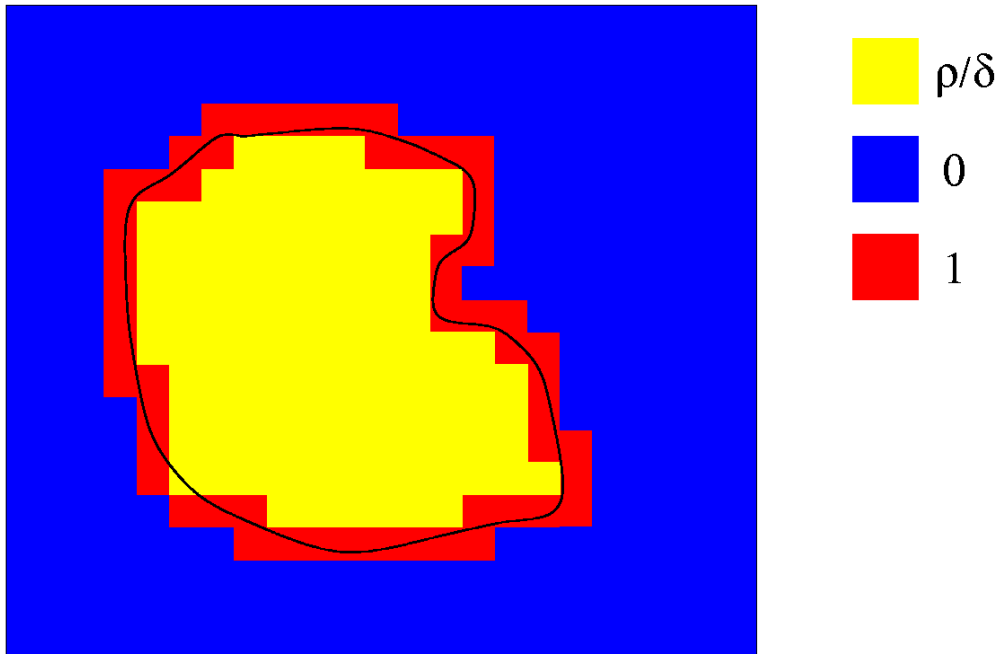

2PTC
No. 1
RMSD = 1.94 Å

Results: Docking of 2PTC



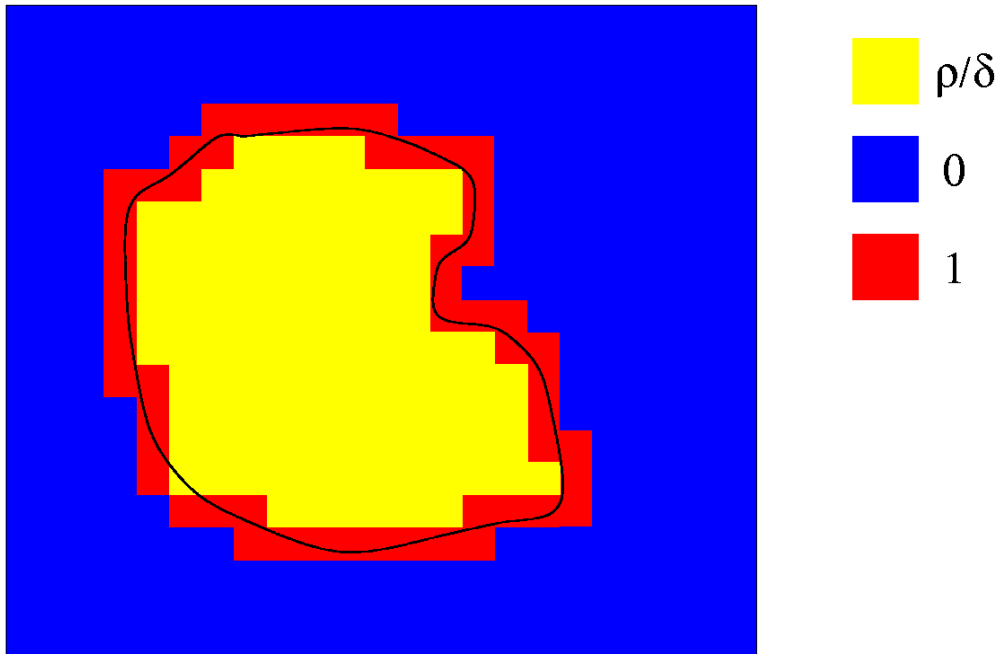
Basic Ideas

- Protein on grid



Basic Ideas

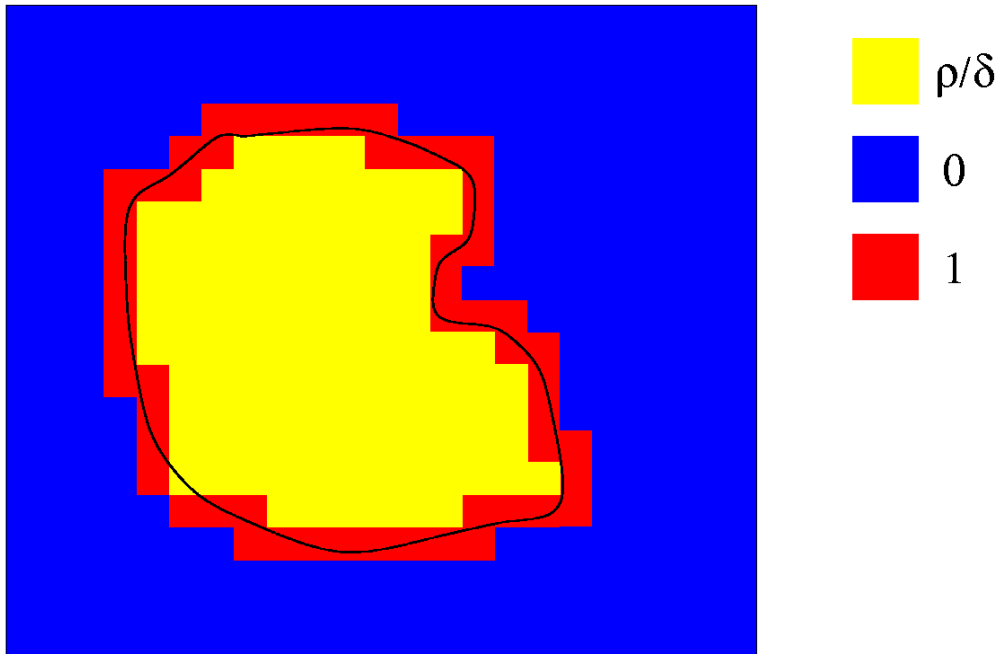
- Protein on grid
- Assign values
  - $a_{i,j,k}$ =
    - 1 at the surface of A
    - ρ << 0 inside A
    - 0 outside A
  - $b_{i,j,k}$ =
    - 1 at the surface of B
    - δ > 0 inside B
    - 0 outside B



A \ B	inside	surface	outside
inside	ρ* δ < 0	ρ < 0	0
surface	δ > 0	1	0
outside	0	0	0

Basic Ideas

- Protein on grid
- Assign values
  - $a_{i,j,k}$ =
    - 1 at the surface of A
    - ρ << 0 inside A
    - 0 outside
  - $b_{i,j,k}$ =
    - 1 at the surface of B
    - δ > 0 inside B
    - 0 outside B



A \ B	inside	surface	outside
inside	ρ* δ < 0	ρ < 0	0
surface	δ > 0	1	0
outside	0	0	0

Calculate Correlation $c_{\alpha,\beta,\gamma}$

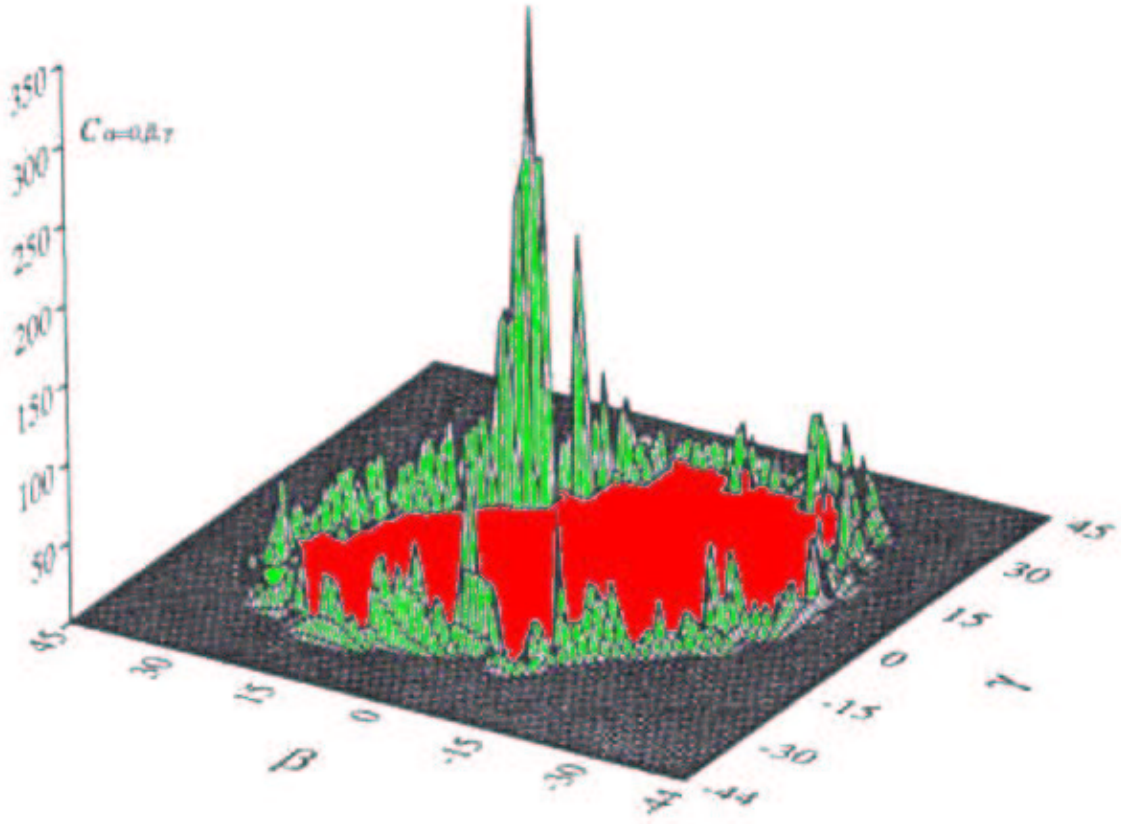
- Correlation of *a* and *b* identifies those translations where **A** and **B** are in contact:

$$c_{\alpha,\beta,\gamma} = \sum_{i,j,k} a_{i,j,k} \cdot b_{i+\alpha,\, j+\beta,\, k+\gamma}$$

- Repeat for all translation vectors (α, β, γ)

  ⇨ **Run time $O(N^6)$!**

Cross Section $c_{\alpha=0,\beta,\gamma}$



From: Katchalski-Katzir et al., PNAS 1992, 2195

Algorithm

- Calculate $a_{i,j,k}$ and $A^* = [\mathrm{FFT}(a)]^*$
- For all rotations of **B**:
  - Calculate $b_{i,j,k}$ and $B = \mathrm{FFT}(b)$
  - Calculate $C = A * B$
  - Calculate $c_{i,j,k} = \mathrm{IFT}(C)$
  - Identify tentative transformations (α, β, γ) as strongly positive peaks in c

Comparison of Algorithms

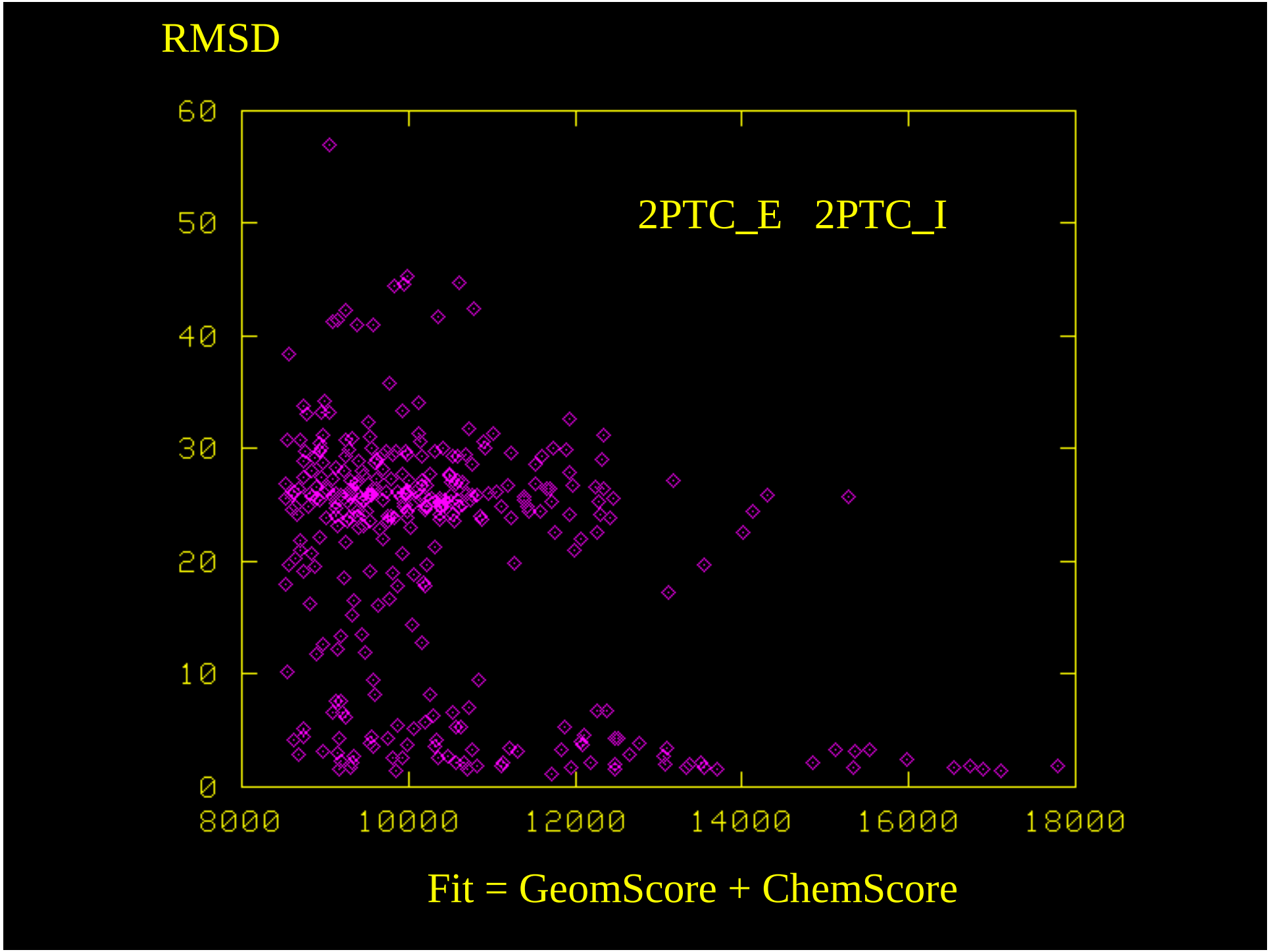
PDB ID	NLW+95	NLW+95	NLW+95	MWS96	AHP+96	L97
1CHO	2	2	2	1	3	
4CPA	106	3	3	1	5	
1FDL	124309	2900	3455	33	12	
2HFL	6792	24	25	28		
3HFM	17637	104	106	>125	5	
4HVP	2	2	2	1	10	
2KAI	672	125	229	1	9	
2MHB	49	2	2	1	8	
2PTC	161	2	2	1	1	
2SEC	3518	249	249	1	3	
4SGB	13691	72	72	1	13	
2SNI	81	3	3	1	5	
1TEC	95	6	6	1	8	
2TGP	180	2	2	1		
1TGS	552	2	2	1	3	
4TPI	11	2	2	1	3	

Comparison of Algorithms

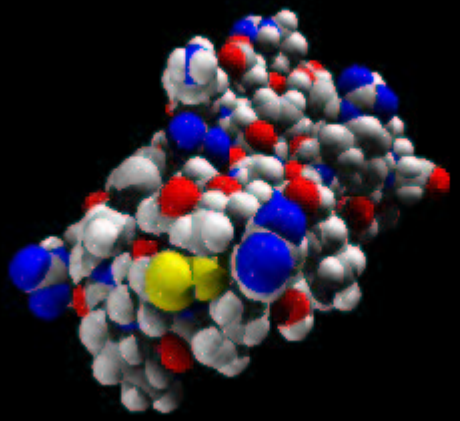
PDB ID	NLW+95	NLW+95	NLW+95	MWS96	AHP+96	L97
1CHO	2	2	2	1	3	1
4CPA	106	3	3	1	5	1
1FDL	124309	2900	3455	33	12	17
2HFL	6792	24	25	28		1
3HFM	17637	104	106	>125	5	1
4HVP	2	2	2	1	10	1
2KAI	672	125	229	1	9	1
2MHB	49	2	2	1	8	1
2PTC	161	2	2	1	1	1
2SEC	3518	249	249	1	3	1
4SGB	13691	72	72	1	13	1
2SNI	81	3	3	1	5	1
1TEC	95	6	6	1	8	1
2TGP	180	2	2	1		1
1TGS	552	2	2	1	3	1
4TPI	11	2	2	1	3	1

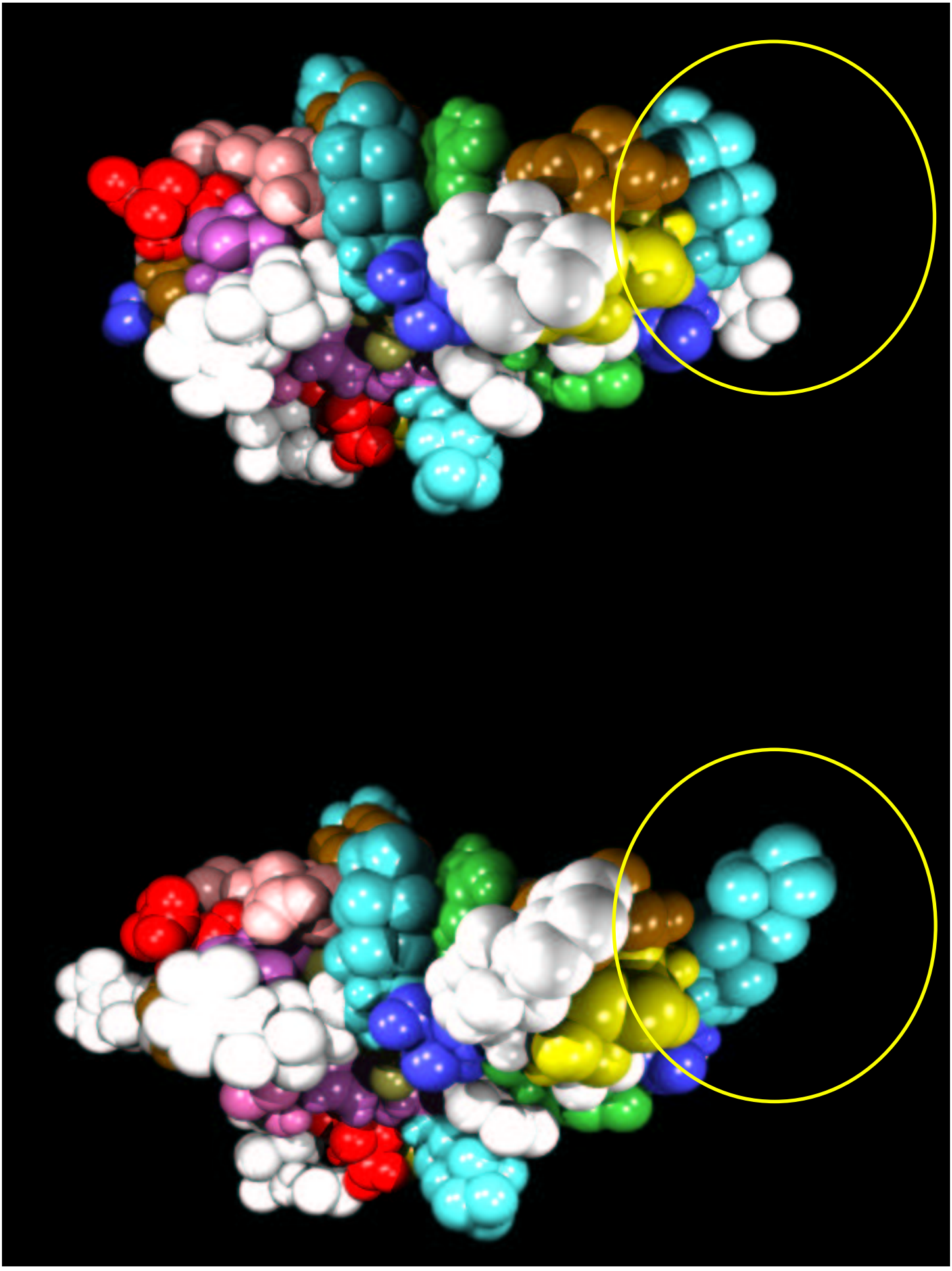
Comparison of Algorithms

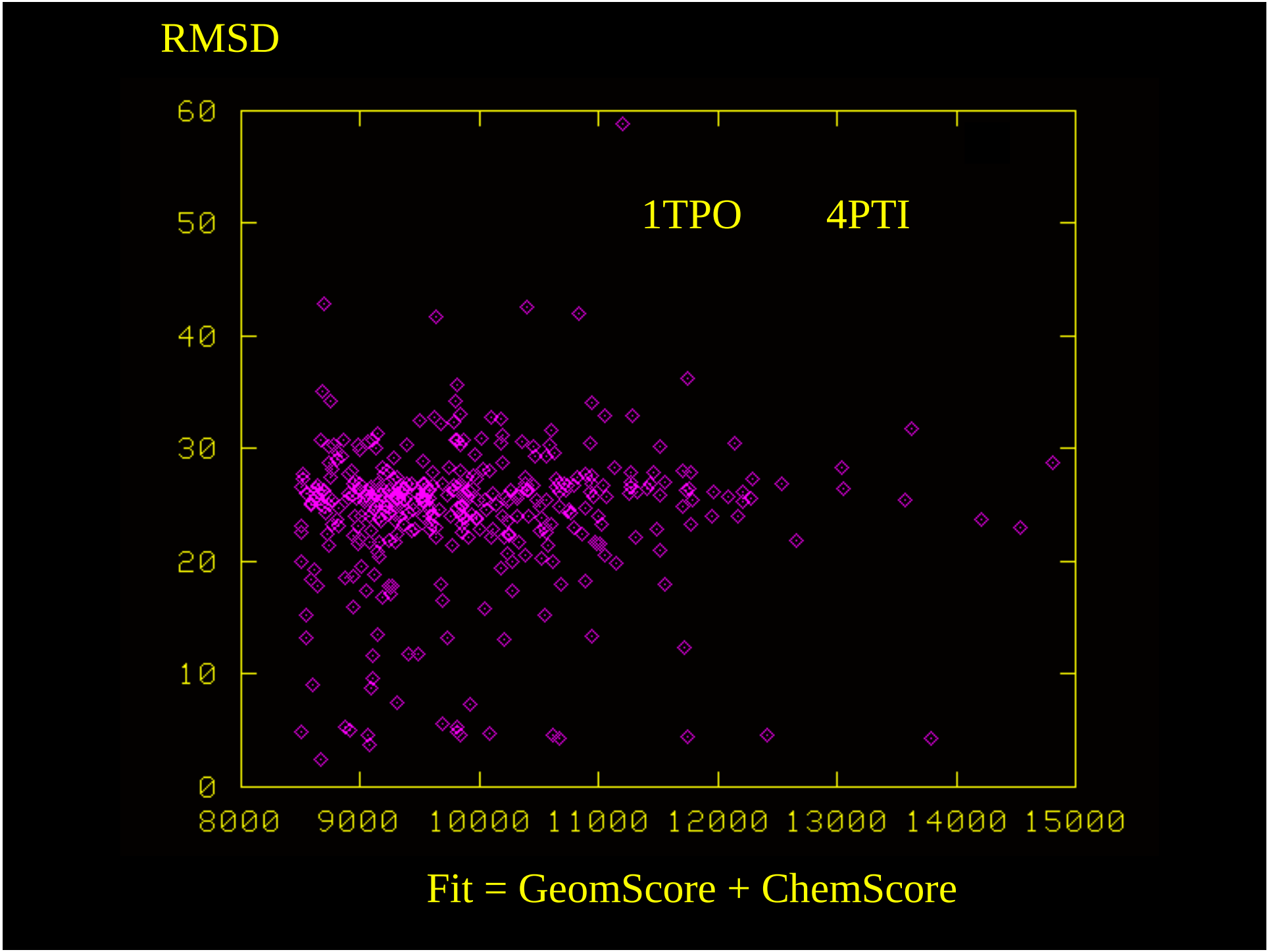
PDB ID	NLW+95	NLW+95	NLW+95	MWS96	AHP+96	L97
1CHO	2	2	2	1	3	1
4CPA	106	3	3	1	5	1
1FDL	124309	2900	3455	33	12	17
2HFL	6792	24	25	28		1
3HFM	17637	104	106	>125	5	1
4HVP	2	2	2	1	10	1
2KAI	672	125	229	1	9	1
2MHB	49	2	2	1	8	1
2PTC	**161**	**2**	**2**	**1**	**1**	**1**
2SEC	3518	249	249	1	3	1
4SGB	13691	72	72	1	13	1
2SNI	81	3	3	1	5	1
1TEC	95	6	6	1	8	1
2TGP	180	2	2	1		1
1TGS	552	2	2	1	3	1
4TPI	11	2	2	1	3	1

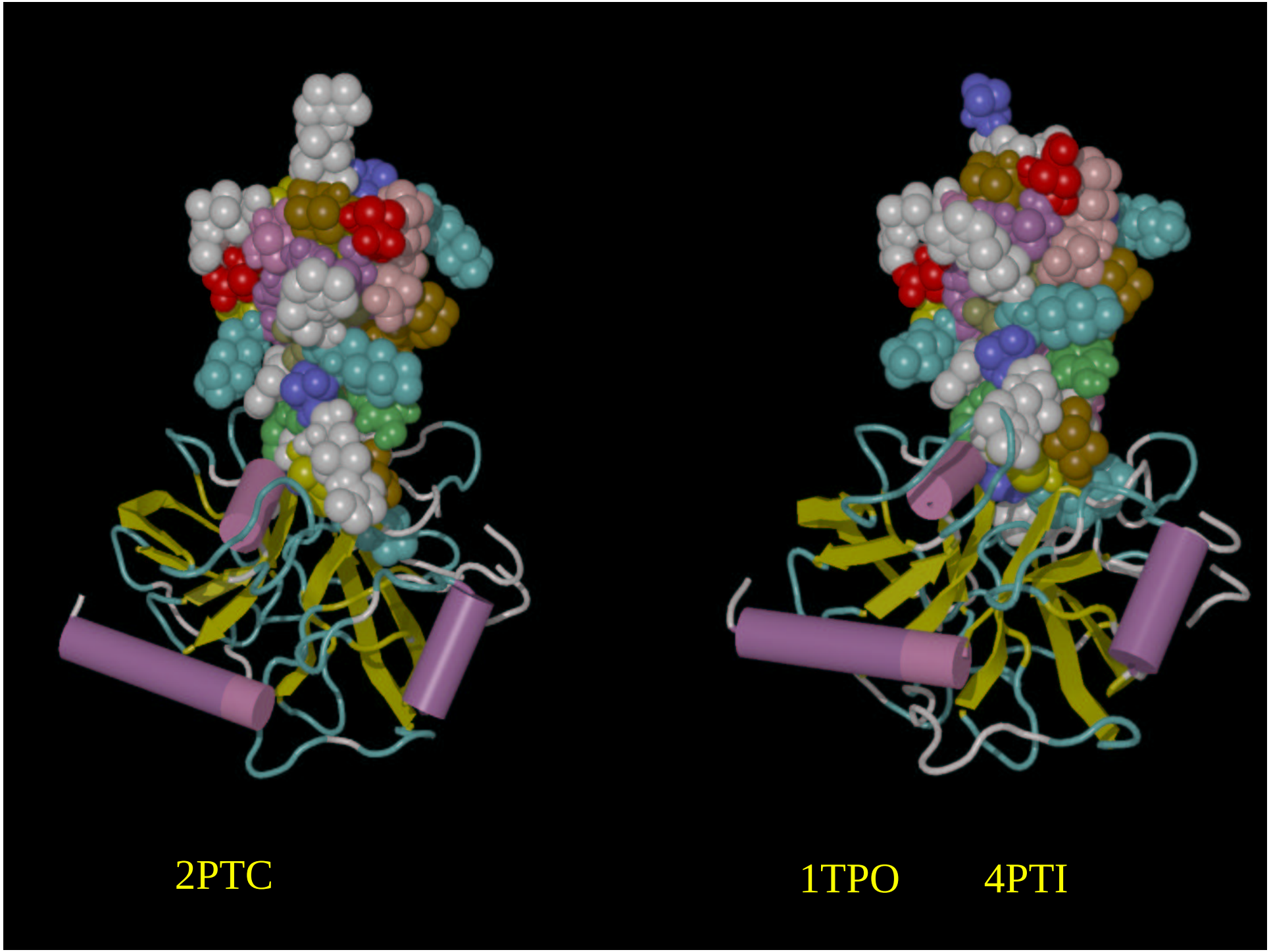

RMSD
60
50
40
30
20
10
0
8000
10000
12000
14000
16000
18000
2PTC_E 2PTC_I
Fit = GeomScore + ChemScore

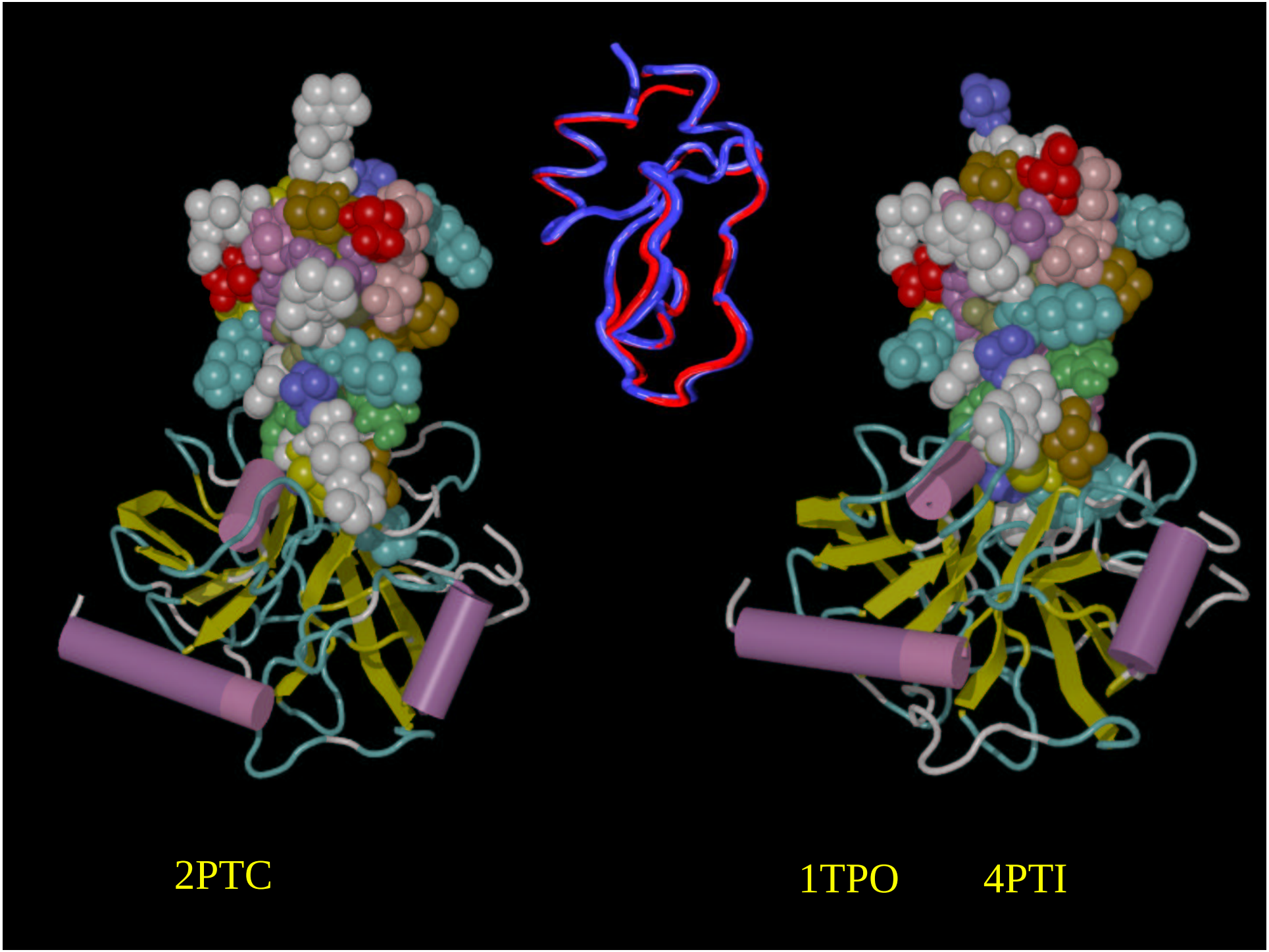
One PS in the Life of PTI

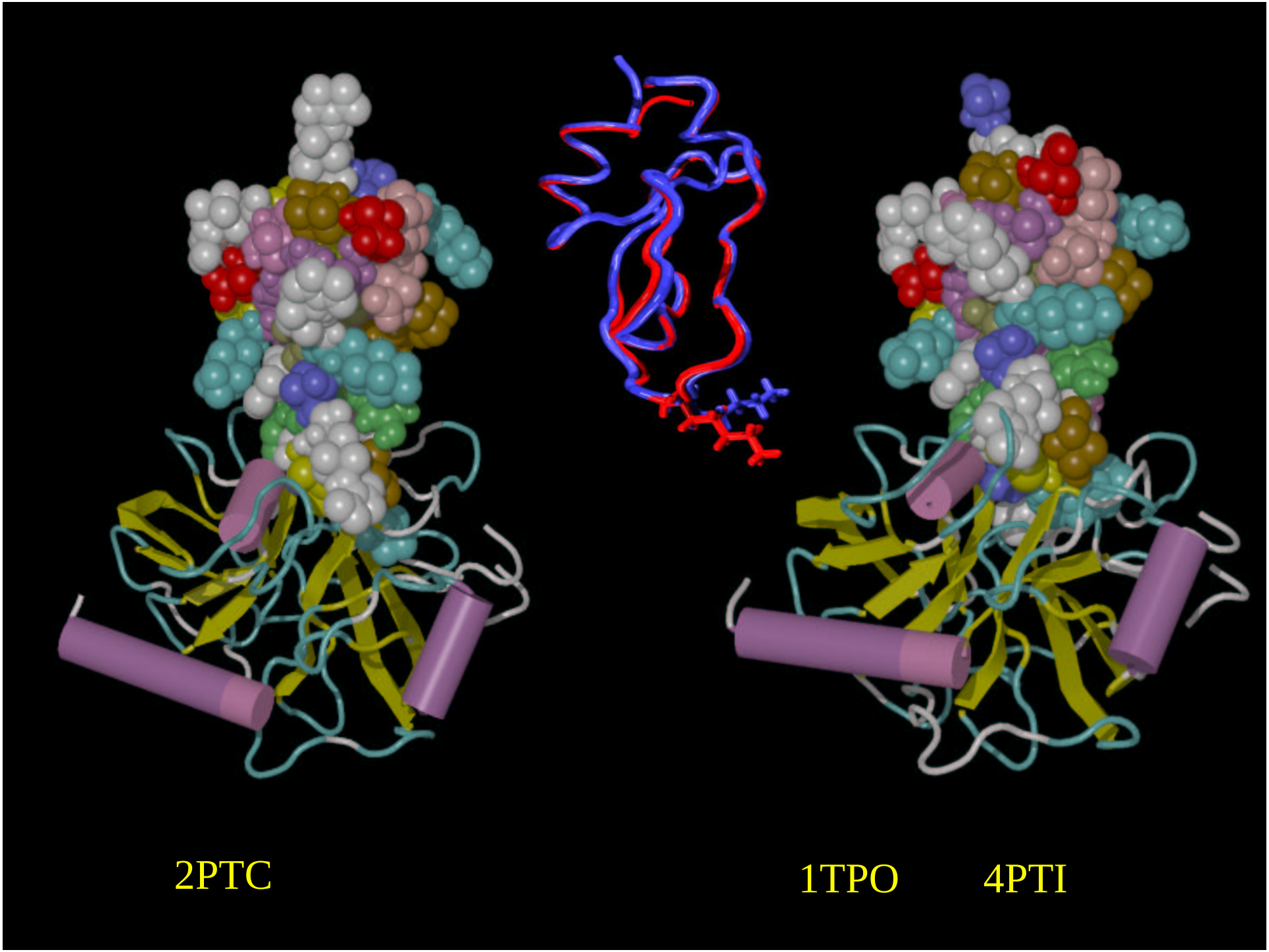





RMSD
60
50
40
30
20
10
0
1TPO
4PTI
8000
9000
10000
11000
12000
13000
14000
15000
Fit = GeomScore + ChemScore


2PTC
1TPO
4PTI


2PTC
1TPO
4PTI


2PTC
1TPO
4PTI

Protein Flexibility

- Domain movements
  - Large scale movements of protein domains
  - Mainly backbone movements
  - Highly flexible “hinge” regions: hinge bending
- Side-chain flexibility
  - Rather rigid backbone
  - Local rearrangements of side chains
  - Flexibility based on side-chain torsions

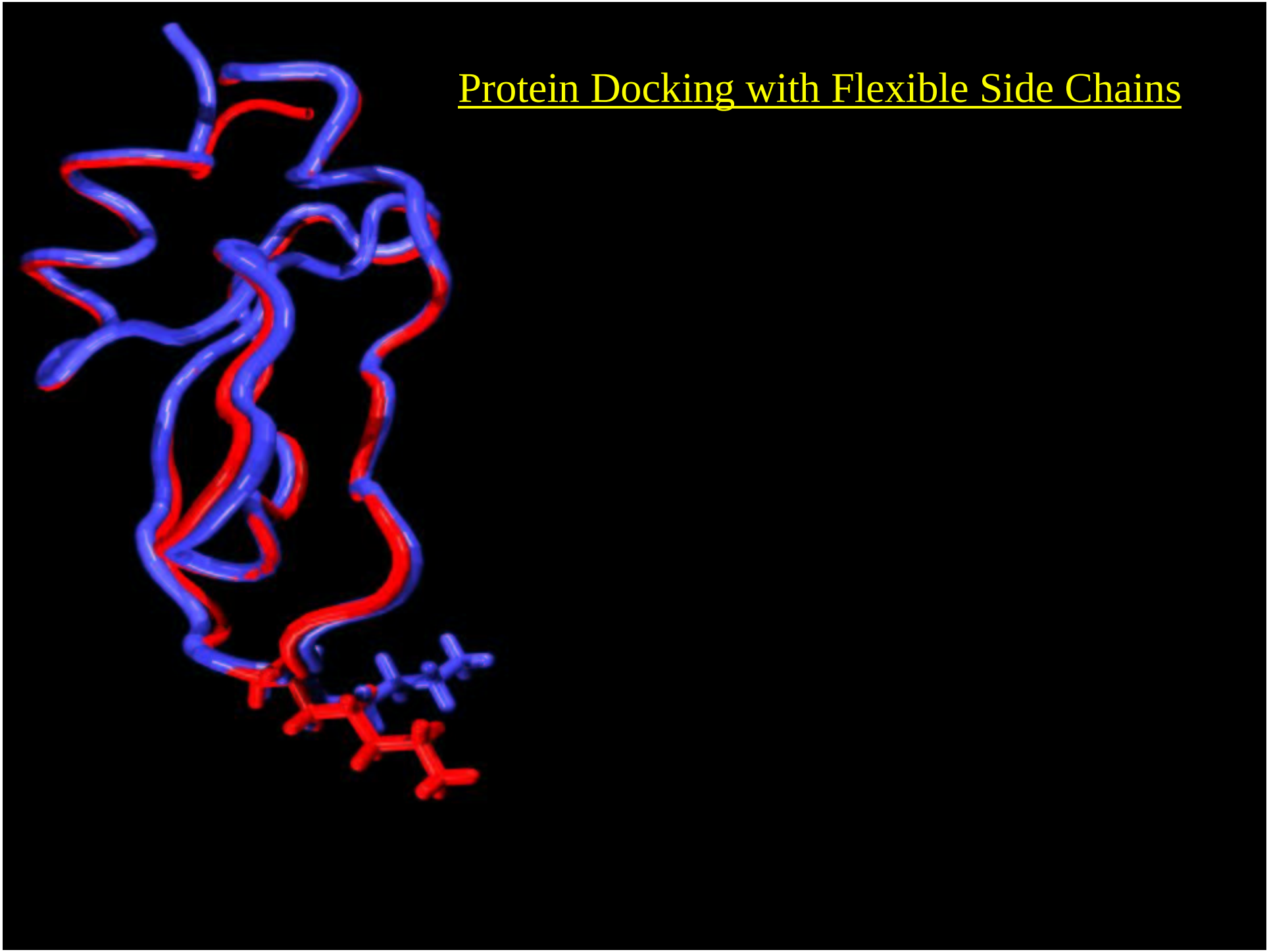
Semi-Flexible Docking

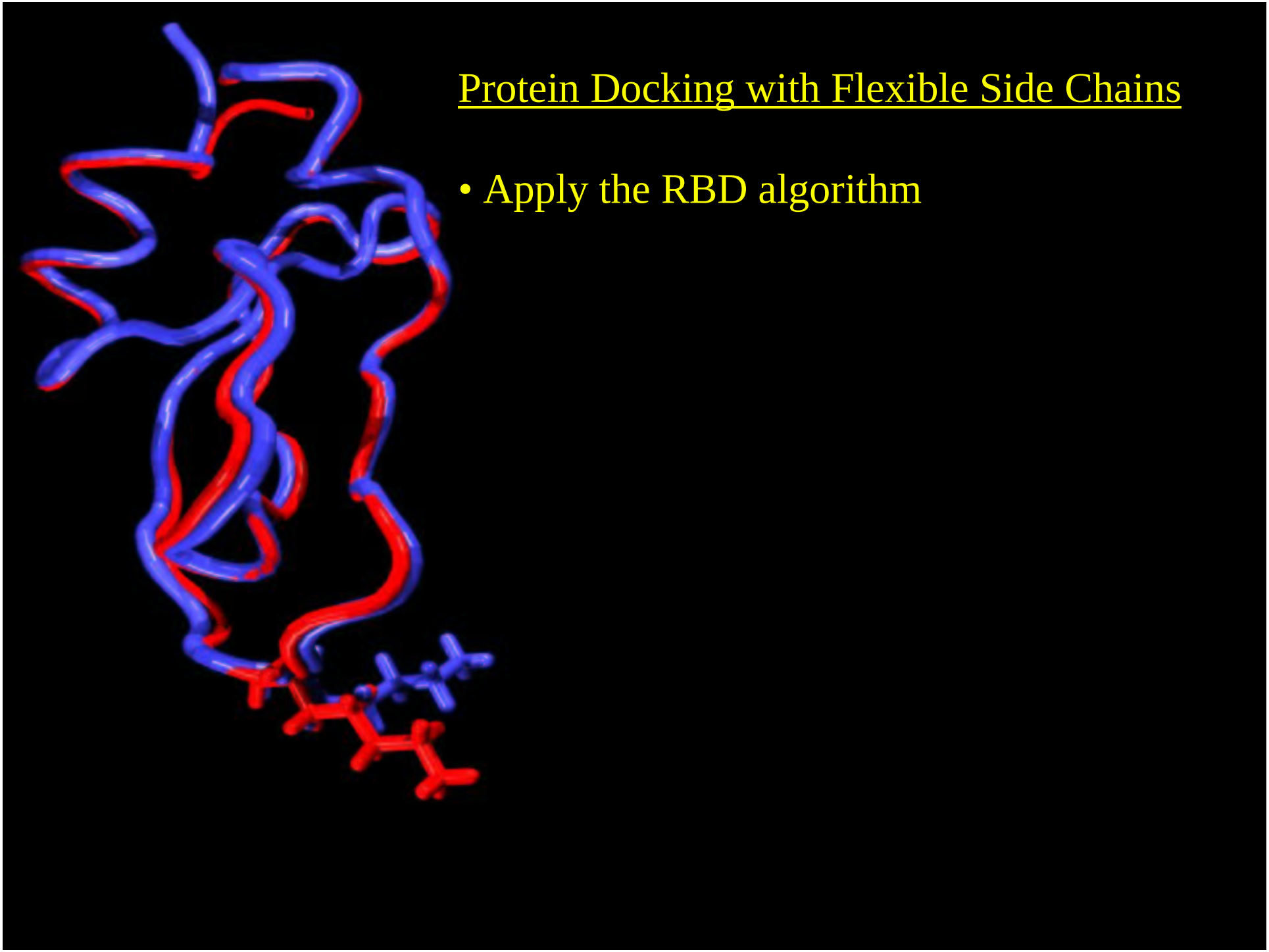
- Assumptions
  - Rigid backbone
  - Flexible side chains

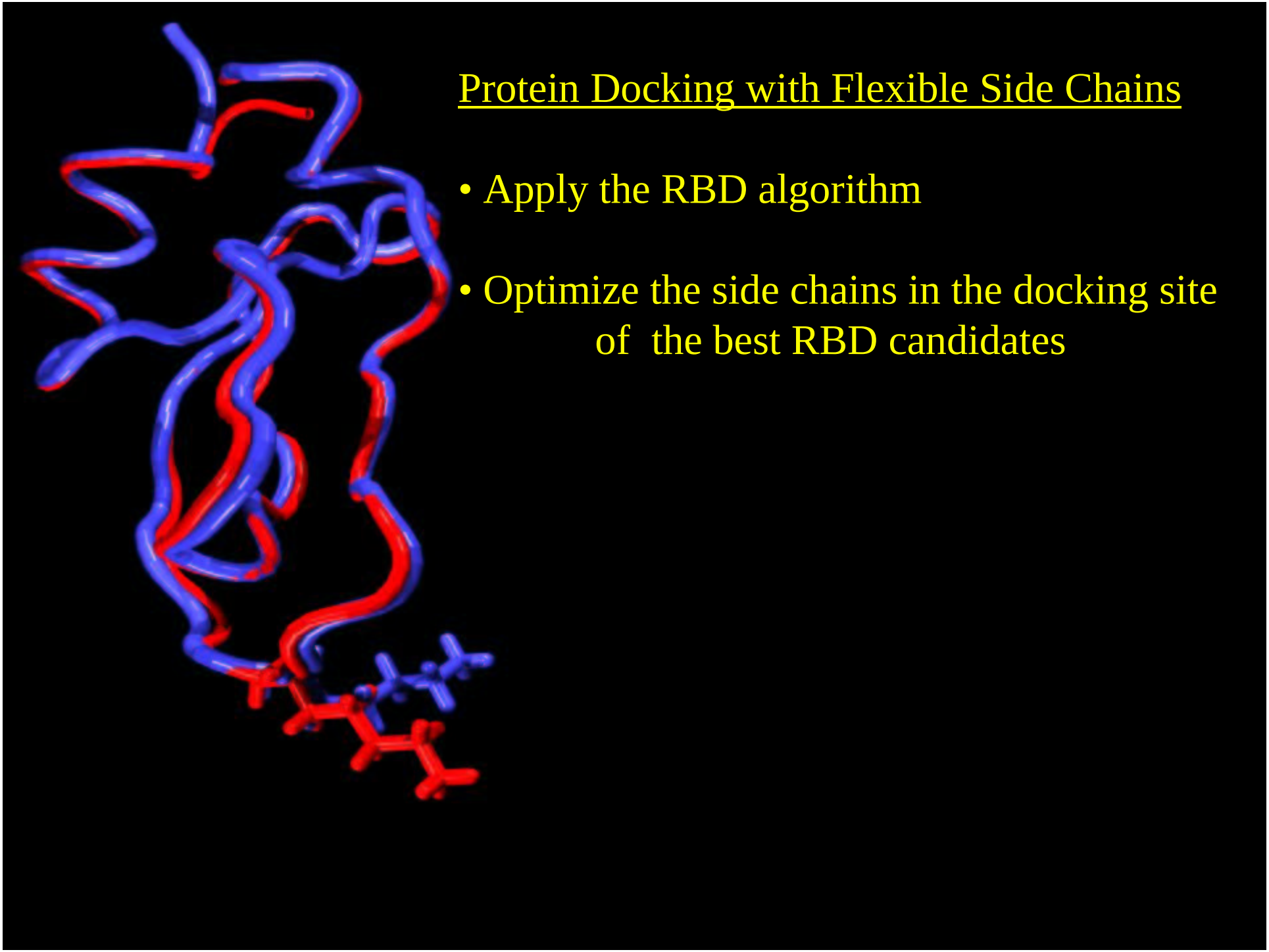
Good approximation for Trypsin/BPTI!

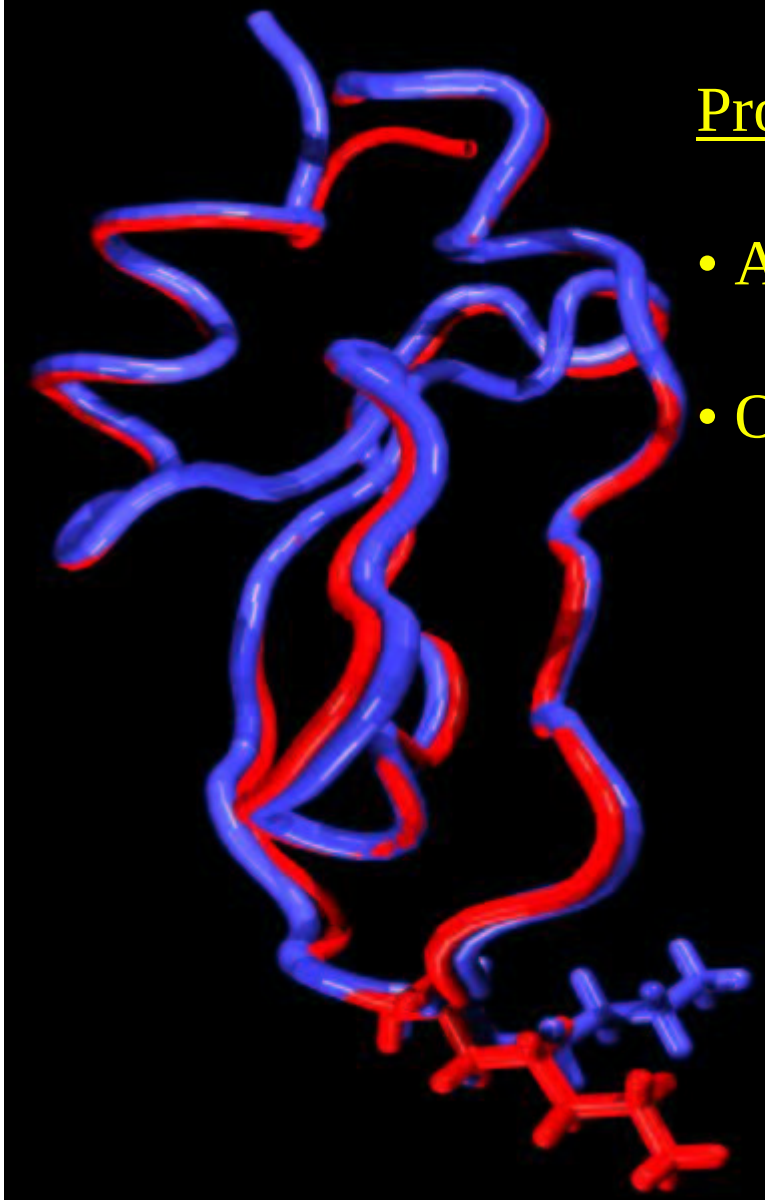
- Basic Algorithm
  - Start with rigid docking candidates
  - Demangle overlapping side chains
  - Calculate energy

Protein Docking with Flexible Side Chains




Protein Docking with Flexible Side Chains
• Apply the RBD algorithm

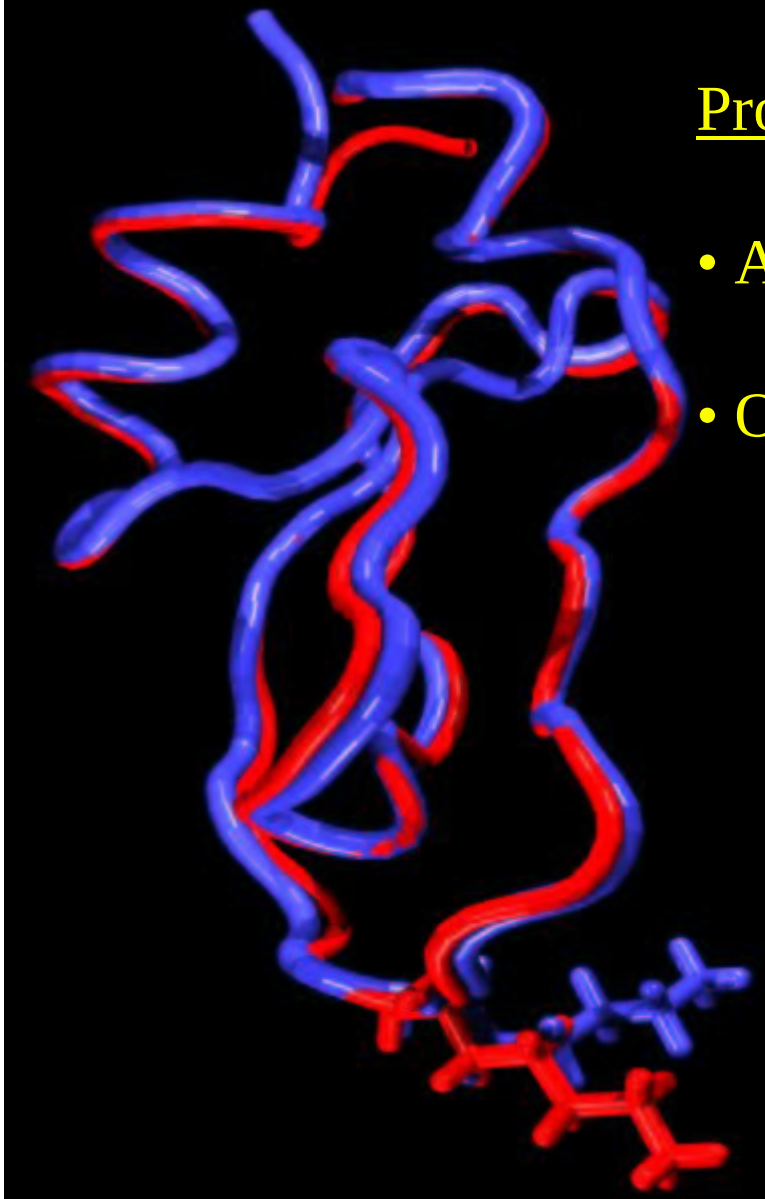

Protein Docking with Flexible Side Chains
• Apply the RBD algorithm
• Optimize the side chains in the docking site of the best RBD candidates



Protein Docking with Flexible Side Chains

- Apply the RBD algorithm
- Optimize the side chains in the docking site of the best RBD candidates

w.r.t. the potential energy (AMBER)

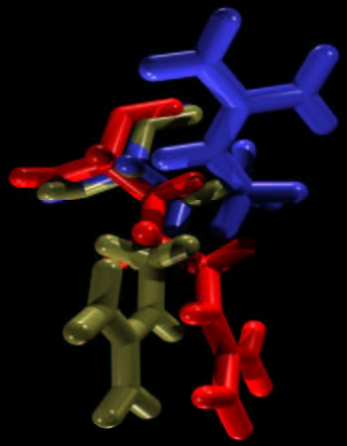


Protein Docking with Flexible Side Chains

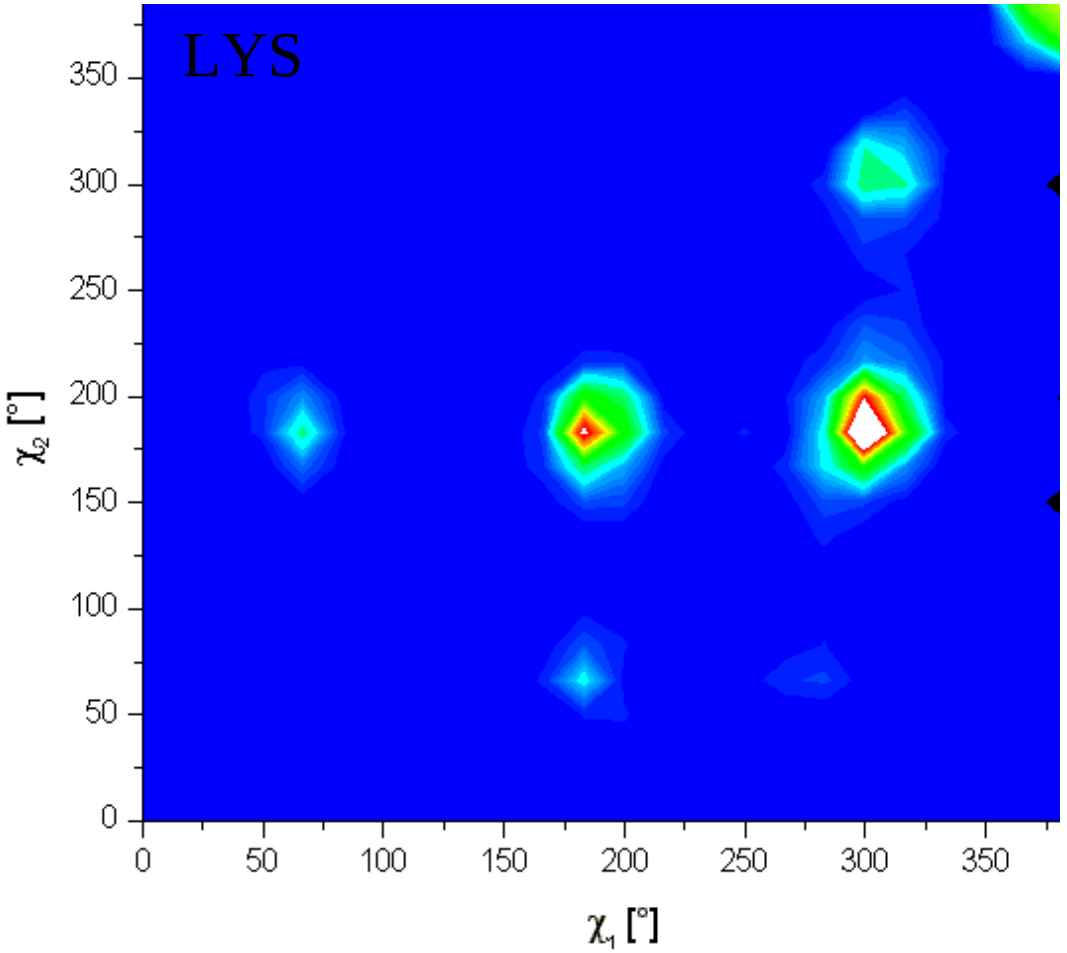
- Apply the RBD algorithm
- Optimize the side chains in the docking site of the best RBD candidates

  w.r.t. the potential energy (AMBER)

  using a rotamer library

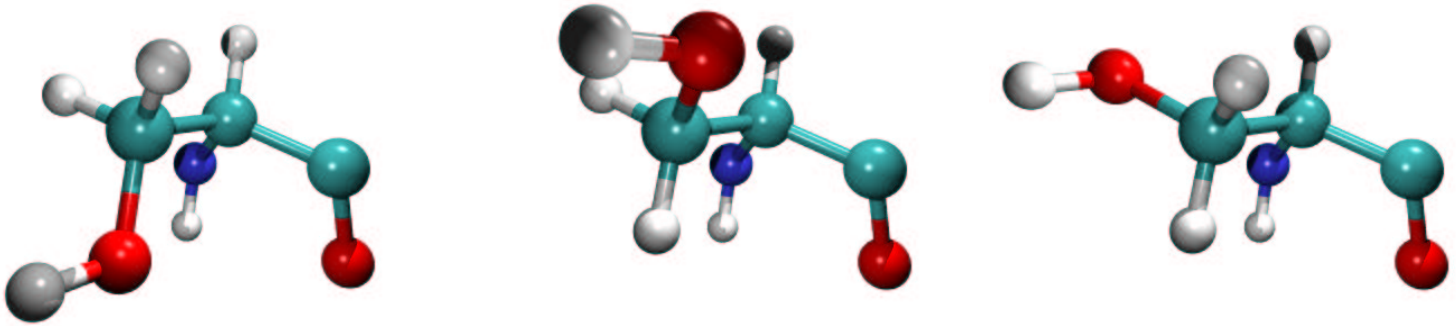


Torsion angle distribution

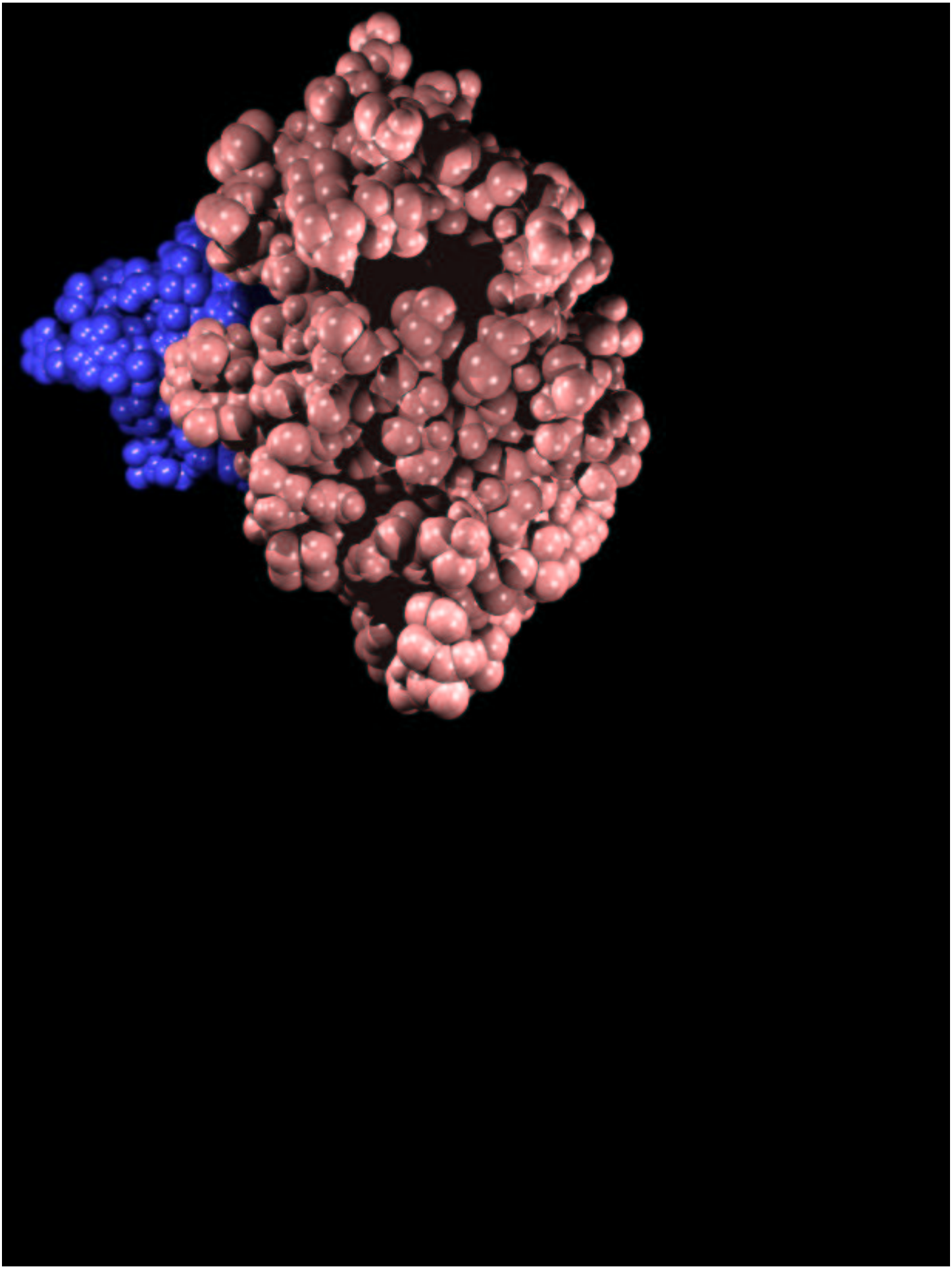


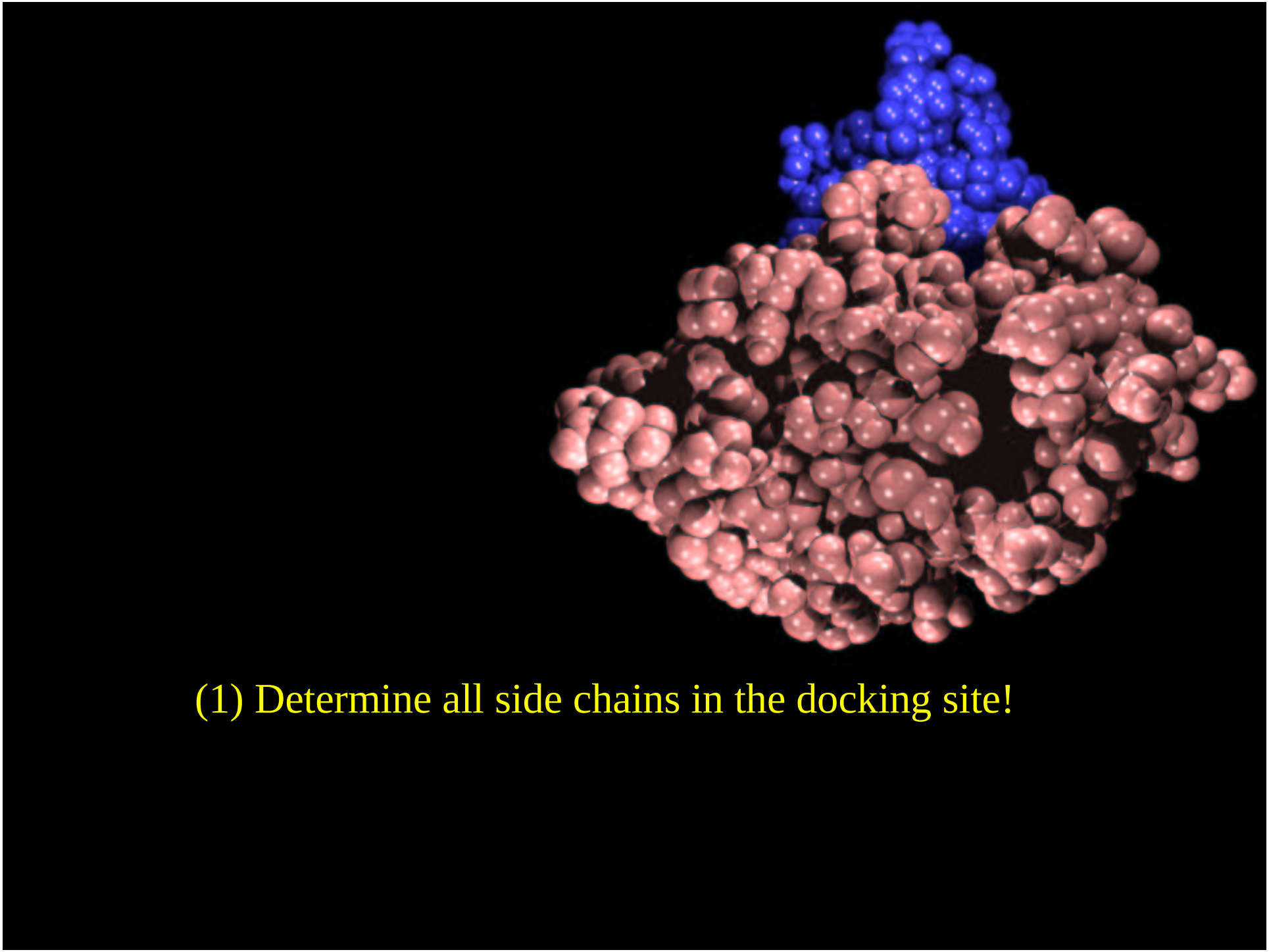
Rotamer Library

Side-chain conformational space is adequately represented by a discrete set of *rotamers*

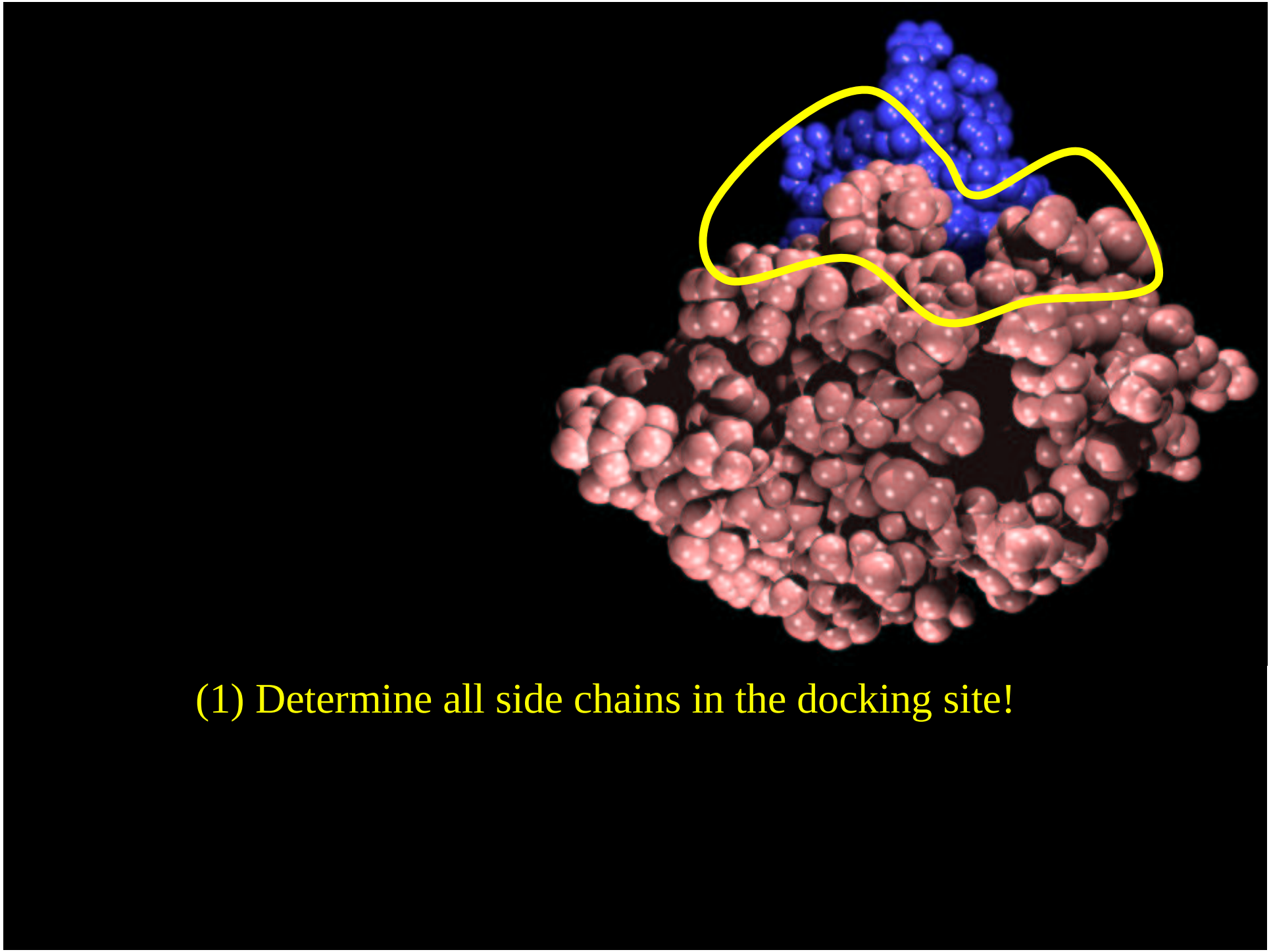


SER





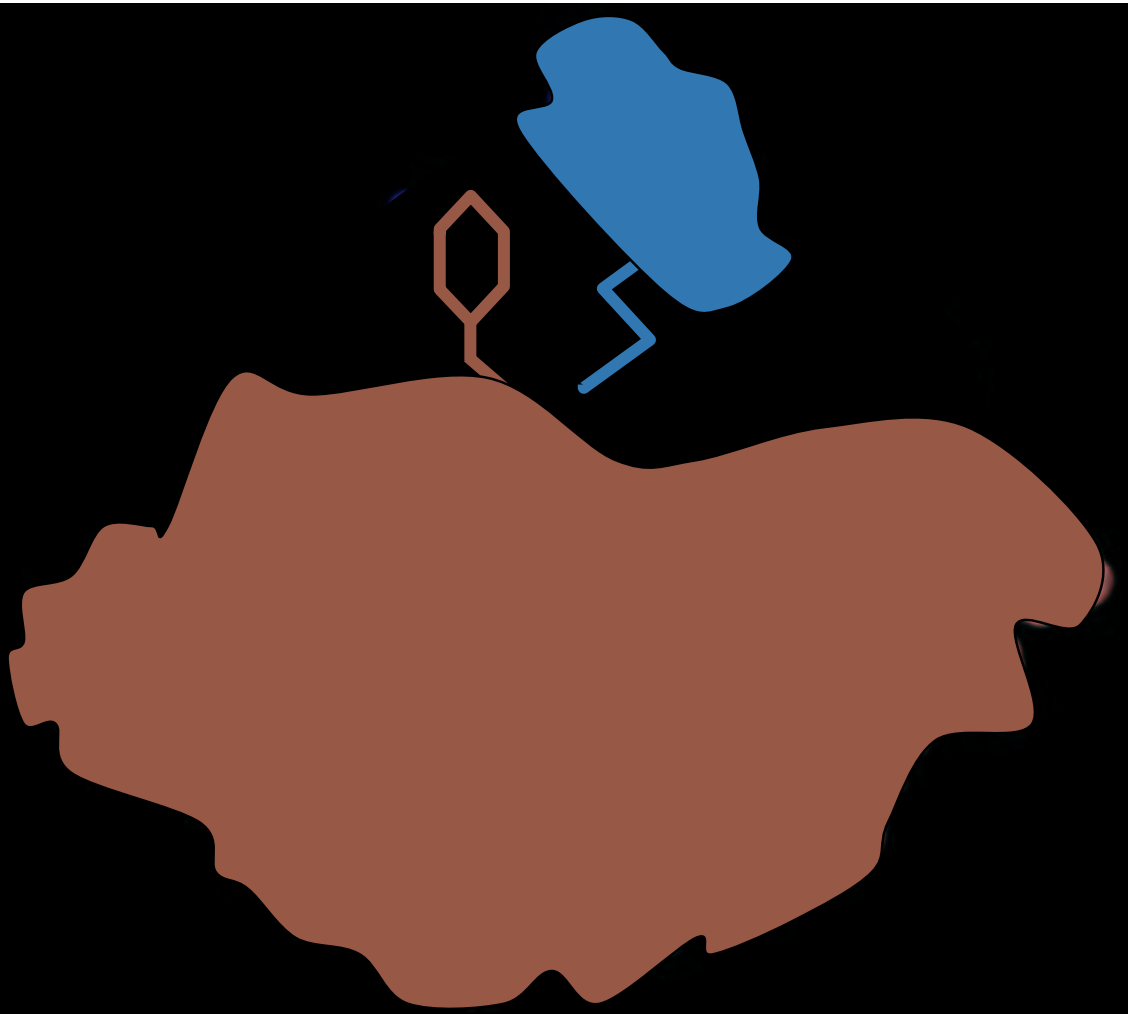
(1) Determine all side chains in the docking site!



(1) Determine all side chains in the docking site!

S1

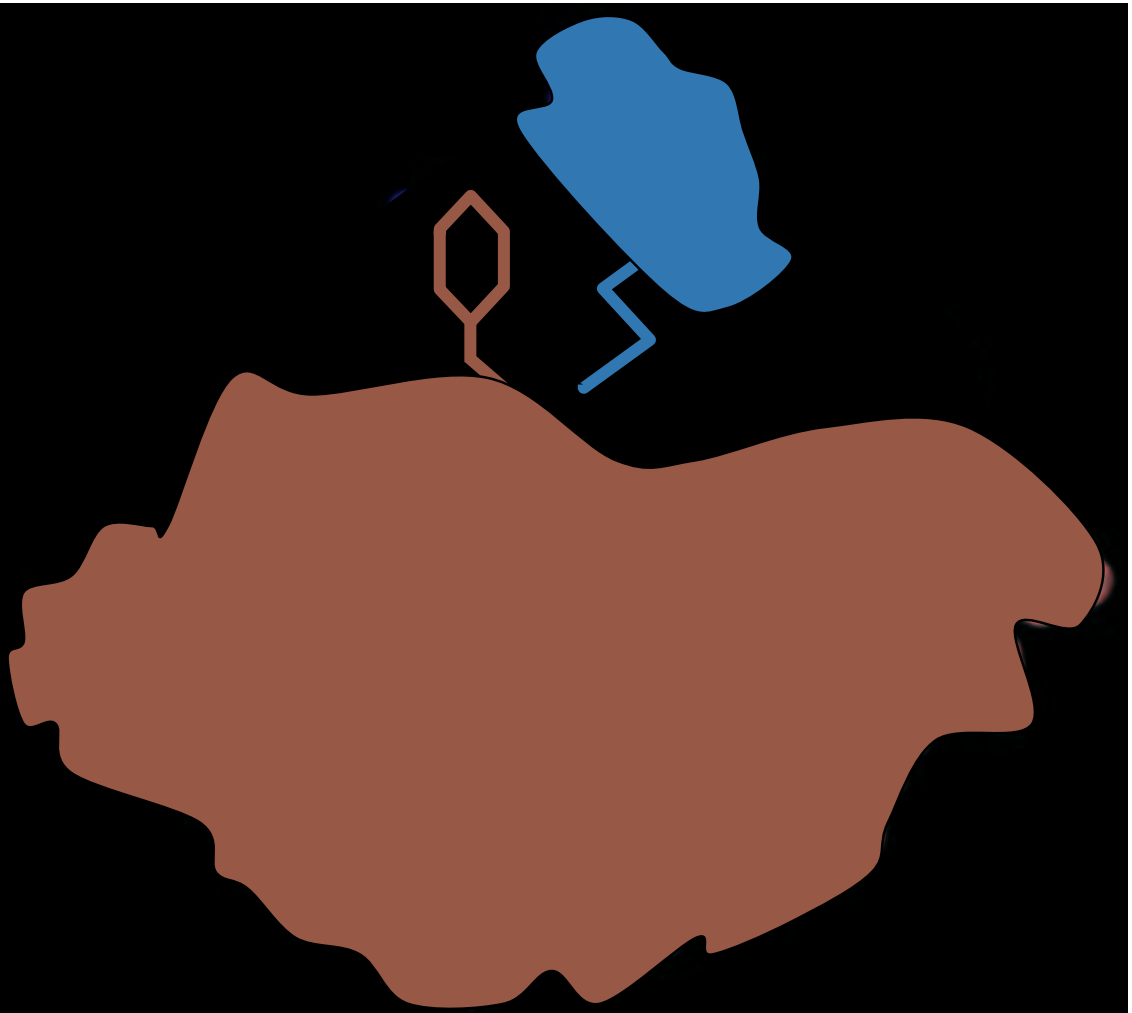
S2



(1) Determine all side chains in the docking site!

S1

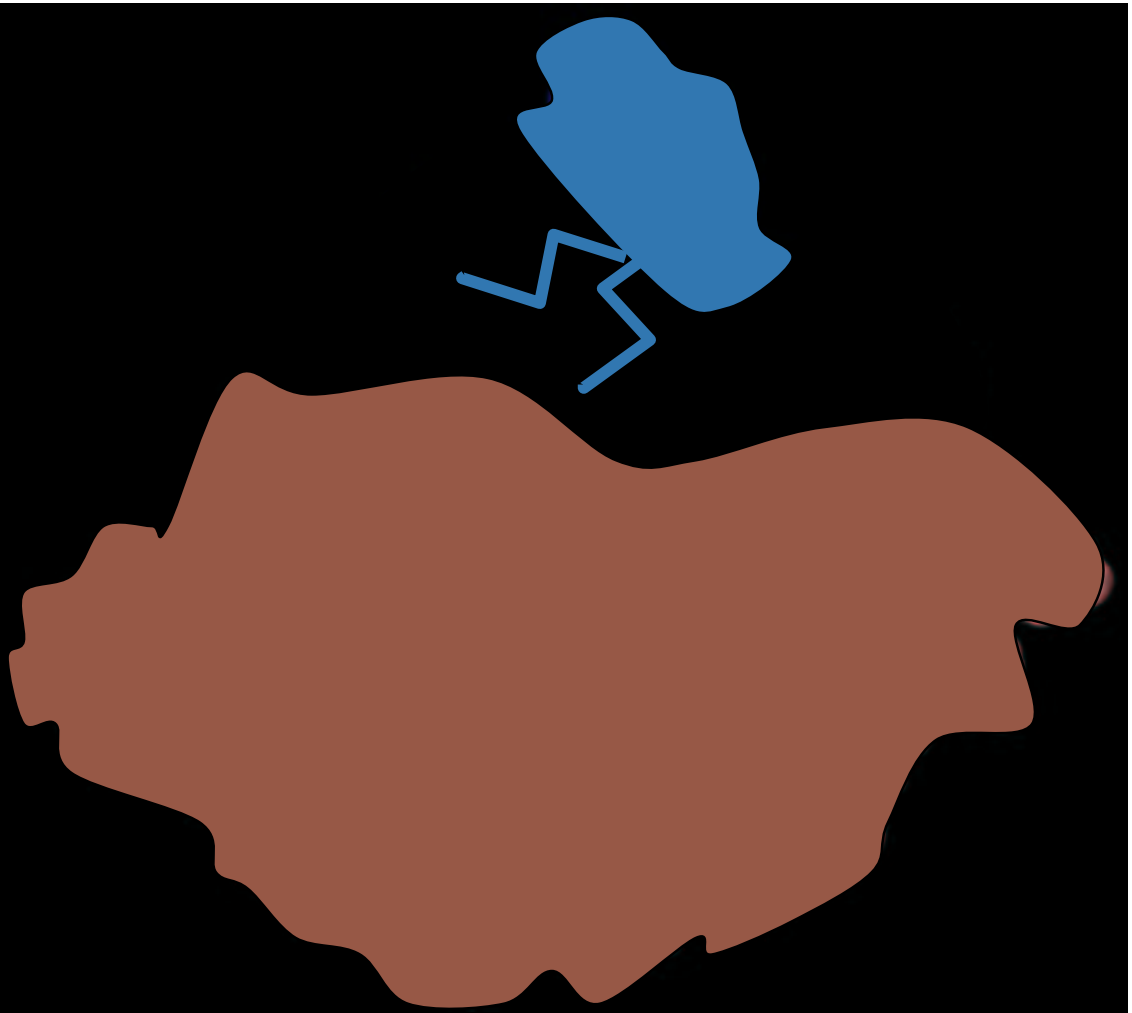
S2



(1) Determine all side chains in the docking site!
(2) Build sets of potential side chain orientation

S1 R1 R2

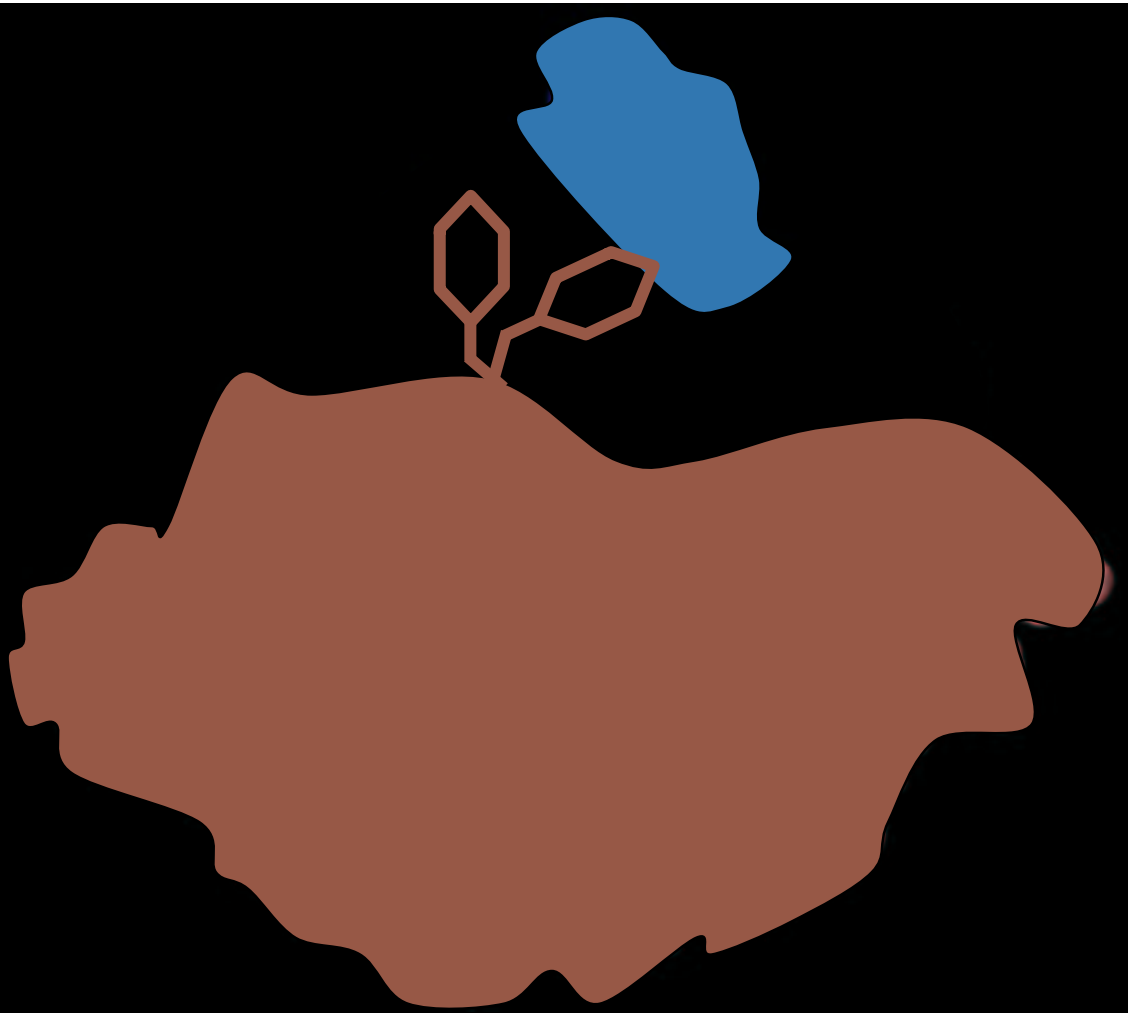
S2



(1) Determine all side chains in the docking site!
(2) Build sets of potential side chain orientation

S1 R1 R2

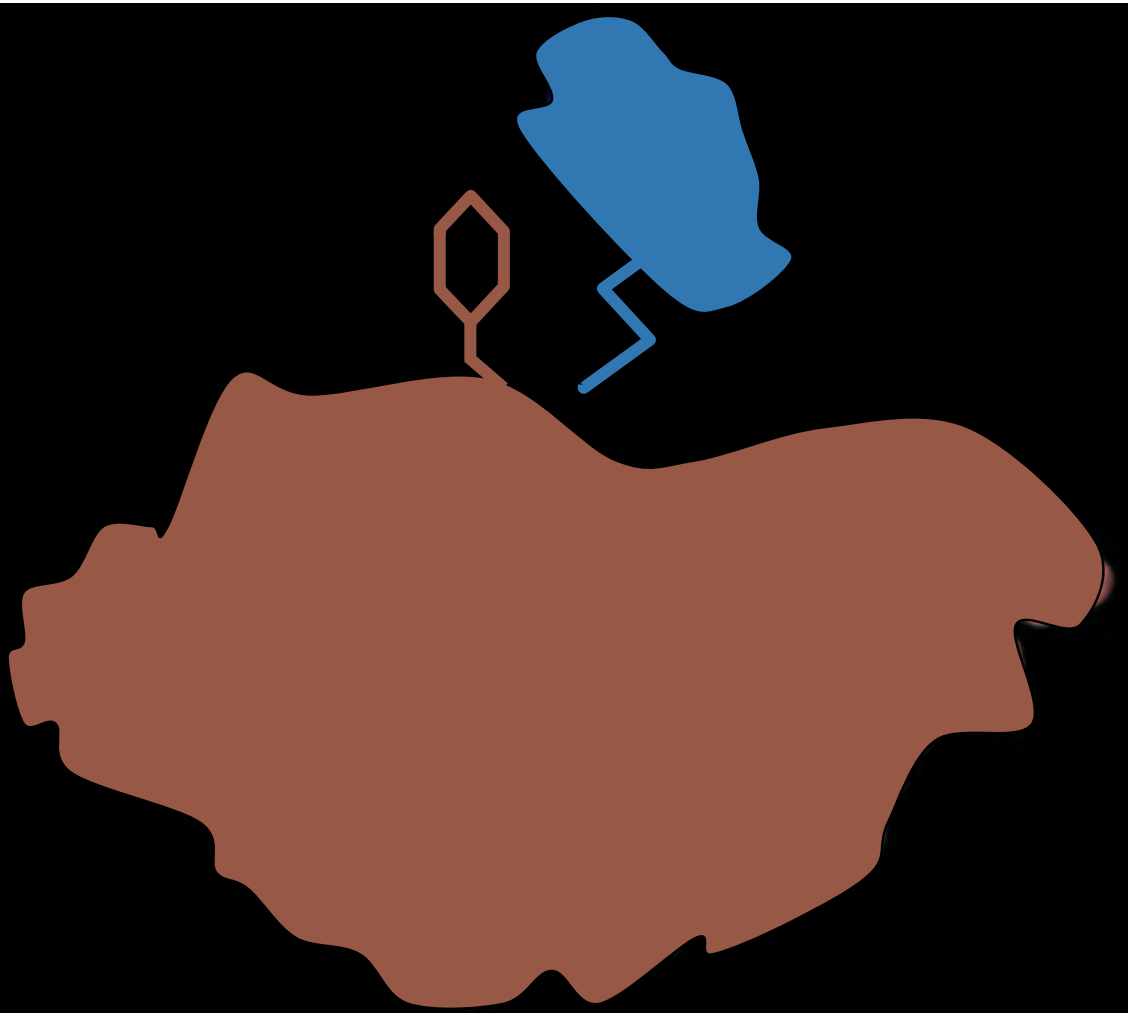
S2 R1 R2



(1) Determine all side chains in the docking site!
(2) Build sets of potential side chain orientation

S1 R1 R2

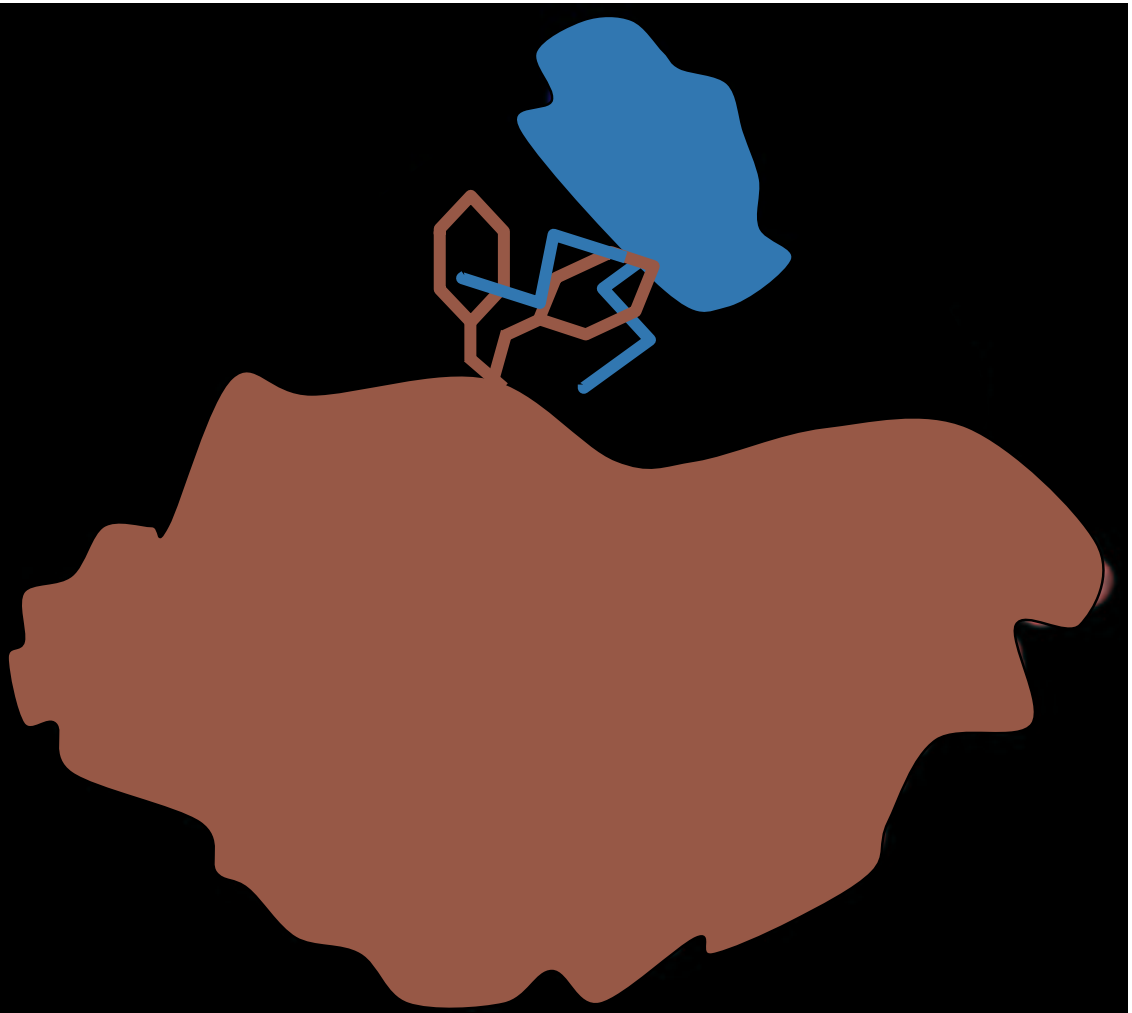
S2 R1 R2



(1) Determine all side chains in the docking site!
(2) Build sets of potential side chain orientation
(3) Determine the rotamer combination with minimal energy!

S1 R1 R2

S2 R1 R2

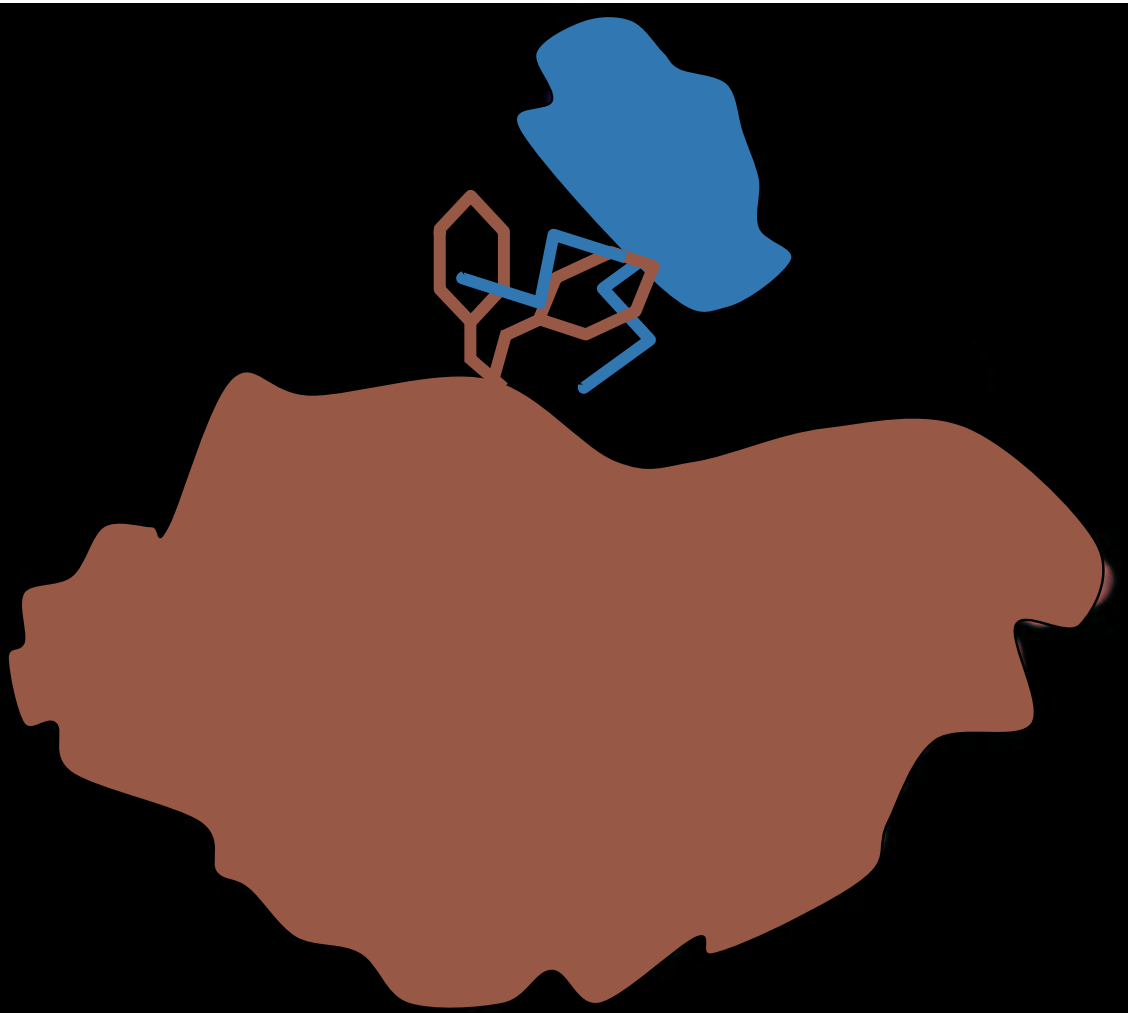


X_{11} X_{12}

S1 R1 R2

S2 R1 R2

X_{21} X_{22}



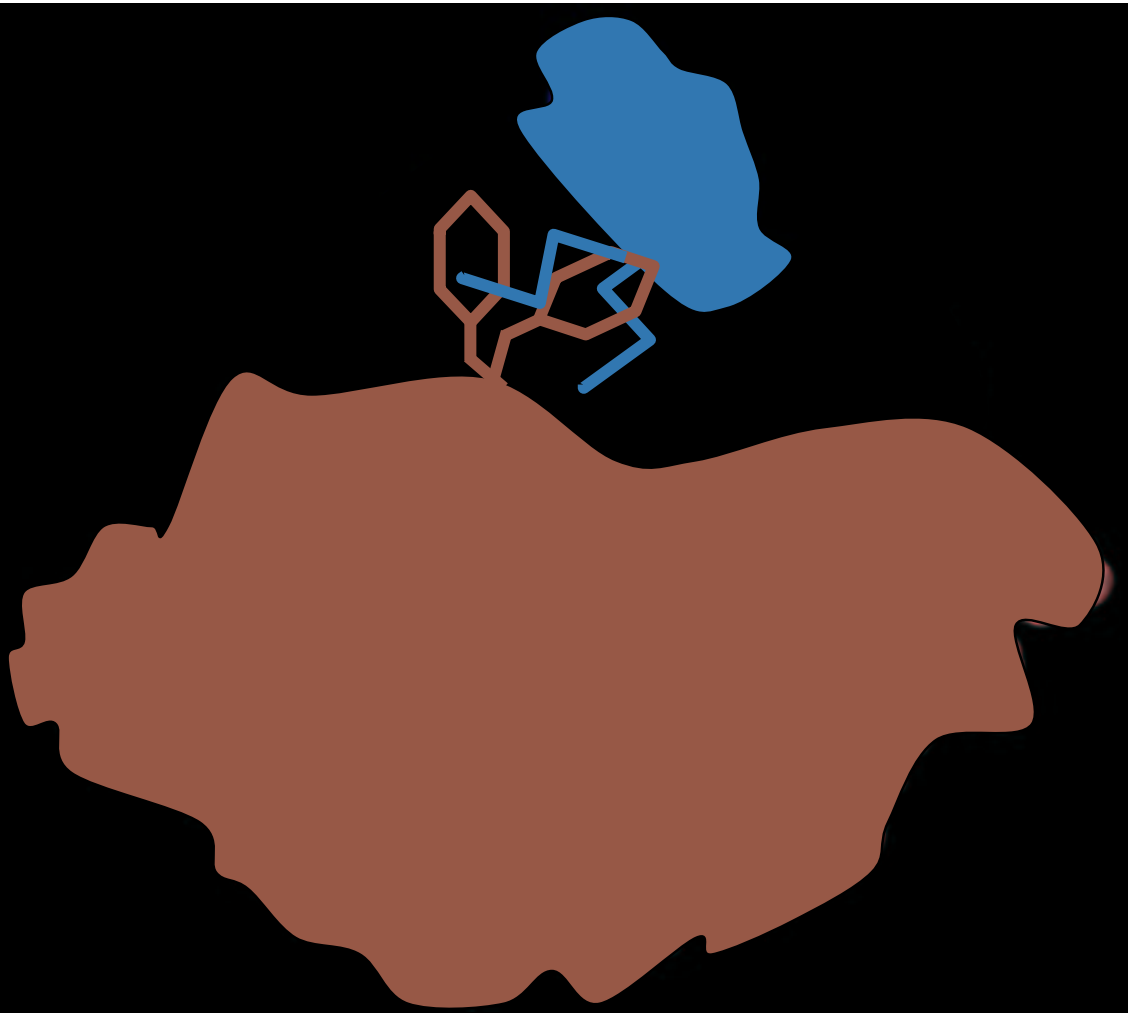
X_{11} X_{12}

S1 R1 R2

S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{ 0, 1\}$



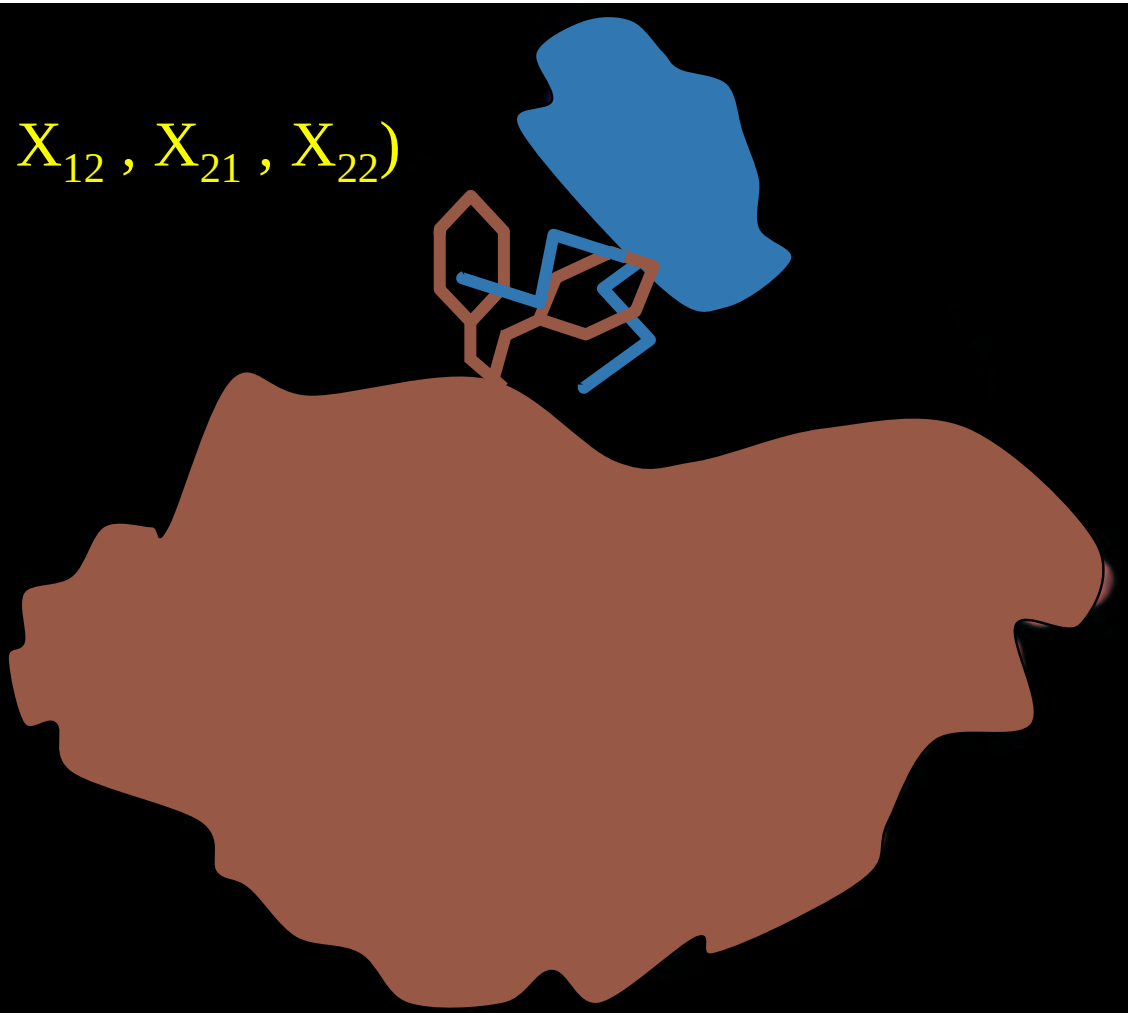
X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

S1 R1 R2

S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{0, 1\}$



X_{11} X_{12}

S1 R1 R2

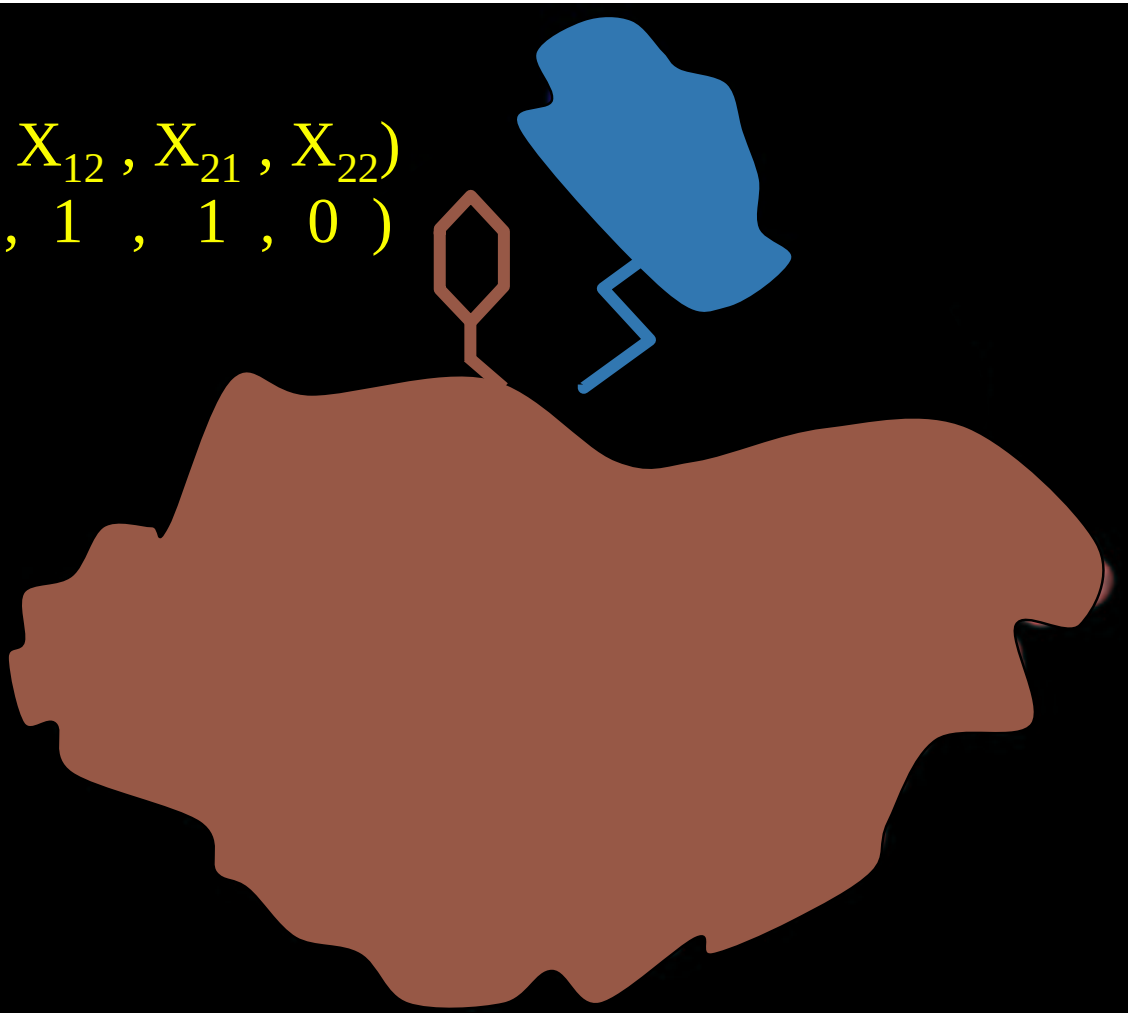
S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{0, 1\}$

$(X_{11}, X_{12}, X_{21}, X_{22})$

(0 , 1 , 1 , 0)



X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

$(\ 0\ ,\ 1\ ,\ 1\ ,\ 0\ )$

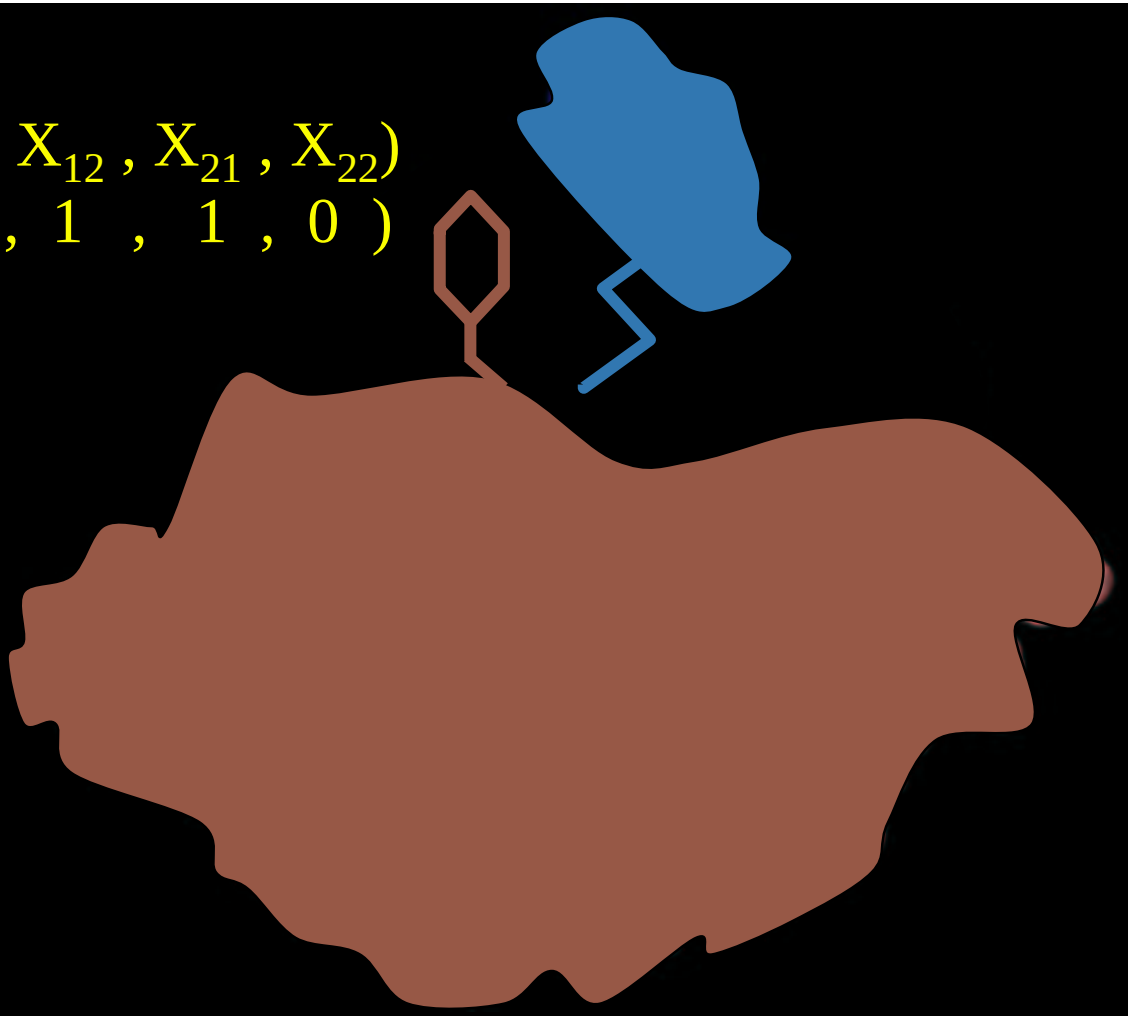
S1 R1 R2

S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{0, 1\}$

$\sum_{j} X_{ij} = 1 \quad \forall\, i$



X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

$(\ 0\ ,\ 1\ ,\ 1\ ,\ 0\ )$

S1 R1 R2

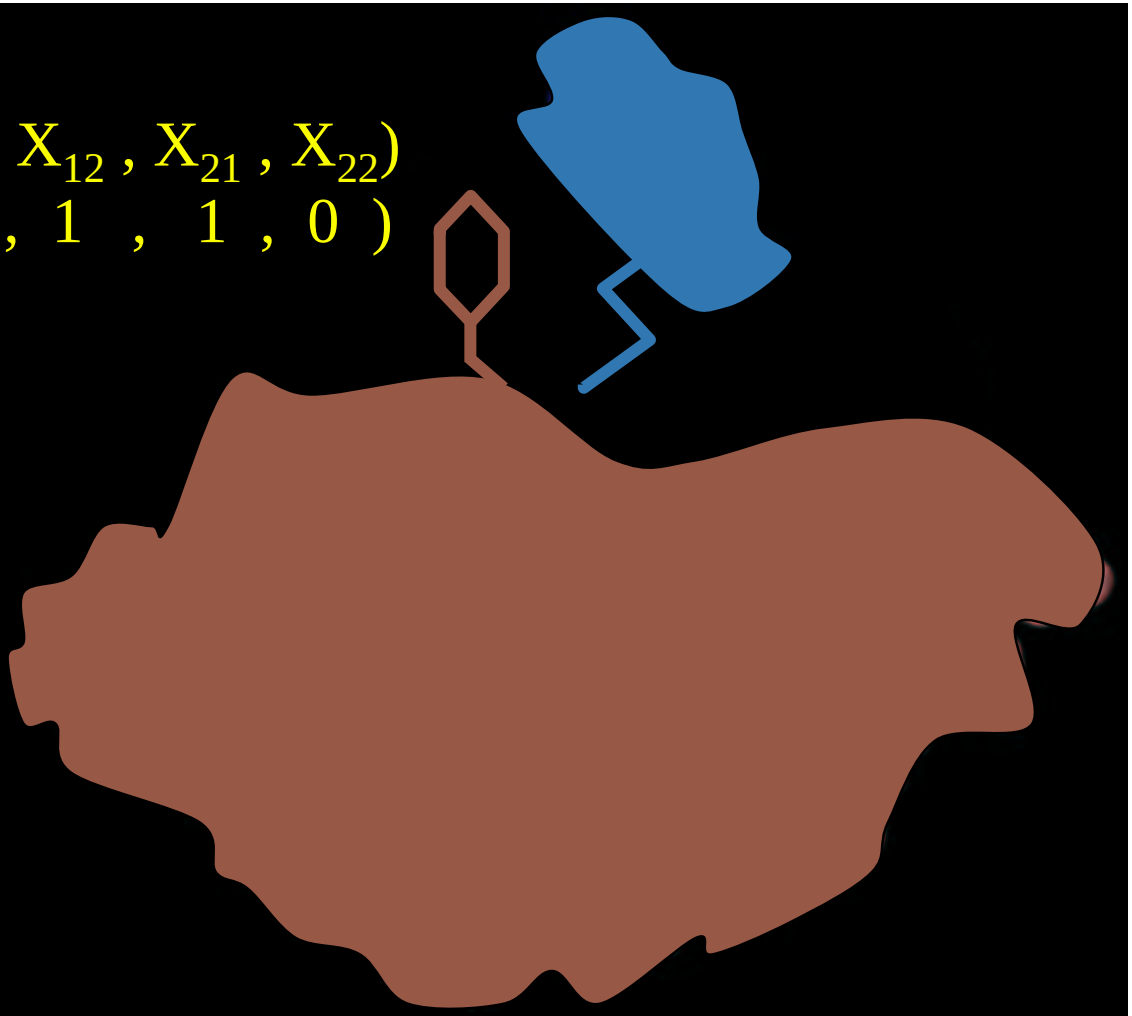
S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{0, 1\}$

$\sum_{j} X_{ij} = 1 \quad \forall\, i$

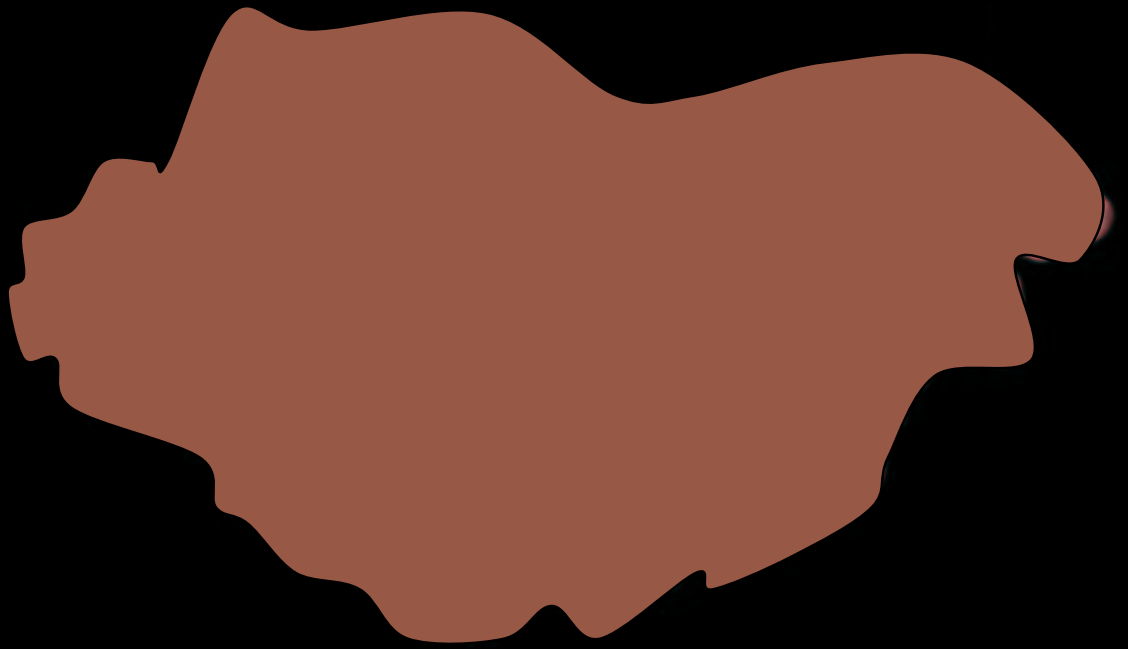
min



X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

$(\ 0\ ,\ 1\ ,\ 1\ ,\ 0\ )$

S1 R1 R2



S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{0, 1\}$

$\sum_{j} X_{ij} = 1 \quad \forall\ i$

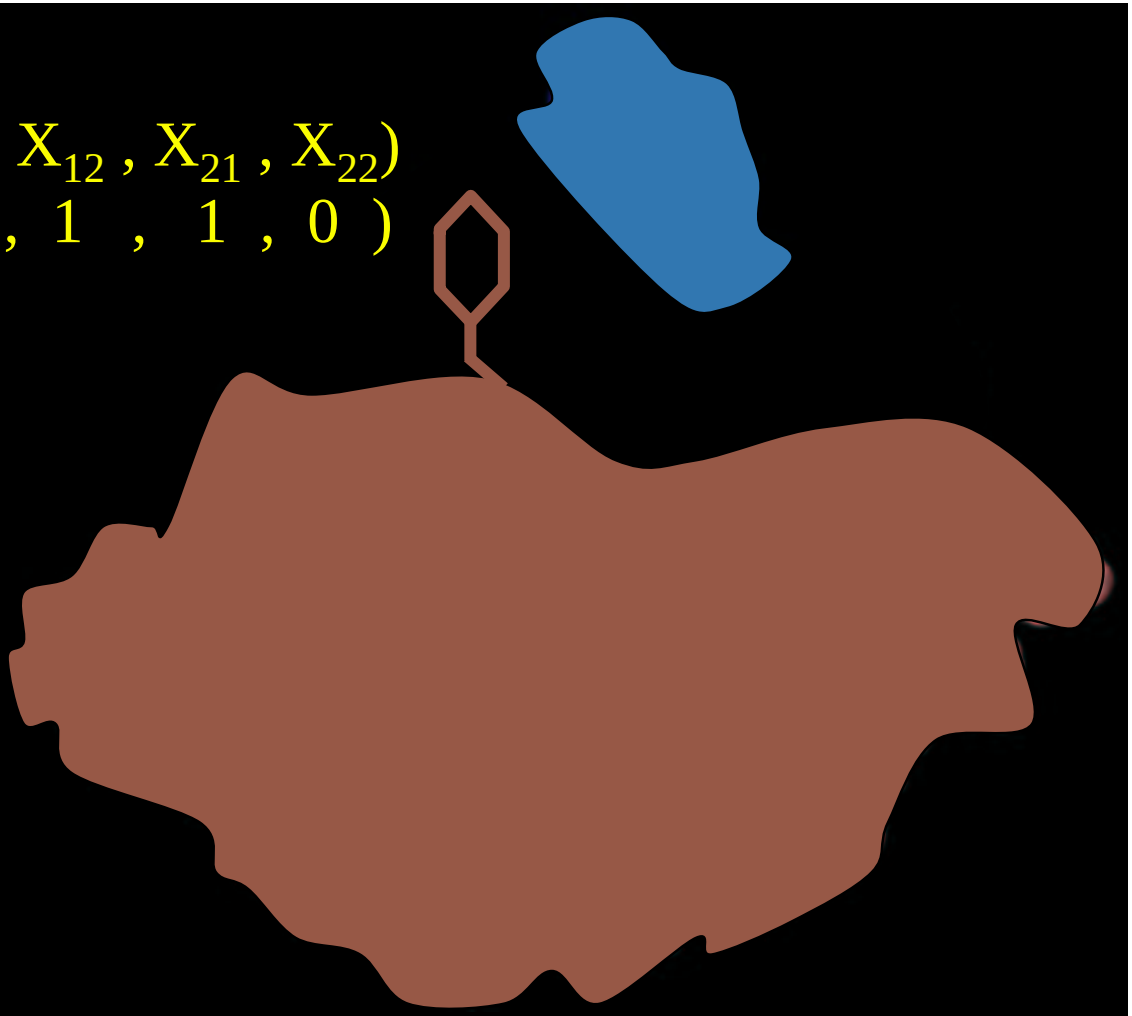
$\min \quad E_T$

X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

S1 R1 R2 (0 , 1 , 1 , 0)

S2 R1 R2

X_{21} X_{22}



$X_{ij} \in \{0, 1\}$

$\sum_{j} X_{ij} = 1 \quad \forall i$

$\min \quad E_T + \sum_{ij} X_{ij} E_{ij}$

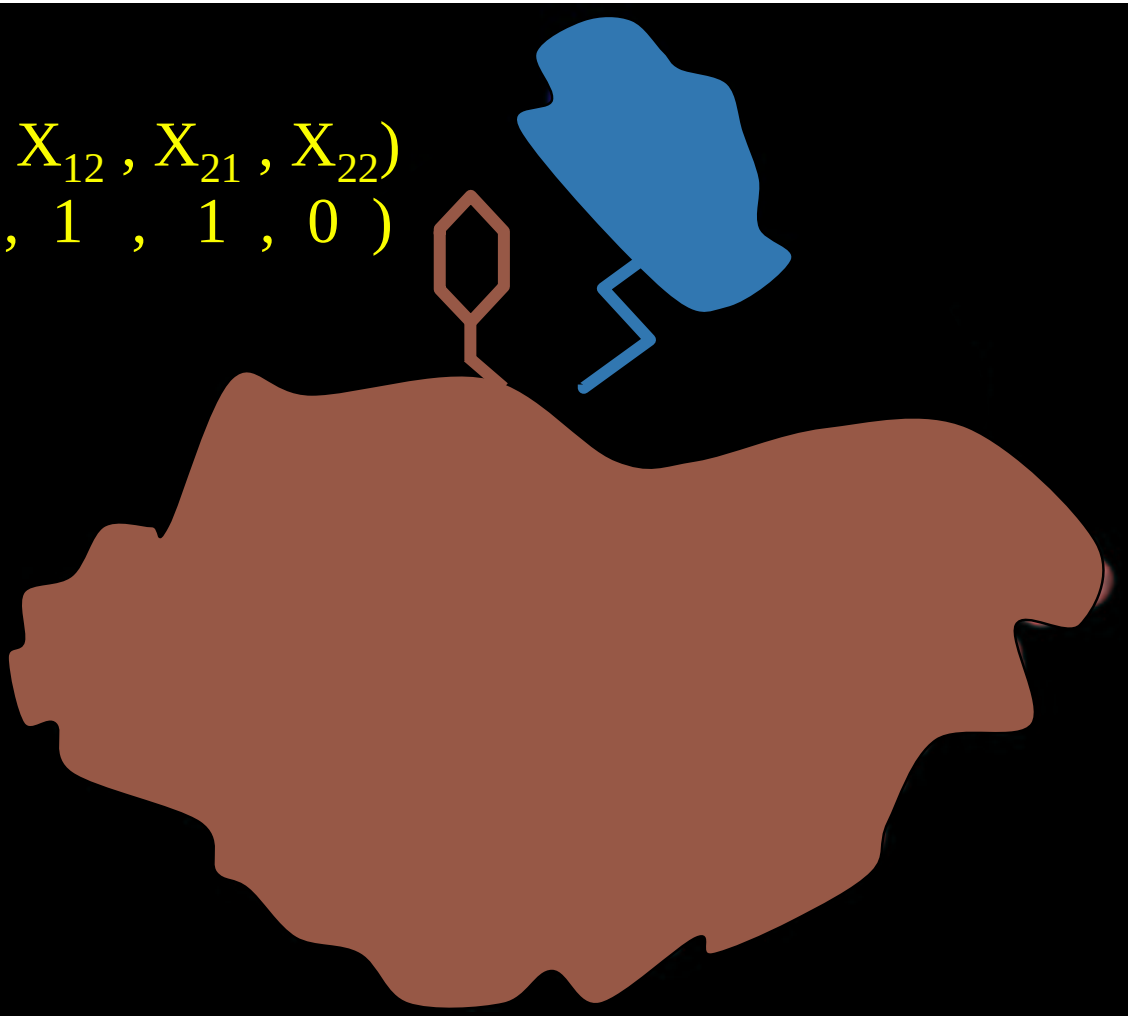
X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

$(\ 0\ ,\ 1\ ,\ 1\ ,\ 0\ )$

S1 R1 R2

S2 R1 R2

X_{21} X_{22}



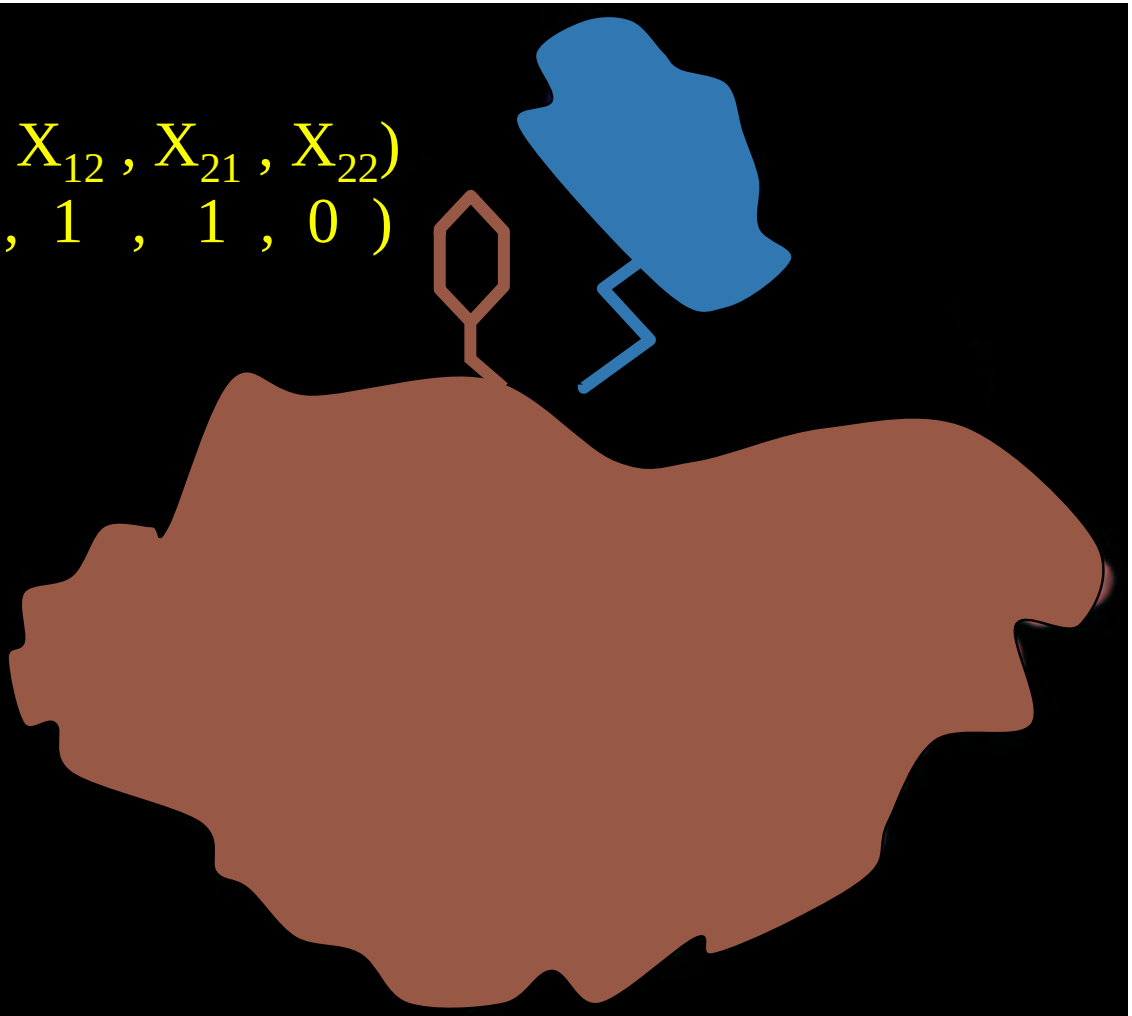
$$X_{ij} \in \{0, 1\}$$

$$\sum_{j} X_{ij} = 1 \quad \forall\, i$$

$$\min \quad E_T + \sum_{ij} X_{ij}\, E_{ij} + \sum_{ij} \sum_{\substack{kl \\ k \neq i}} X_{ij}\, X_{kl}\, E_{ijkl}$$

X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

S1 R1 R2 (0 , 1 , 1 , 0)



S2 R1 R2

X_{21} X_{22}

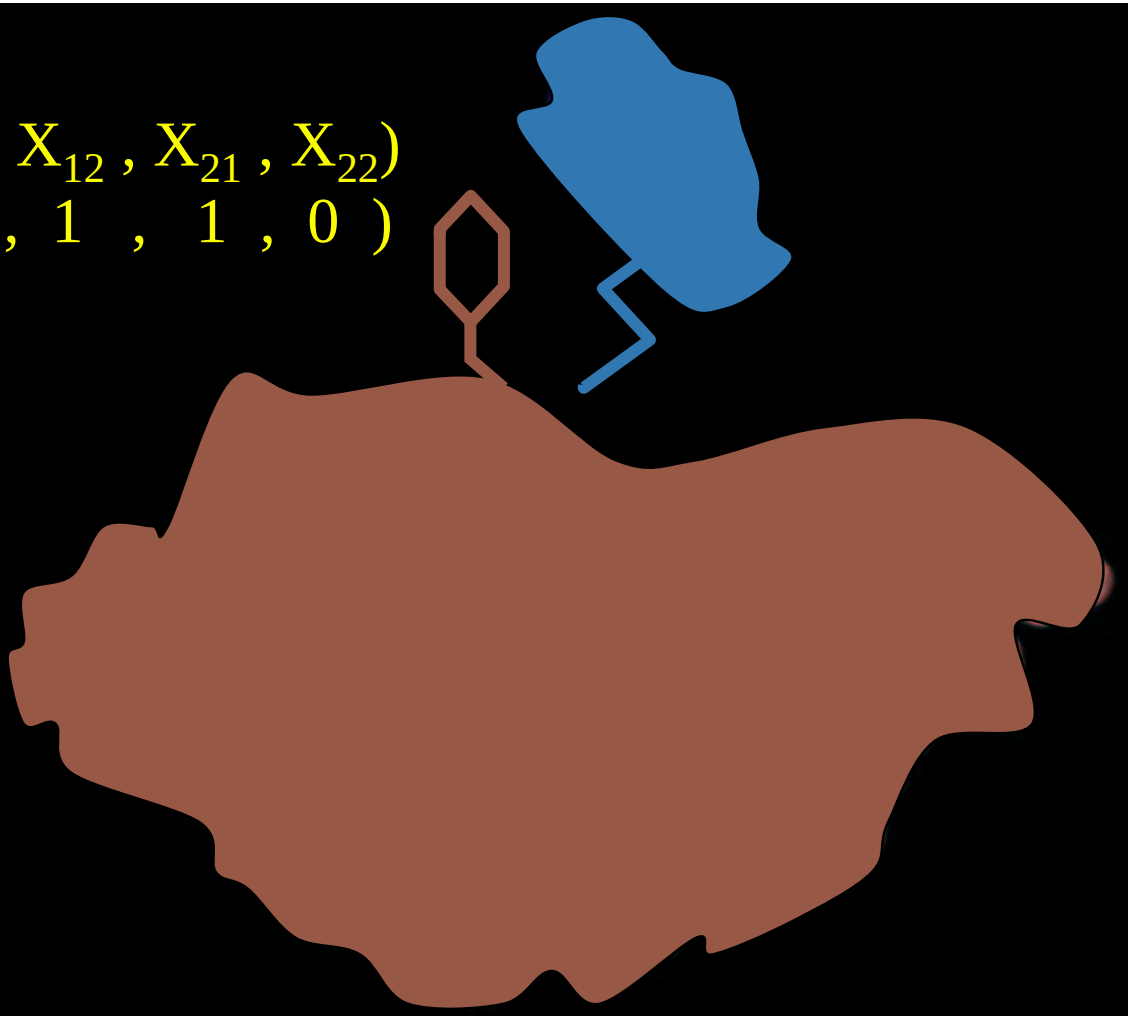
$X_{ij} \in \{0, 1\}$

$$\sum_{j} X_{ij} = 1 \quad \forall\, i$$

$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{\substack{kl \\ k \neq i}} X_{ij} X_{kl} (E_{ijkl} - E_{max})$$

X_{11} X_{12} $(X_{11}, X_{12}, X_{21}, X_{22})$

S1 R1 R2 (0 , 1 , 1 , 0)



S2 R1 R2

X_{21} X_{22}

$X_{ij} \in \{ 0, 1\}$

$$\sum_{j} X_{ij} = 1 \quad \forall i$$

$Y_{ijkl} \leq X_{ij}$, $Y_{ijkl} \leq X_{kl}$ $Y_{ijkl} \in \{ 0, 1\}$

$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{\substack{kl \\ k \neq i}} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \ , \ Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in \{ 0, 1\}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{ 0, 1\}$$

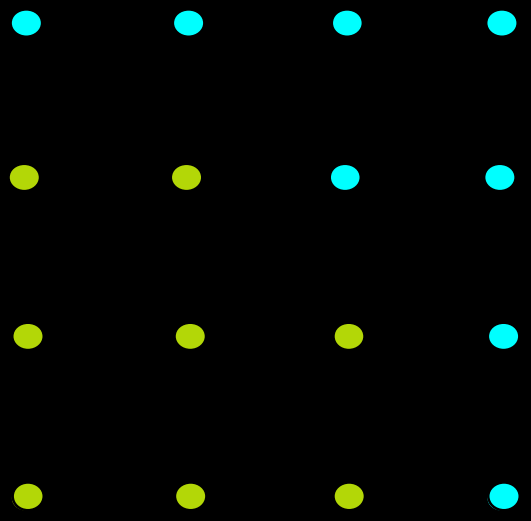
Solve ILP

$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \ , \ Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in \{ 0, 1\}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{ 0, 1\}$$

Solve ILP

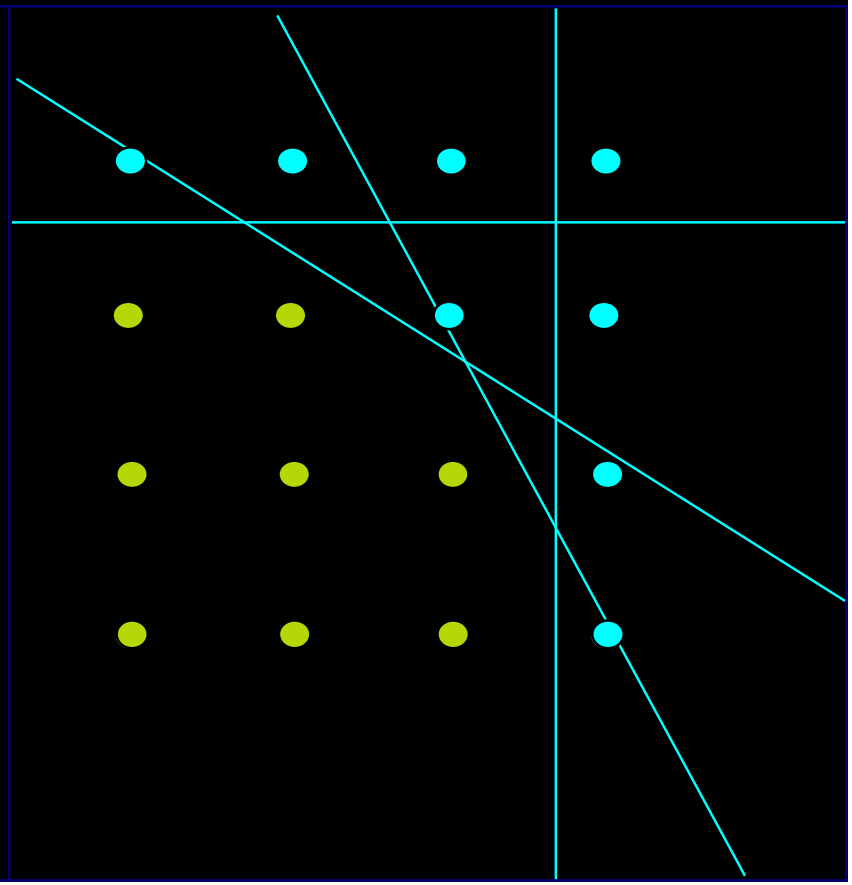


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \ , \ Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in \{ 0, 1 \}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{ 0, 1 \}$$

Solve ILP

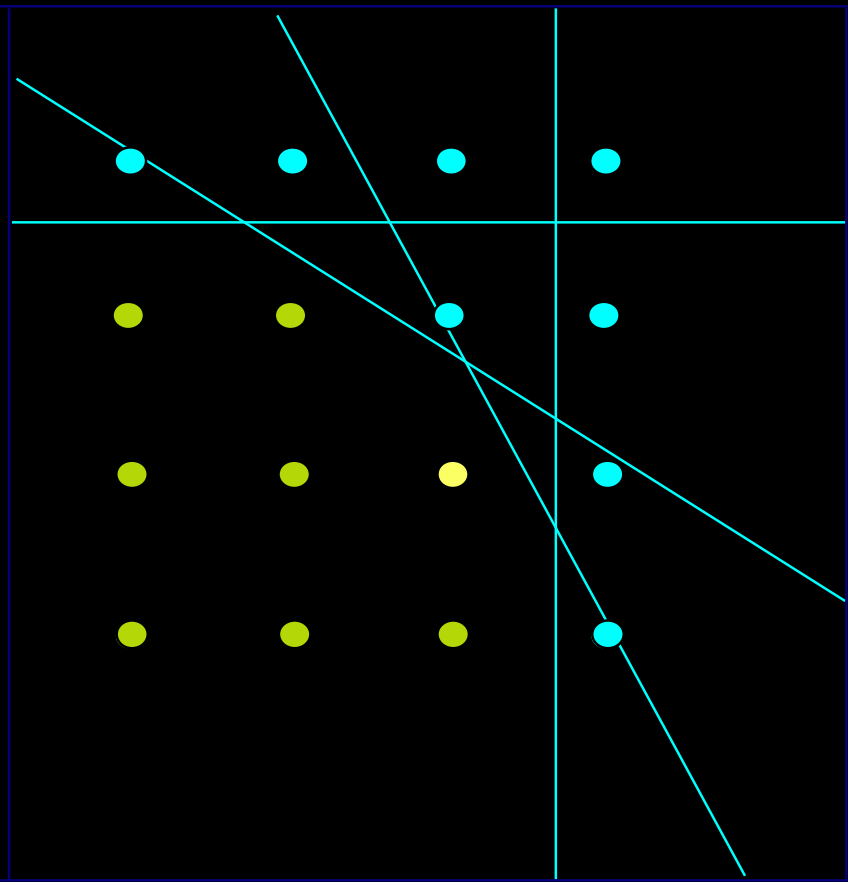


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \ , \ Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in \{ 0, 1 \}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{ 0, 1 \}$$

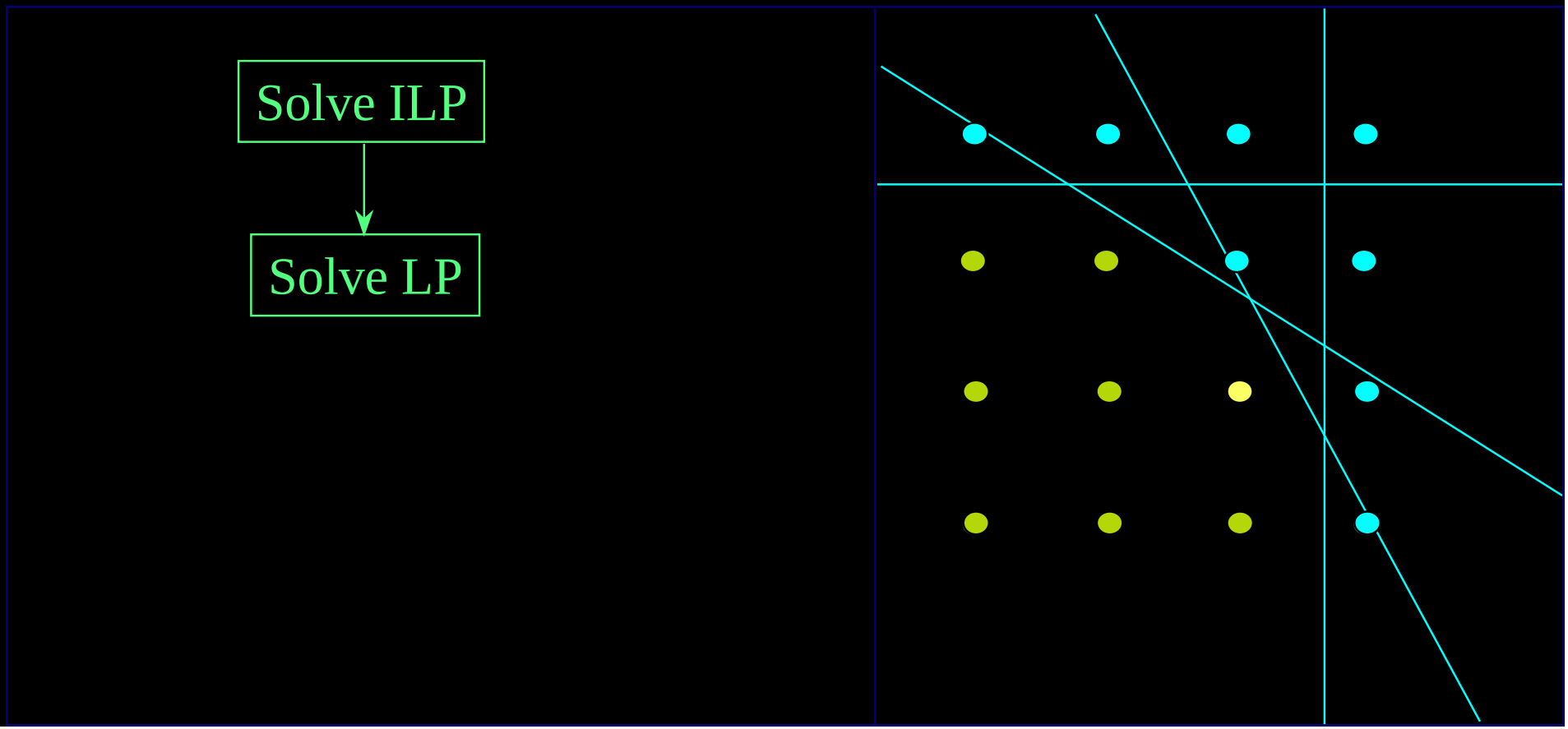
Solve ILP



$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \ , \ Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in \{ 0, 1 \}$$

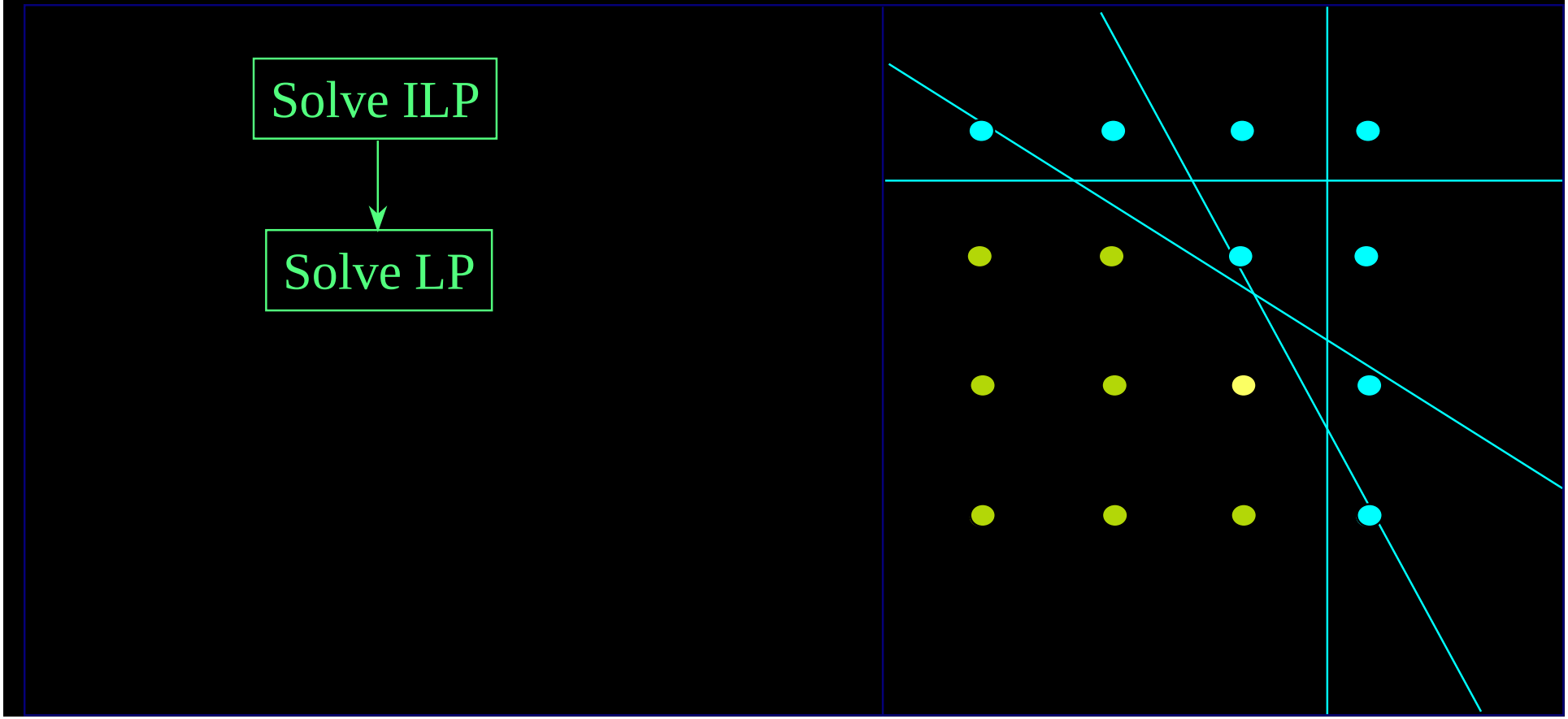
$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{ 0, 1 \}$$



$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \;,\; Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

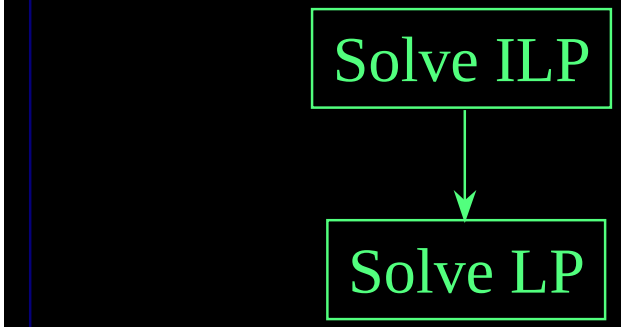
$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

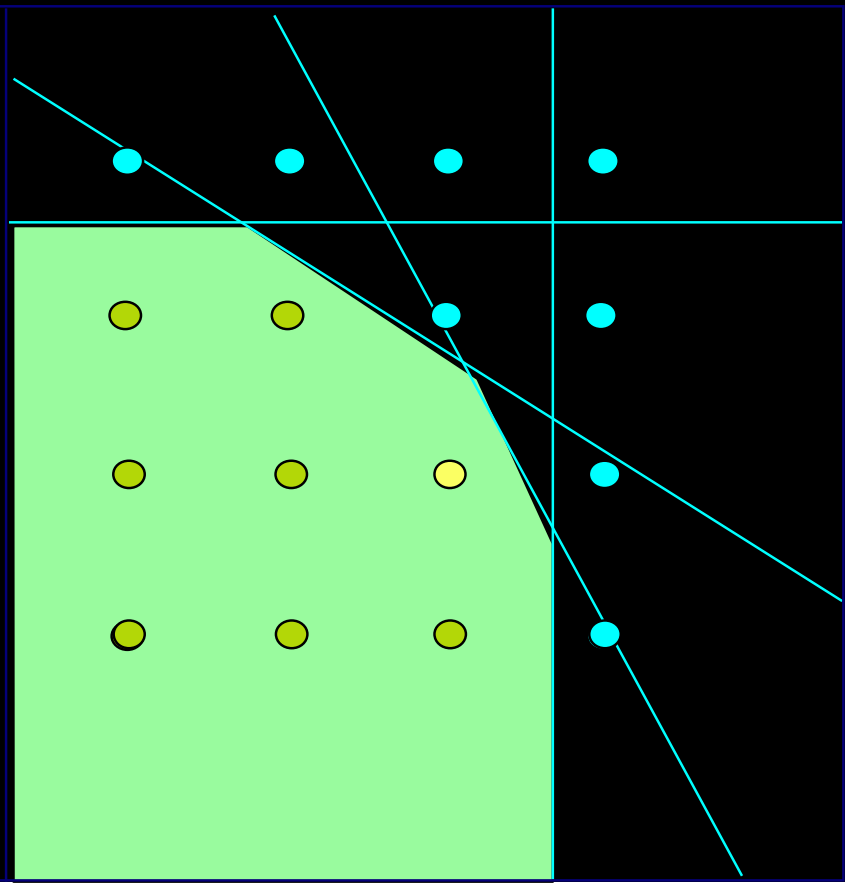


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \;,\; Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

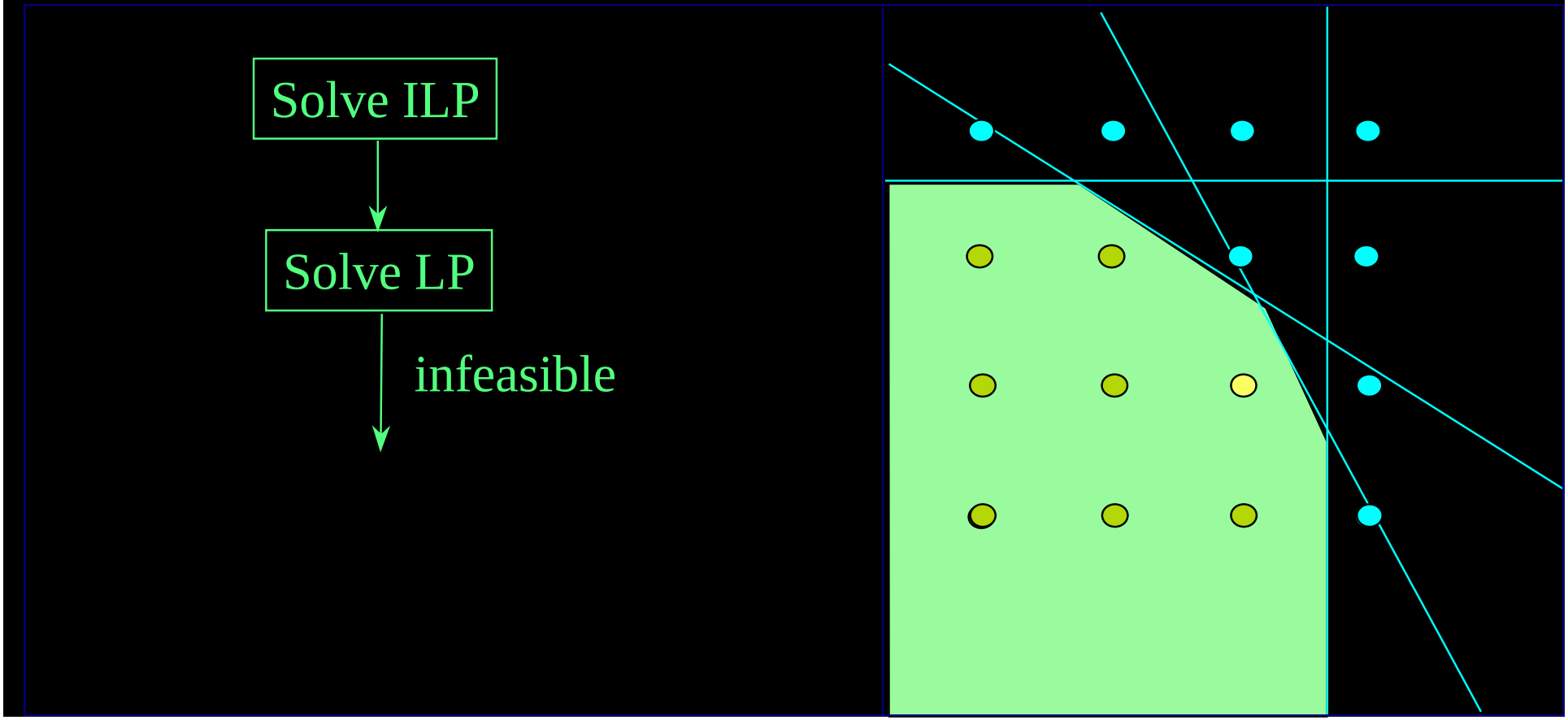




$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \ , \ Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

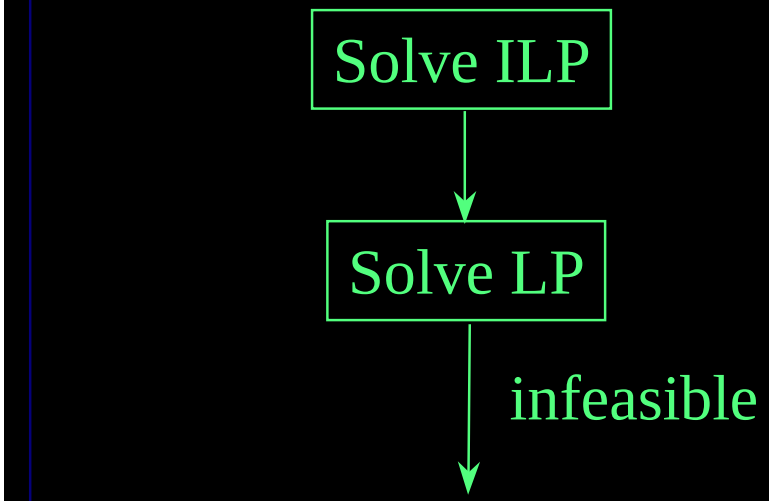
$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

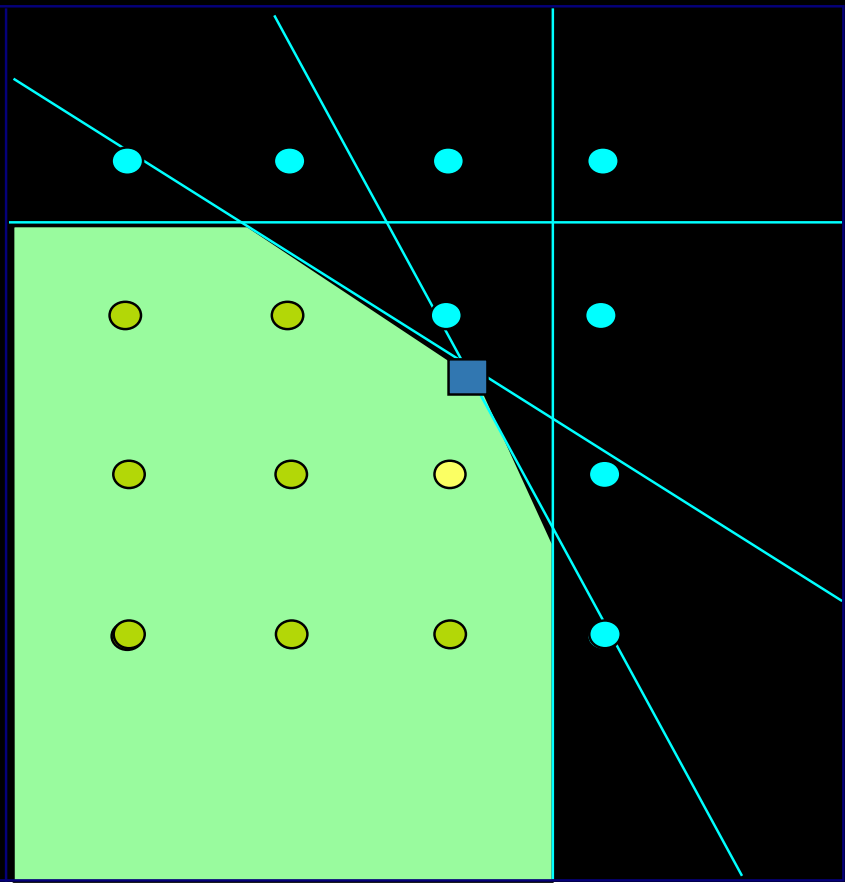


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \ , \ Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

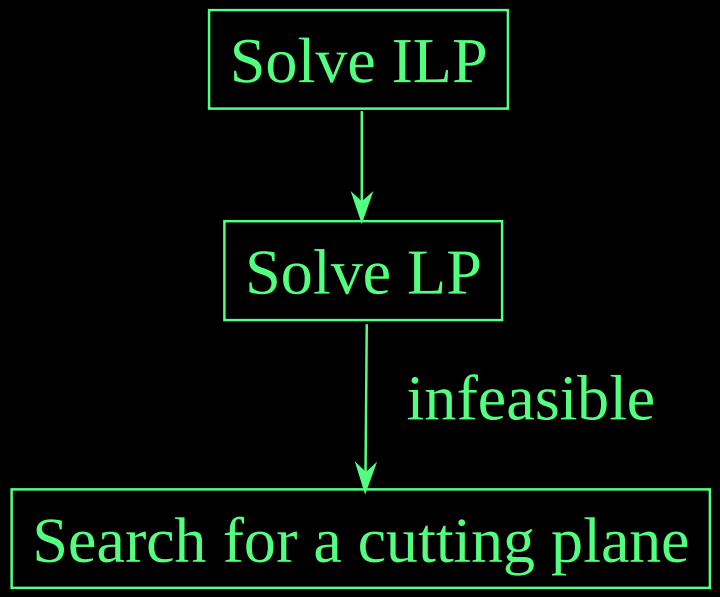


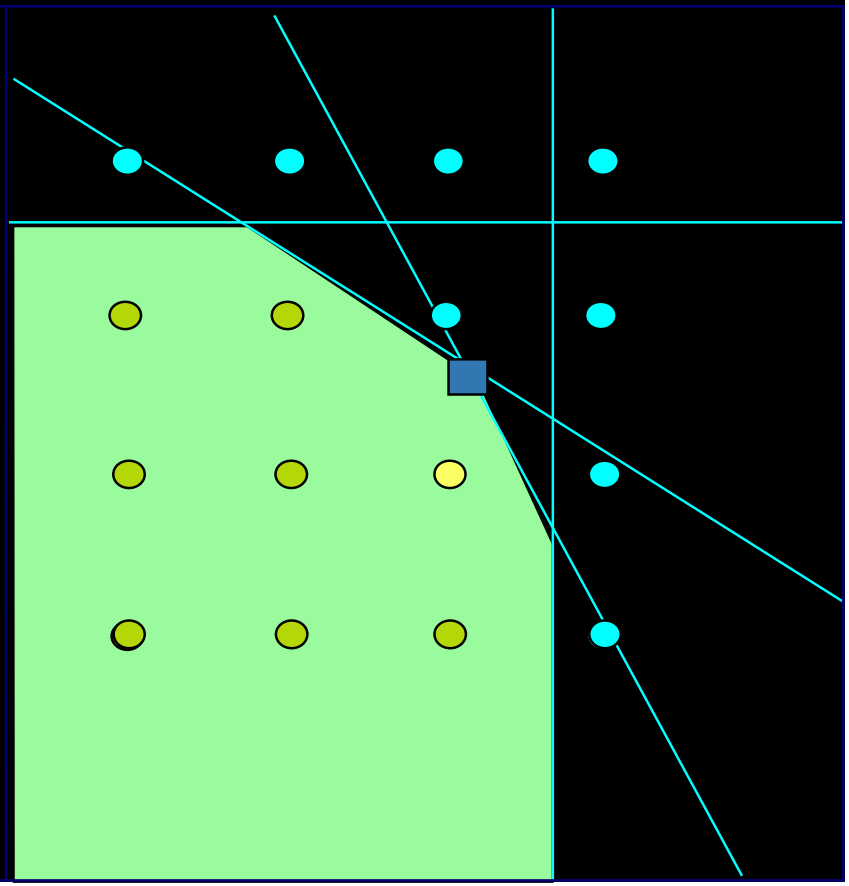


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \ , \ Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

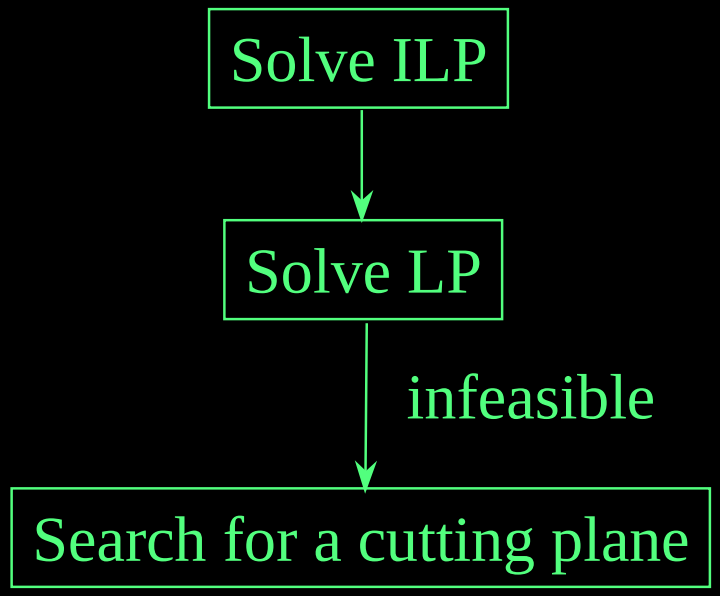


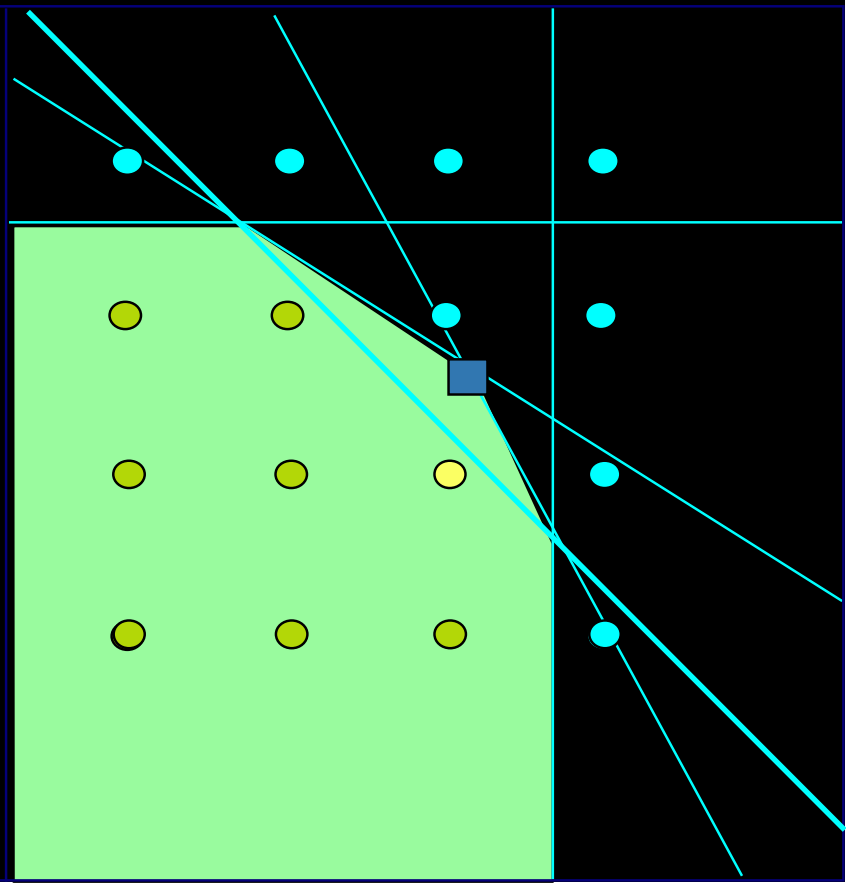


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \ , \ Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

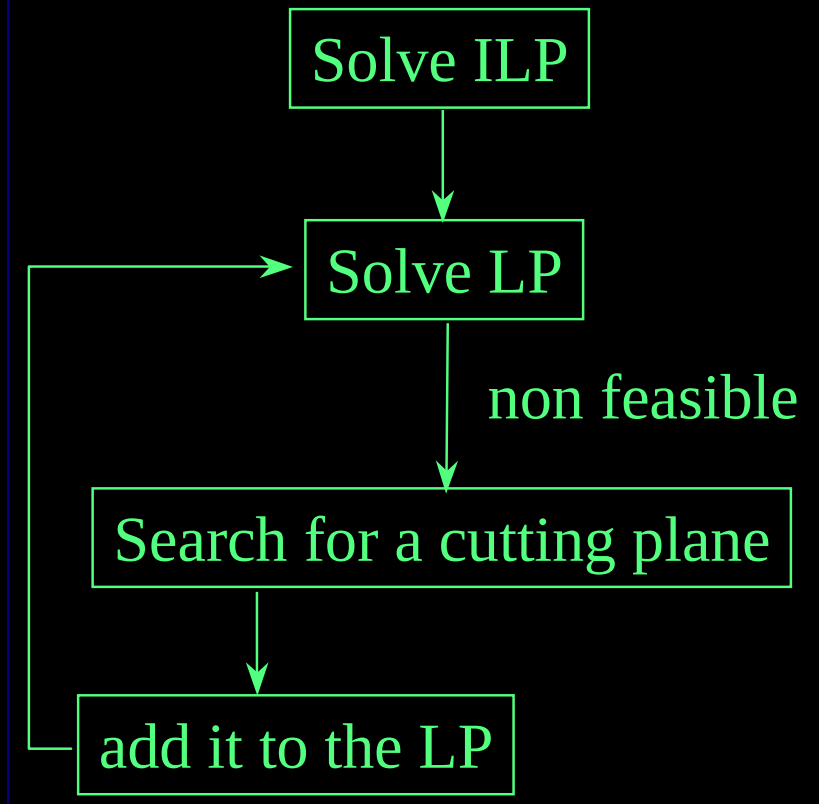


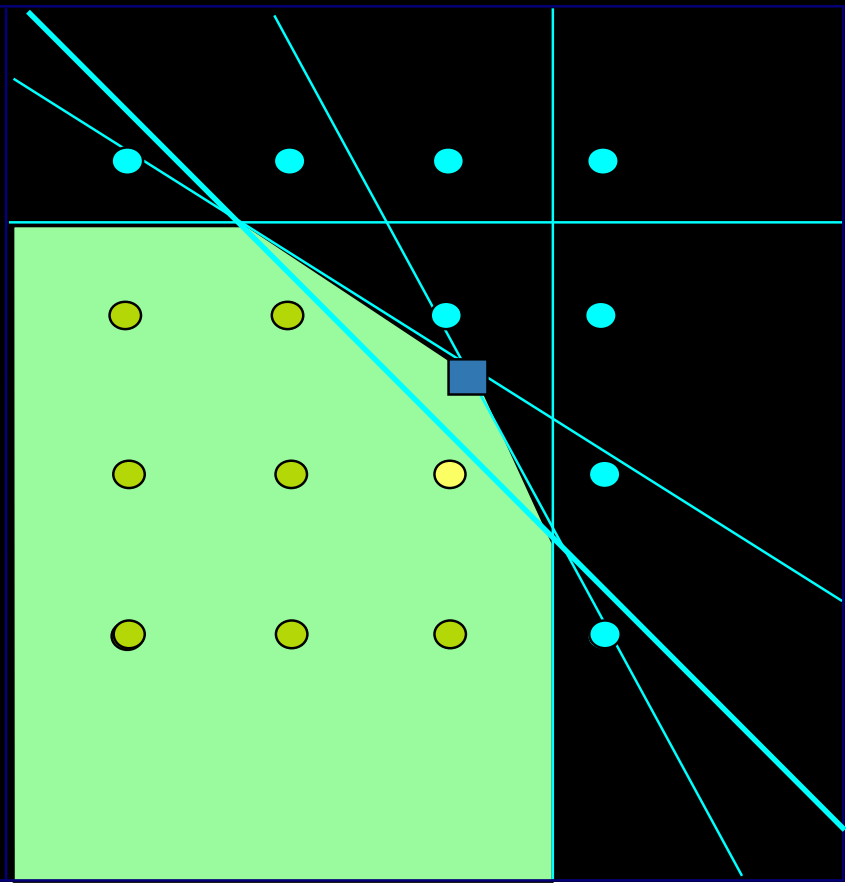


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \leq X_{ij} \;,\; Y_{ijkl} \leq X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

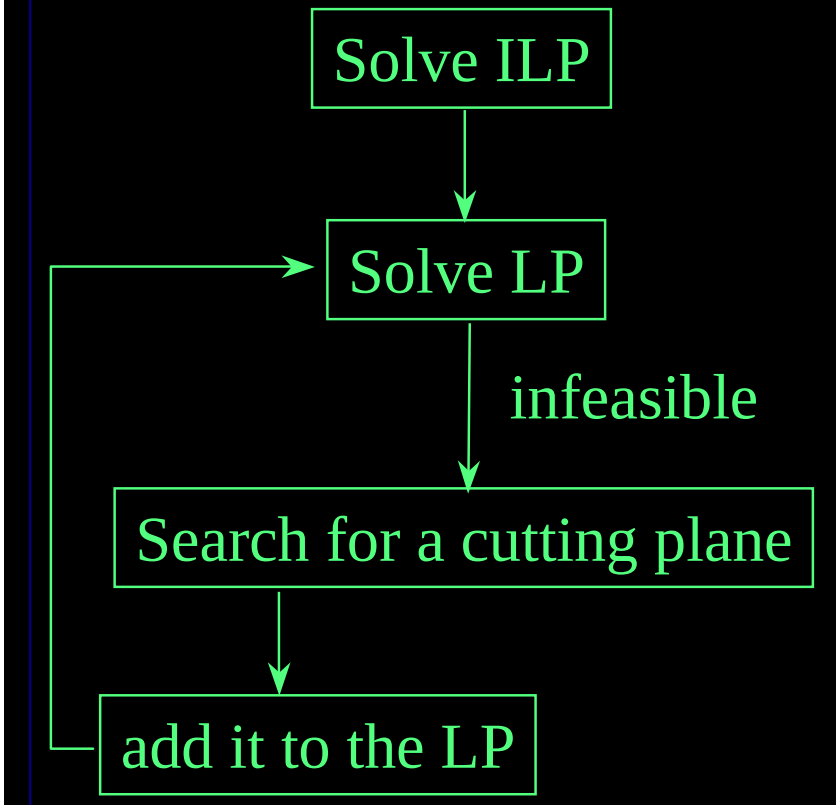


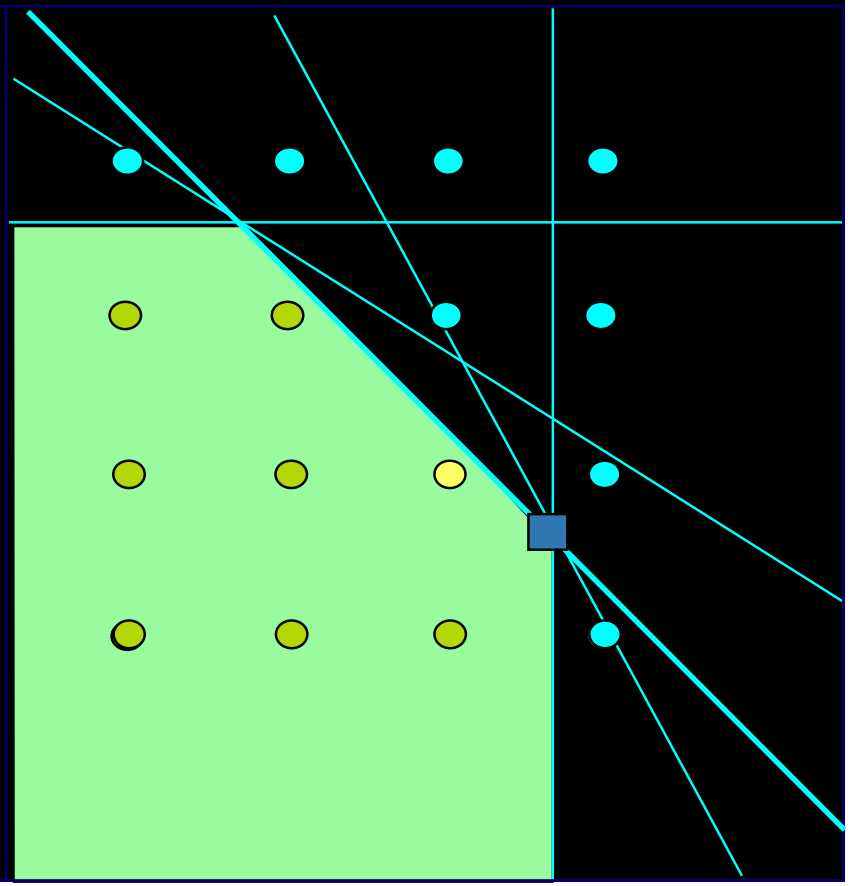


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \;,\; Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

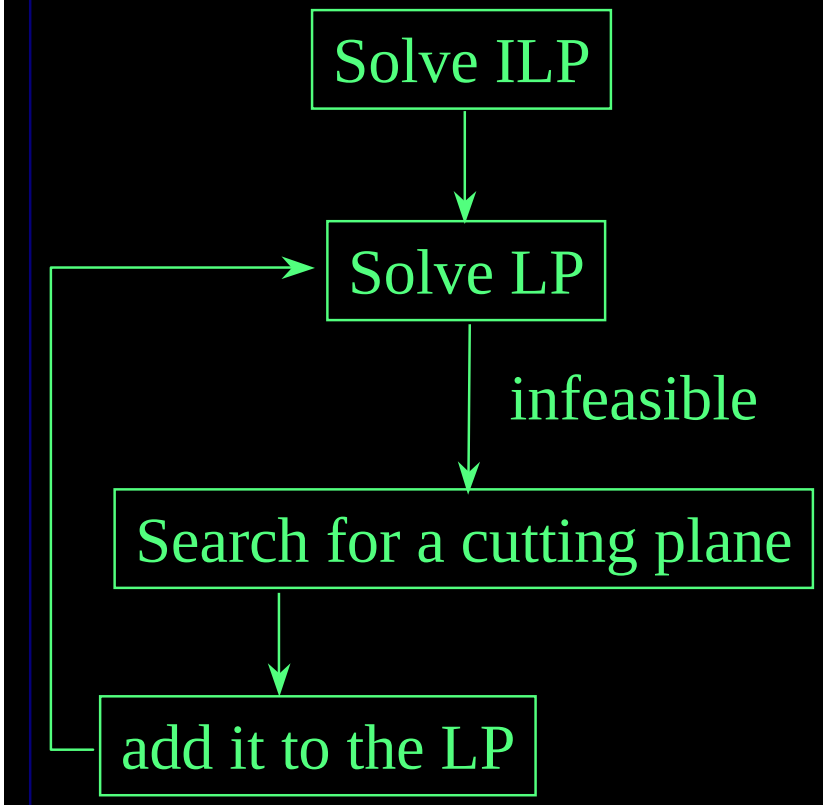


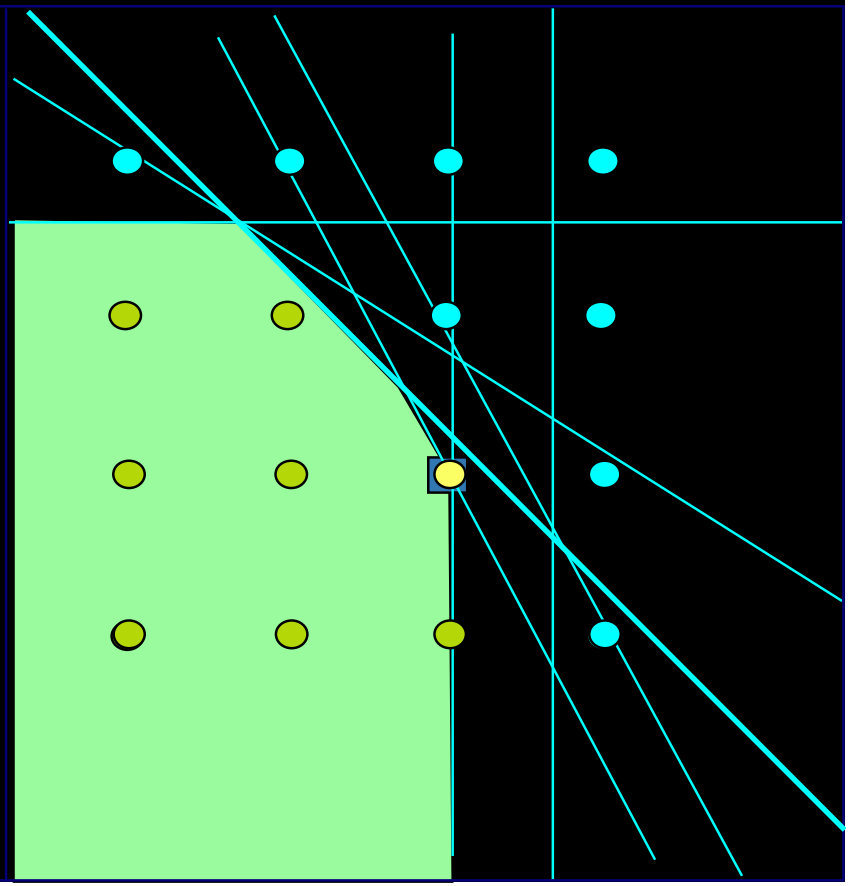


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \;,\; Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

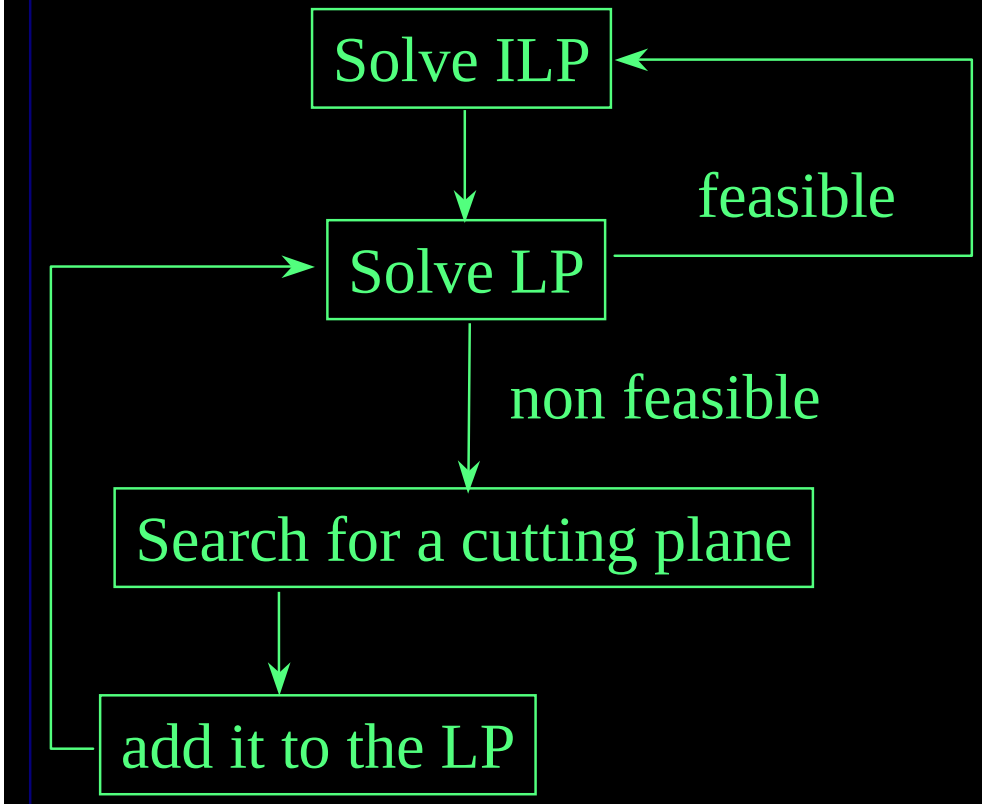


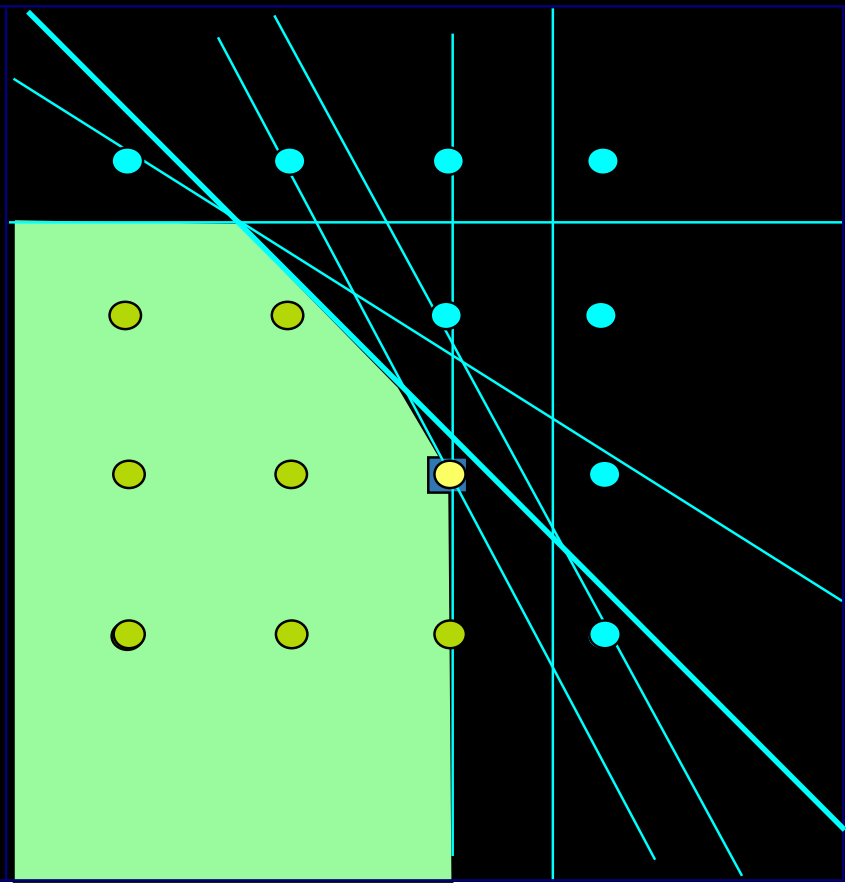


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \ , \ Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in [0, 1]$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in [0, 1]$$

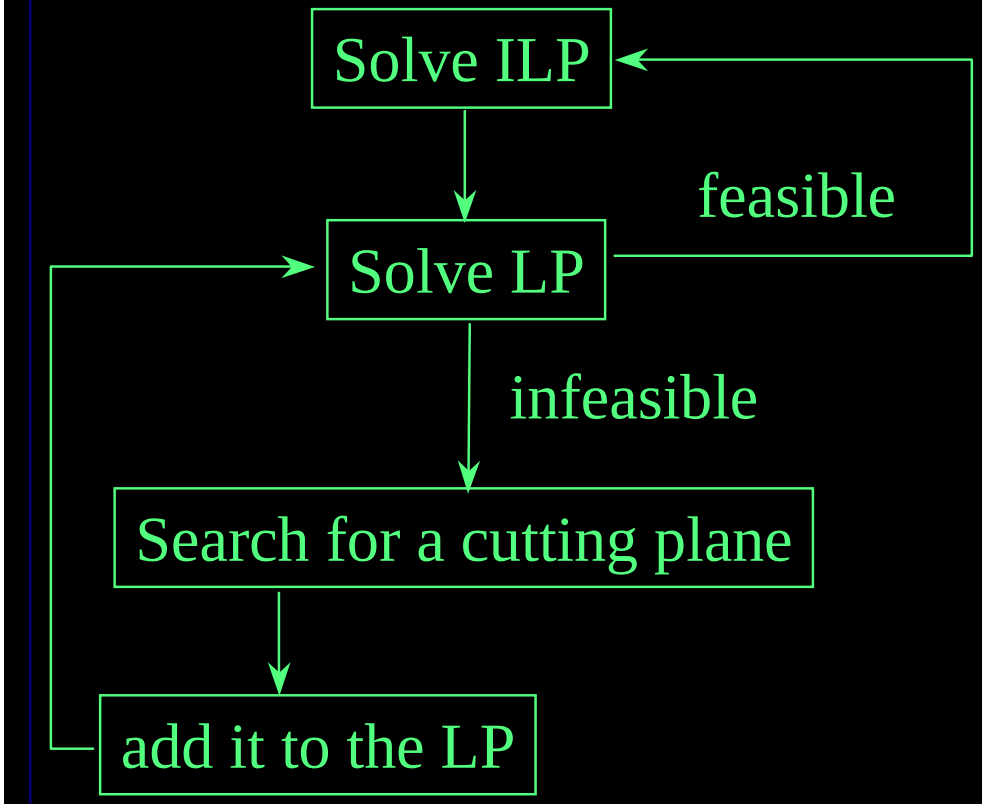


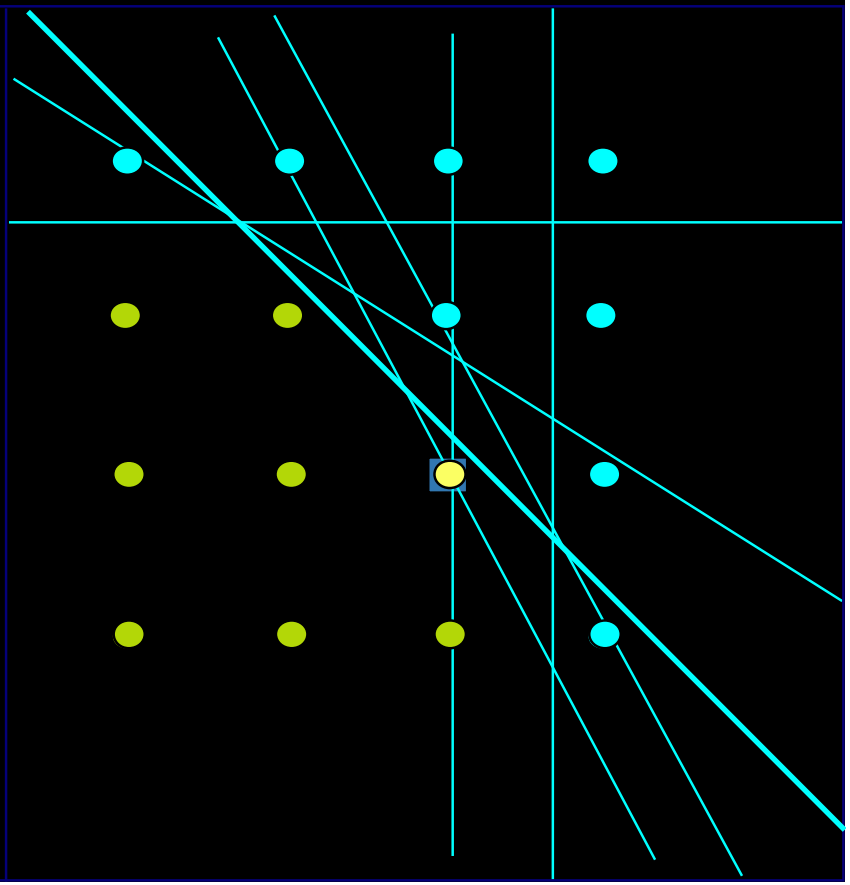


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \;,\; Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in \{0, 1\}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{0, 1\}$$

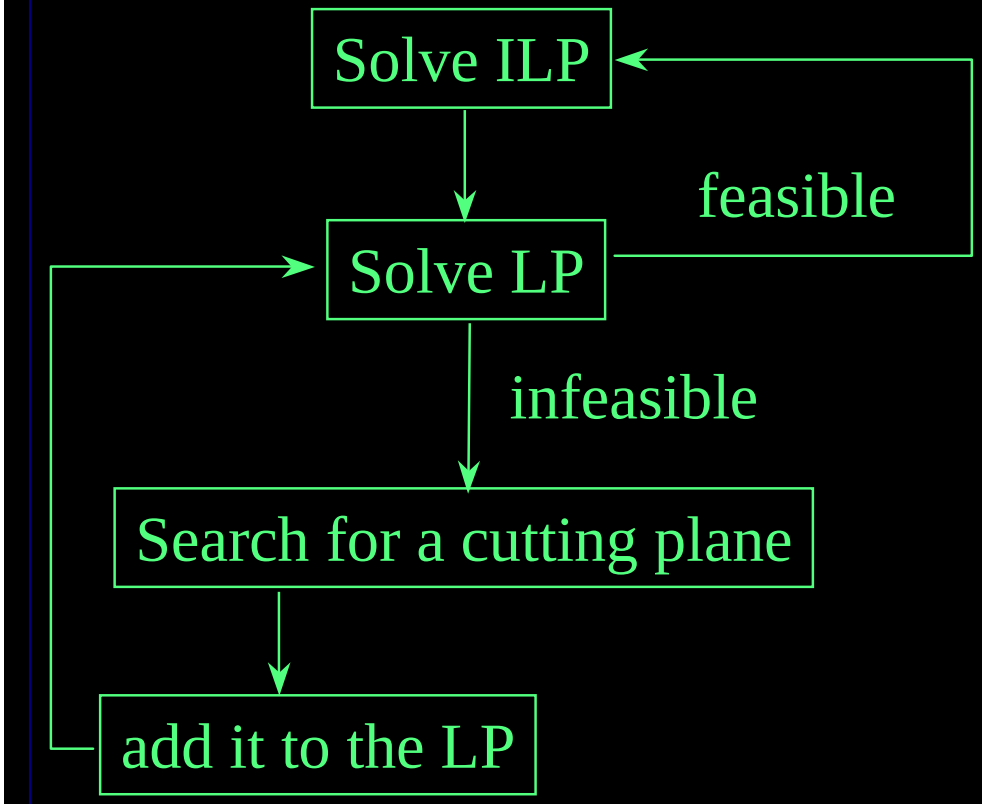


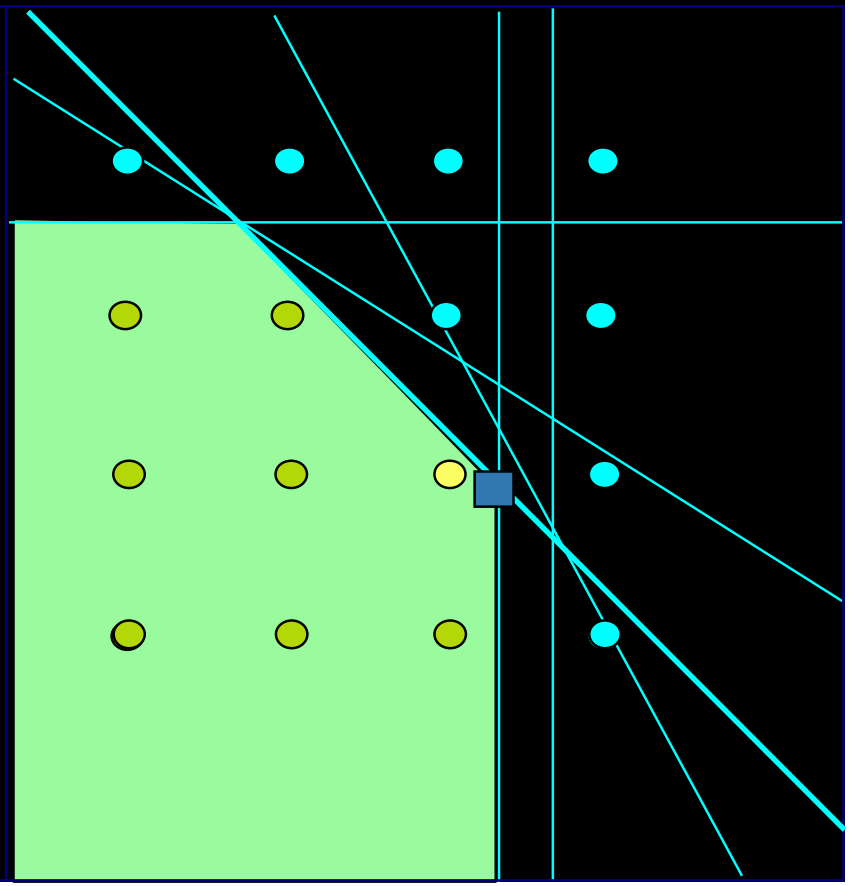


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \ , \ Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in \{ 0, 1 \}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{ 0, 1 \}$$

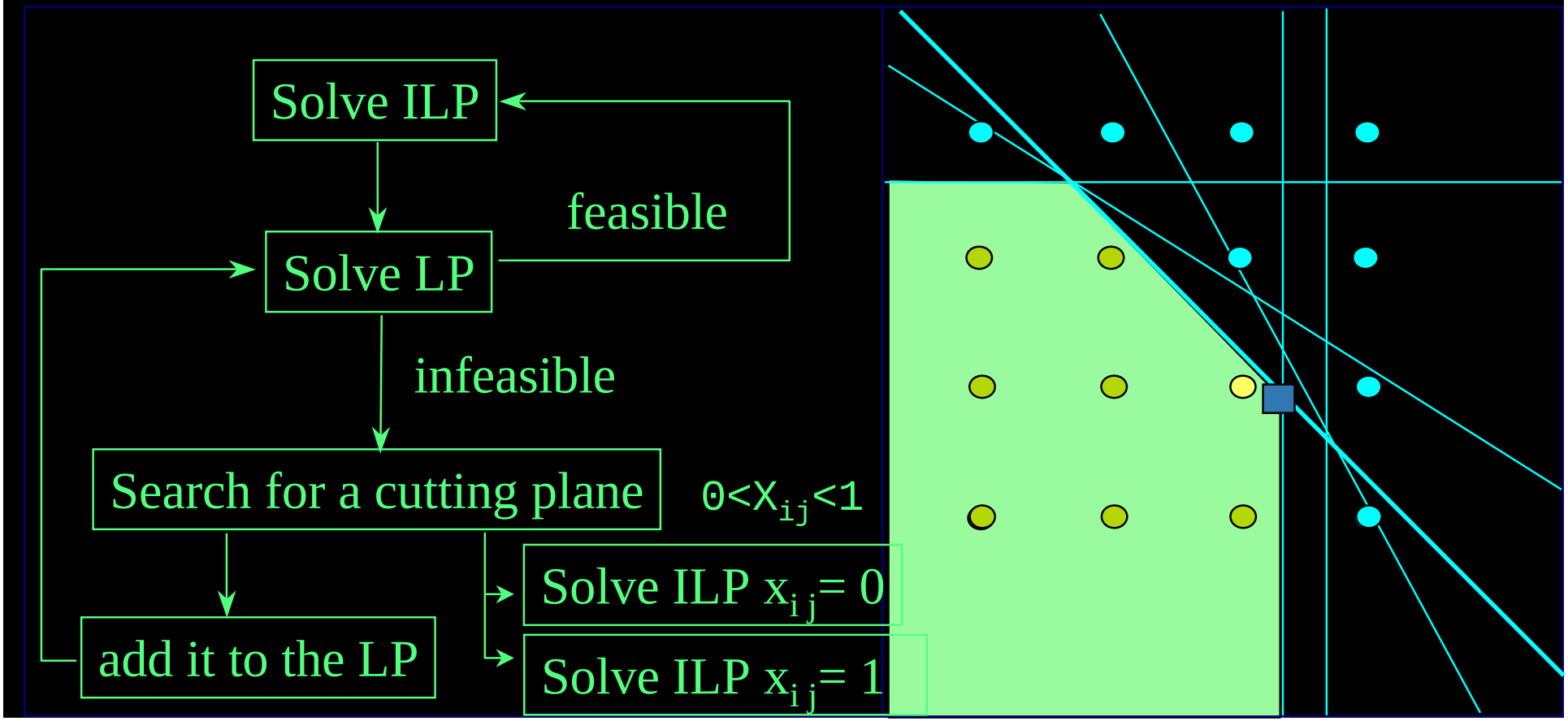




$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

$$Y_{ijkl} \le X_{ij} \;,\; Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in \{0, 1\}$$

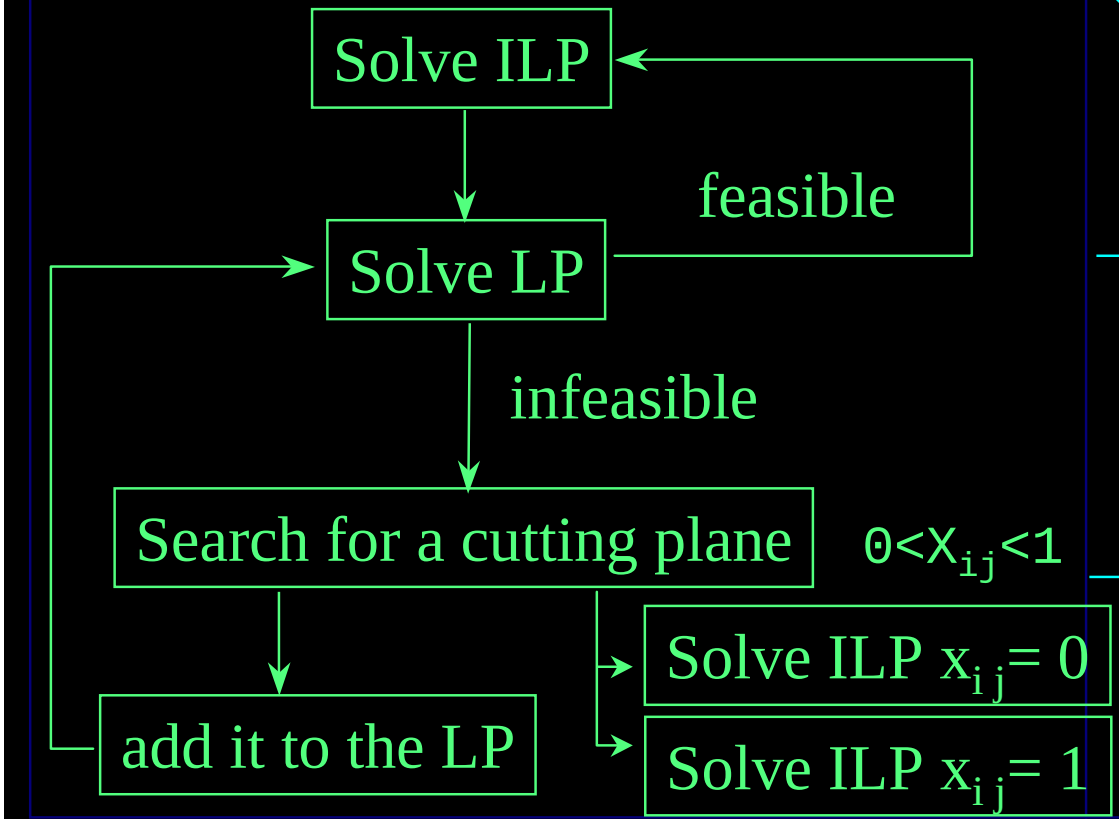
$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{0, 1\}$$

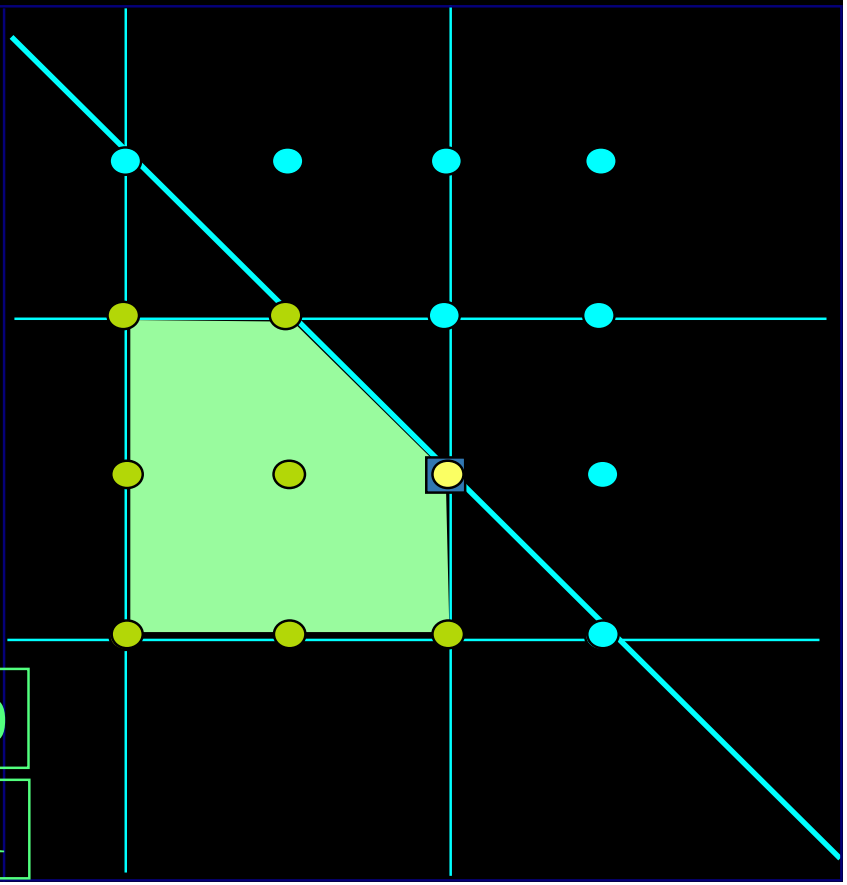


$$\min \quad E_T + \sum_{ij} X_{ij} E_{ij} + \sum_{ij} \sum_{kl} Y_{ijkl} (E_{ijkl} - E_{max})$$

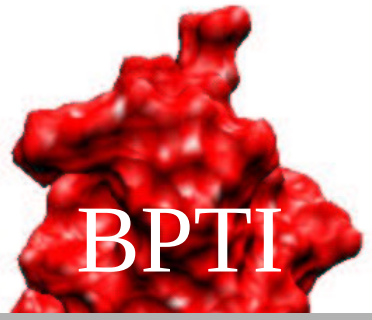
$$Y_{ijkl} \le X_{ij} \;,\; Y_{ijkl} \le X_{kl} \qquad Y_{ijkl} \in \{0, 1\}$$

$$\sum_{j} X_{ij} = 1 \quad \forall i \qquad X_{ij} \in \{0, 1\}$$

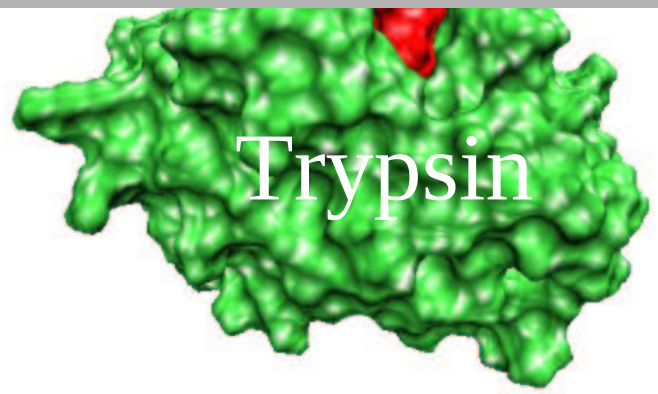




Test Case: Trypsin/BPTI

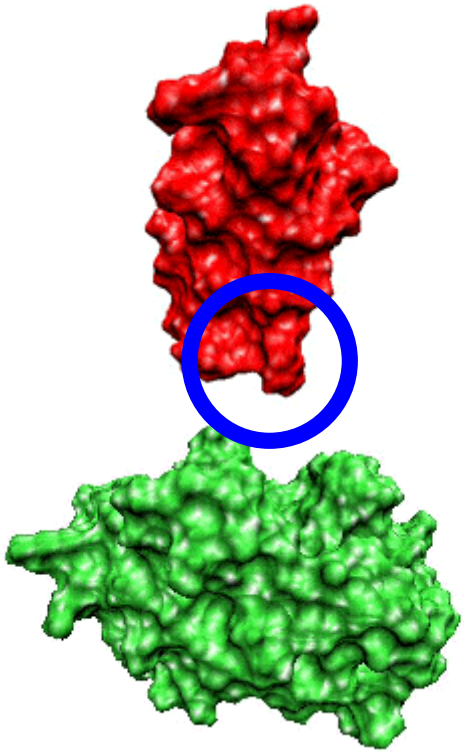


first approximation: rank 3

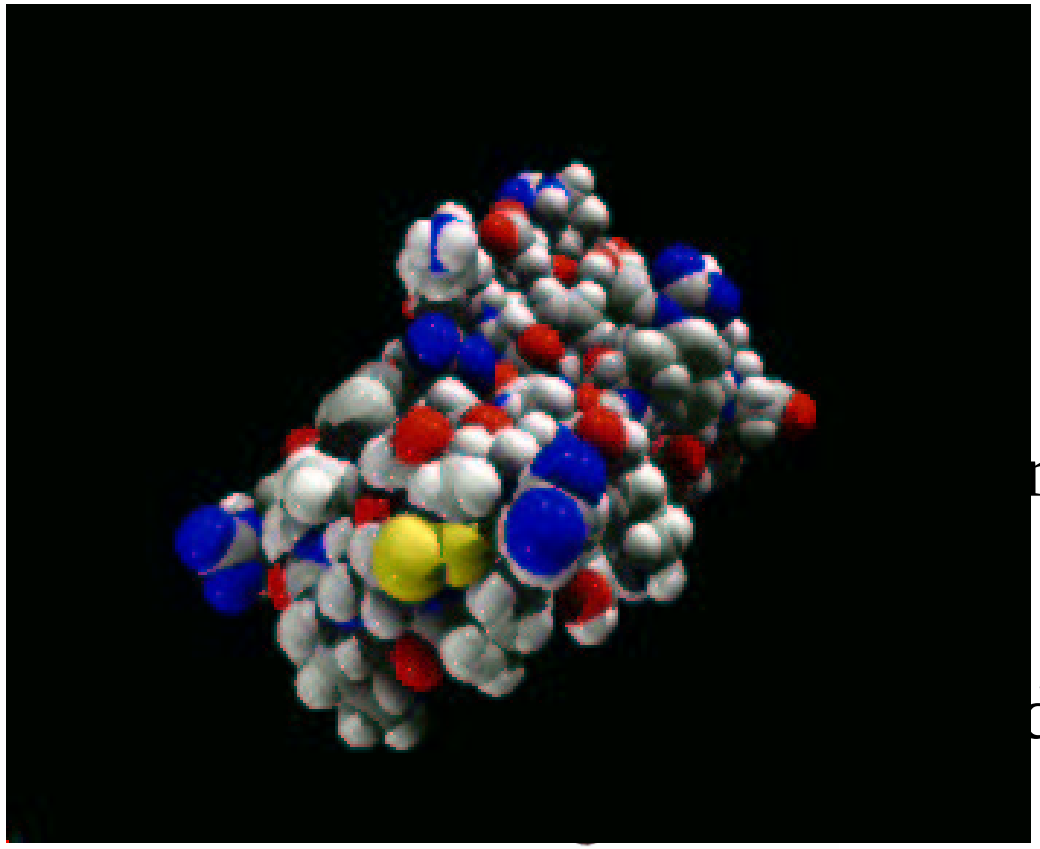


(docking of native structures)

What's the problem?



What's the problem?



15 moves on docking

rlapping atoms

d docking must fail

Semi-Flexible Protein Docking

- Structure generation
- Filtering
- Final energetic evaluation

Semi-Flexible Protein Docking

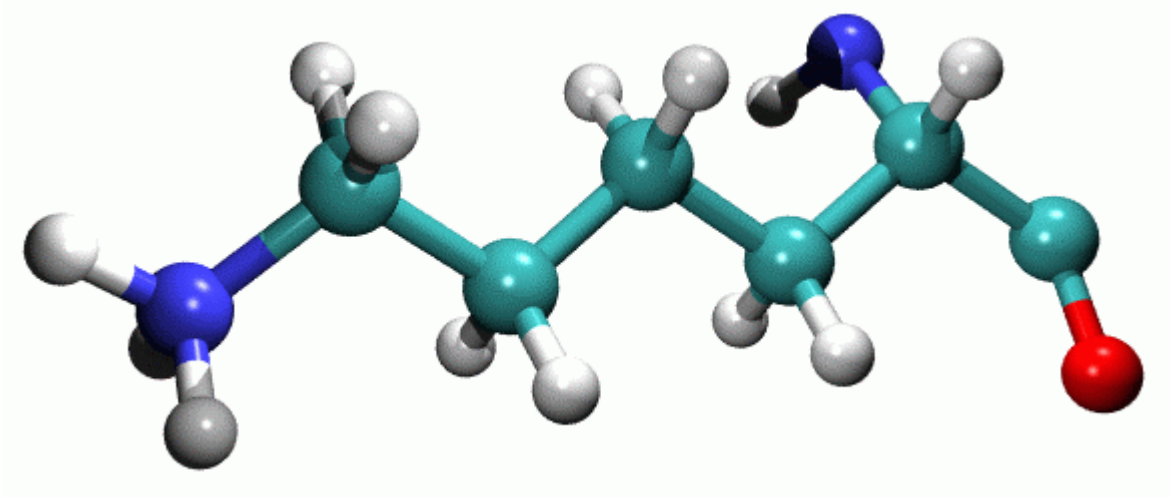
- Structure generation
- Filtering

- Final energetic evaluation

Semi-Flexible Protein Docking

- Structure generation
- Filtering
- Side-chain demangling
- Final energetic evaluation

Side-Chain Flexibility

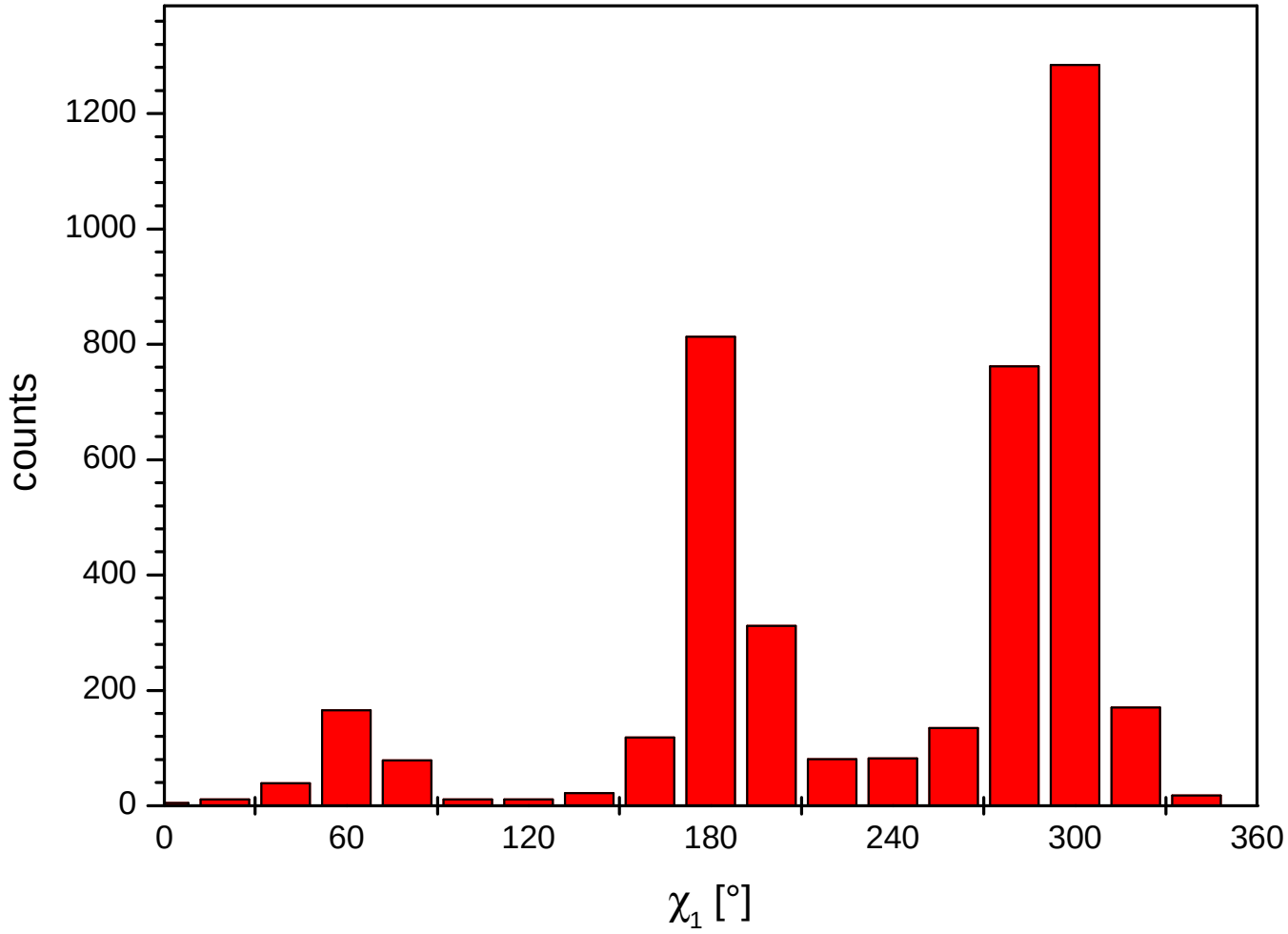


- Bond distances and bond angles constant
- Torsion angles variable

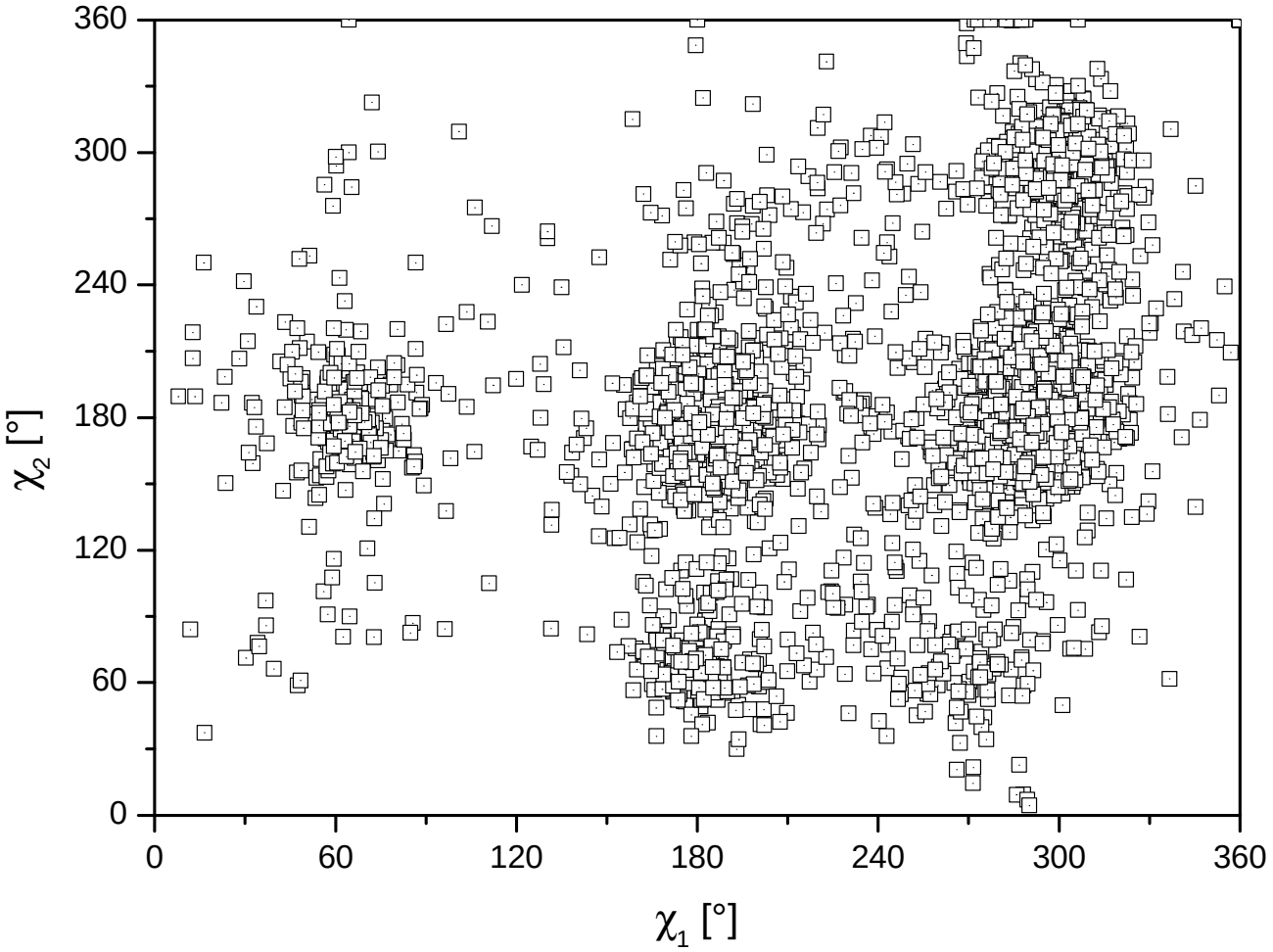
Torsion Angle Space

- Instead of 10 - 20 atoms with 3 coordinates
- Up to 4 torsion angles
- Typical binding site:
  - 50 amino acids, 600 atoms
  - 200 instead of 1800 degrees of freedom

Torsion angle distribution



Torsion angle distribution



Combinatorial Problem

- Identify set of rotamers with lowest energy
- Typical binding site: 50 side-chains
- Binding site defined via distance between **A**/**B**
- Number of possible combinations: ~10^{60}

Scoring Function

- No sophisticated energetic evaluation feasible
- Simple and fast scoring function: AMBER
- Decomposition into three components

$$E^{total} = E^{tpl} + \sum_{i} E^{tpl}_{i_r} + \sum_{i} \sum_{j<i} E^{pw}_{i_r, j_s}$$

Earlier Approaches

- Leach (1994): A* algorithm for side-chain optimization (ligand docking)
- Desmet et al. (1992): Dead End Elimination
  - Simple inequality
  - Iteratively applied

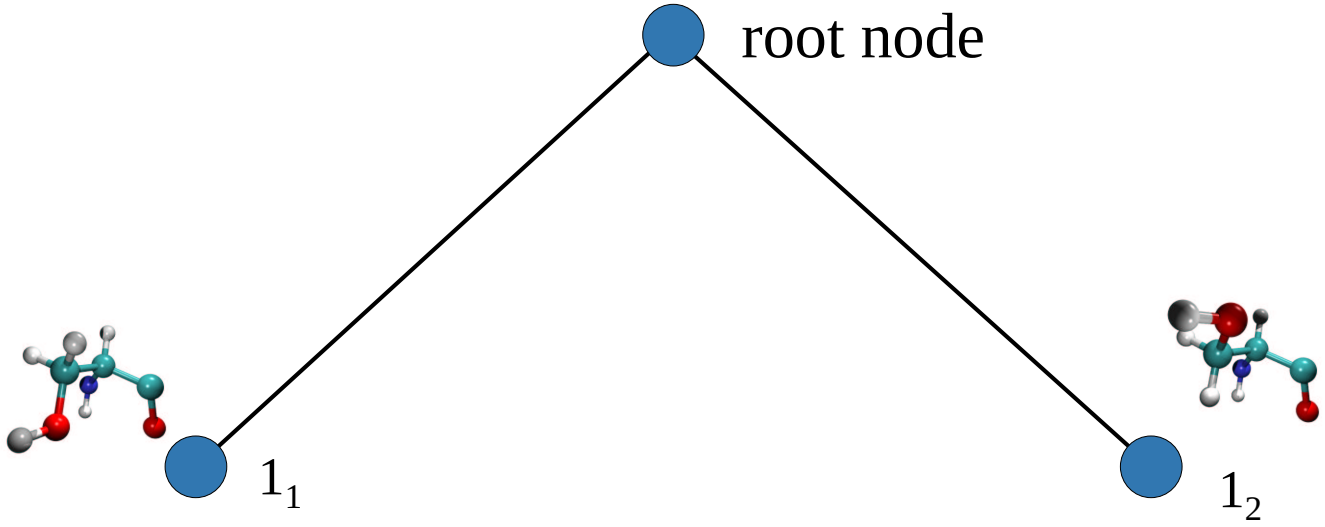
How to solve the problem?

- Multi Greedy method
  - Fast and simple heuristic
  - Suboptimal solutions
- ILP formulation
  - Optimal solution

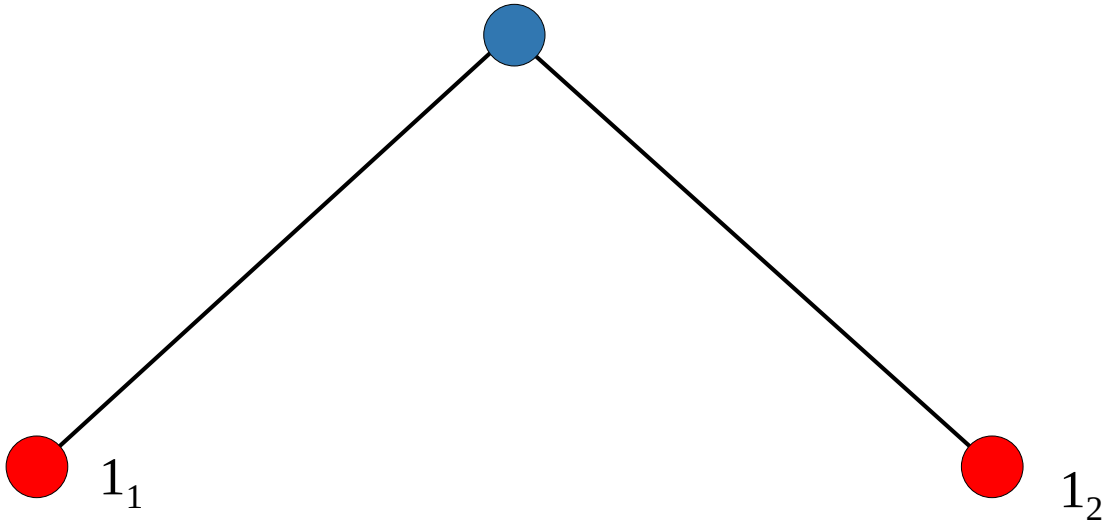
Multi Greedy Method

root node

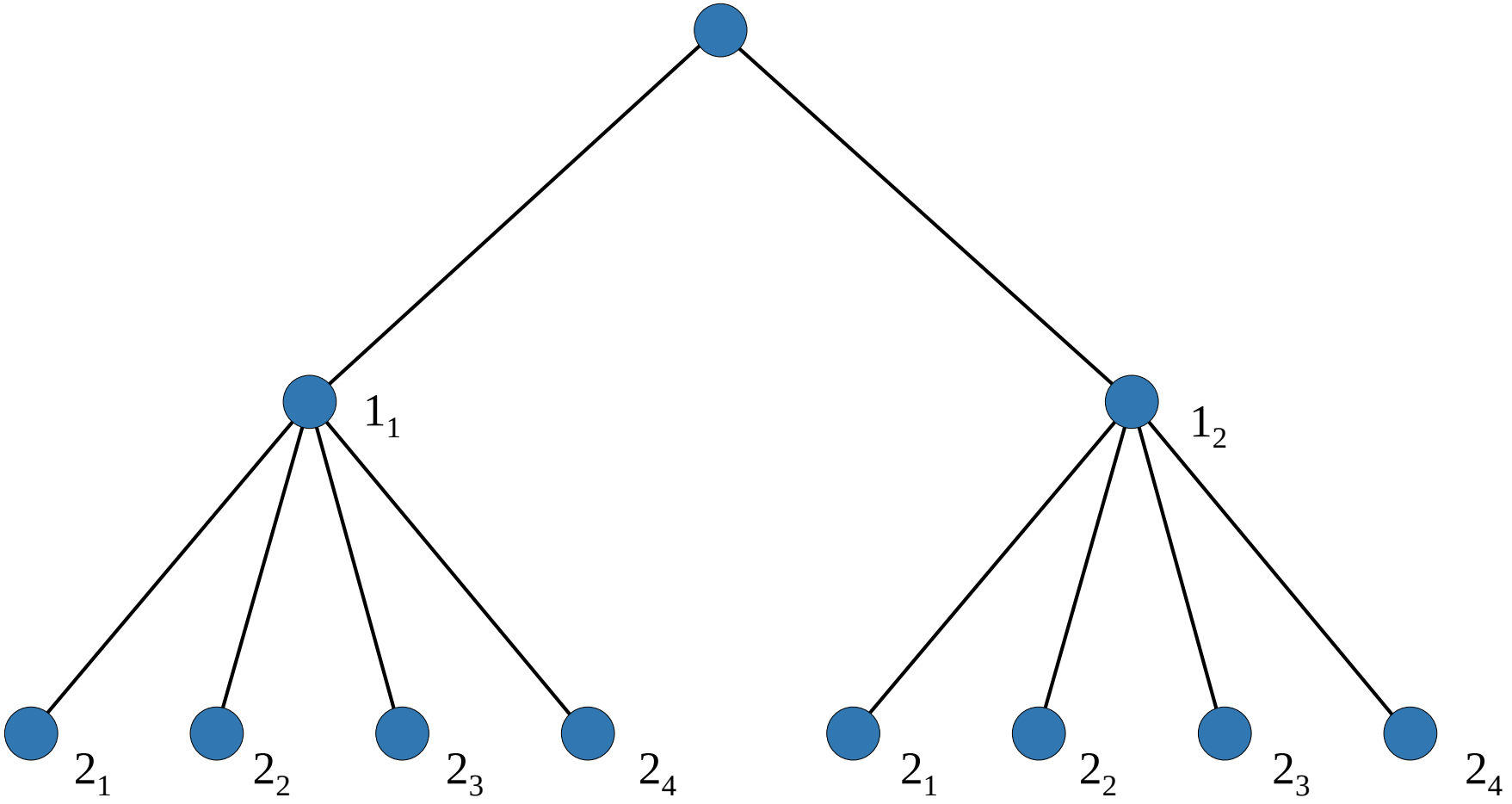
Multi Greedy Method



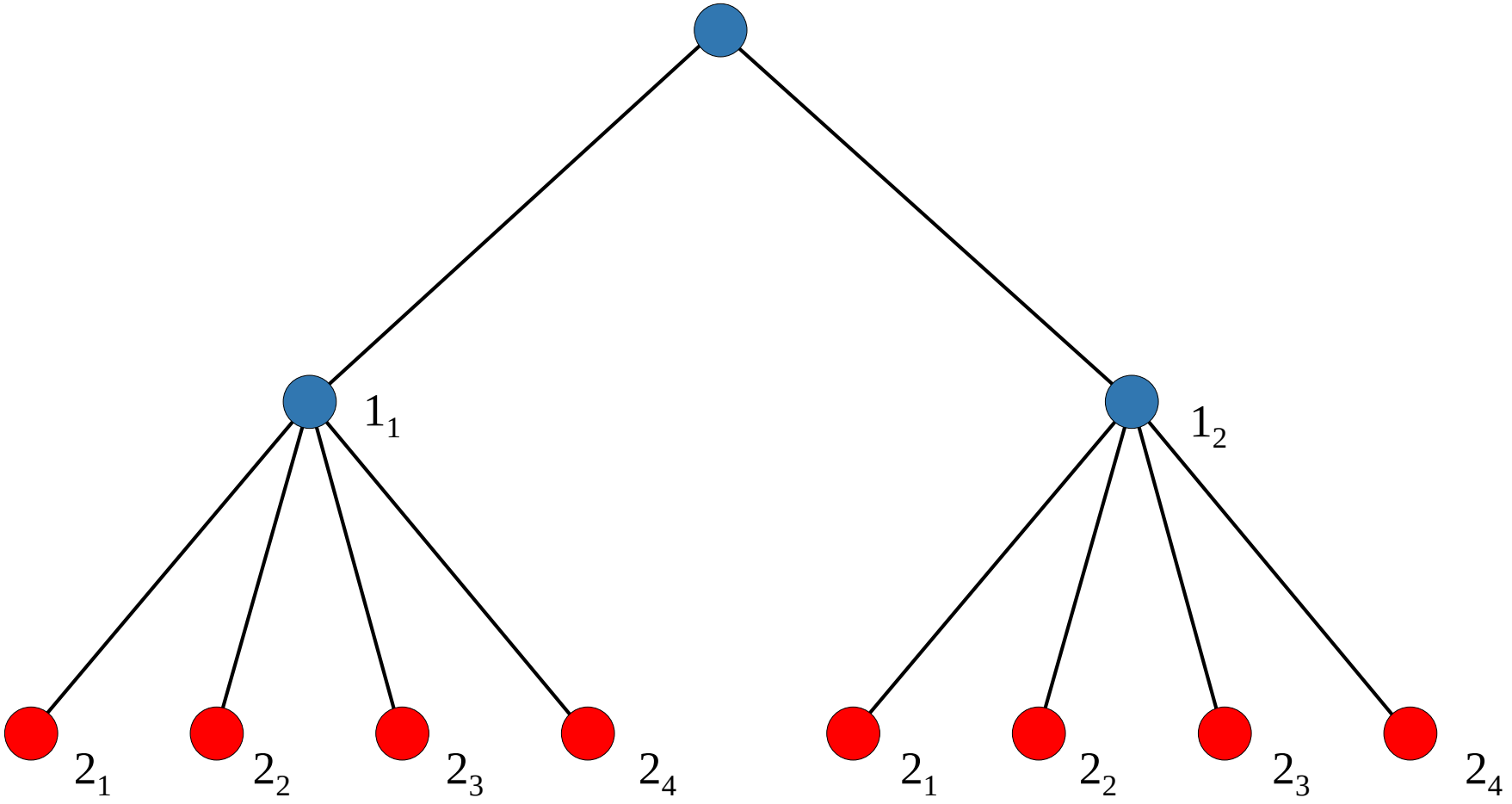
Multi Greedy Method



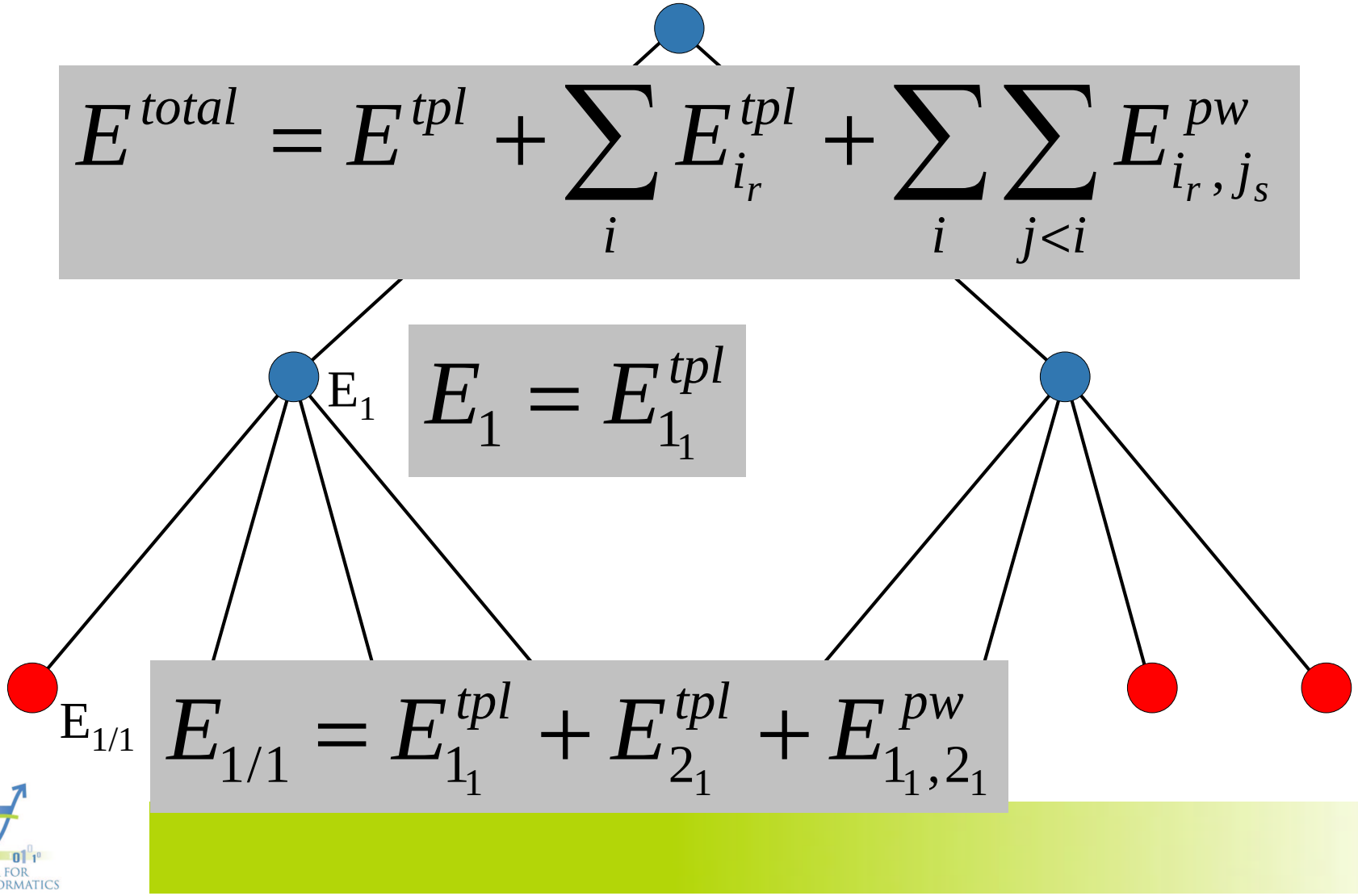
Multi Greedy Method



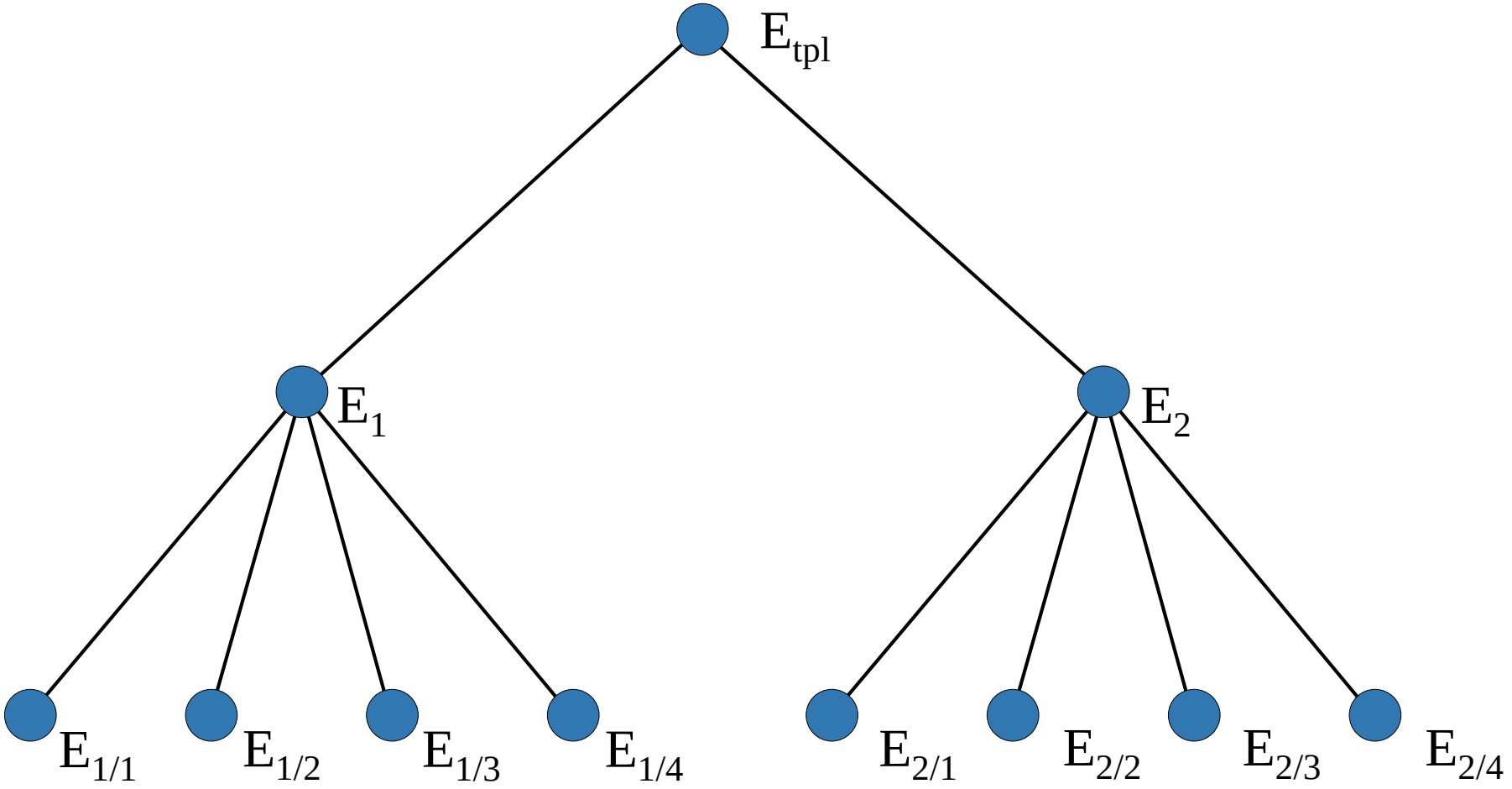
Multi Greedy Method



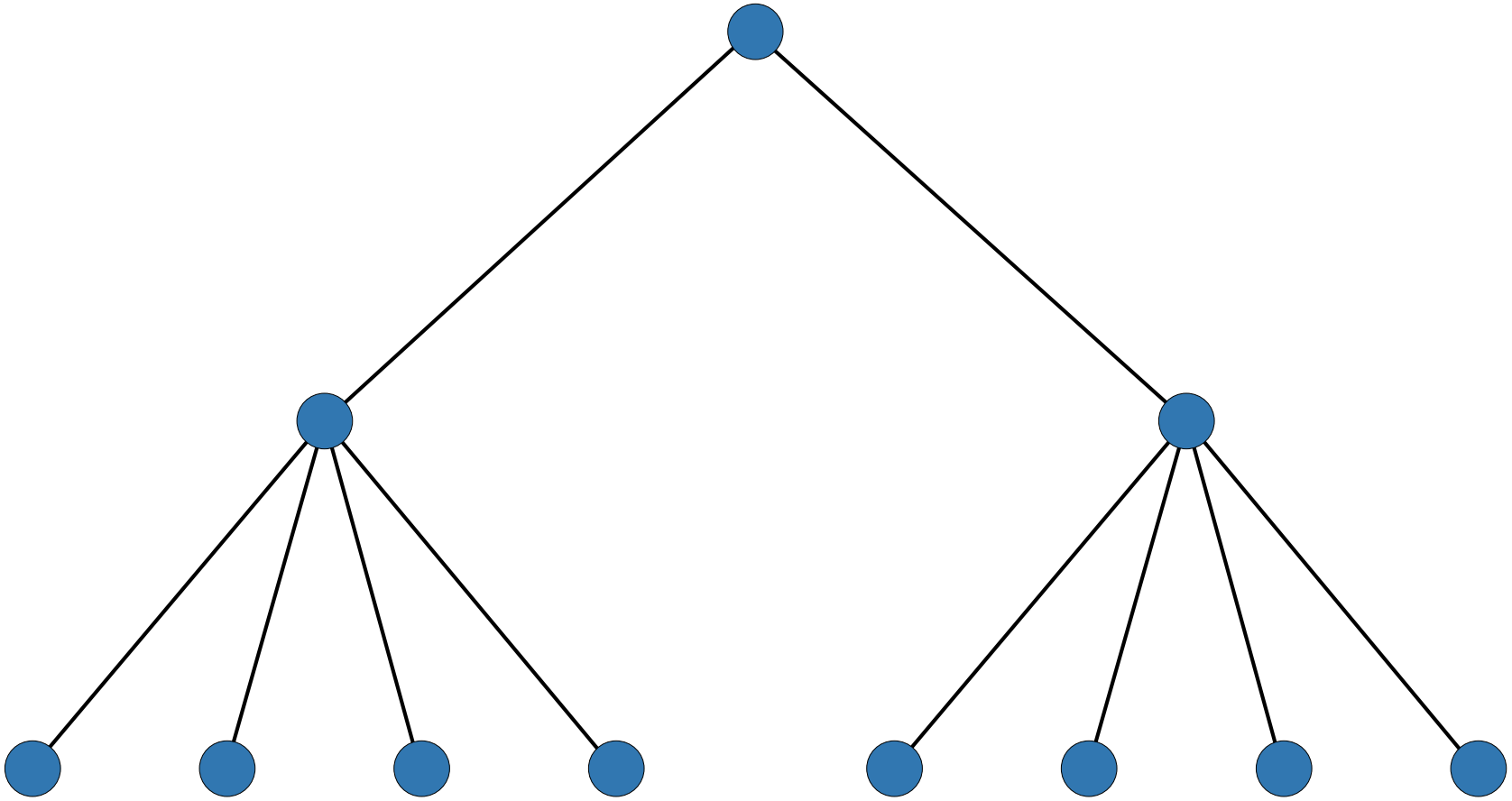
Multi Greedy Method



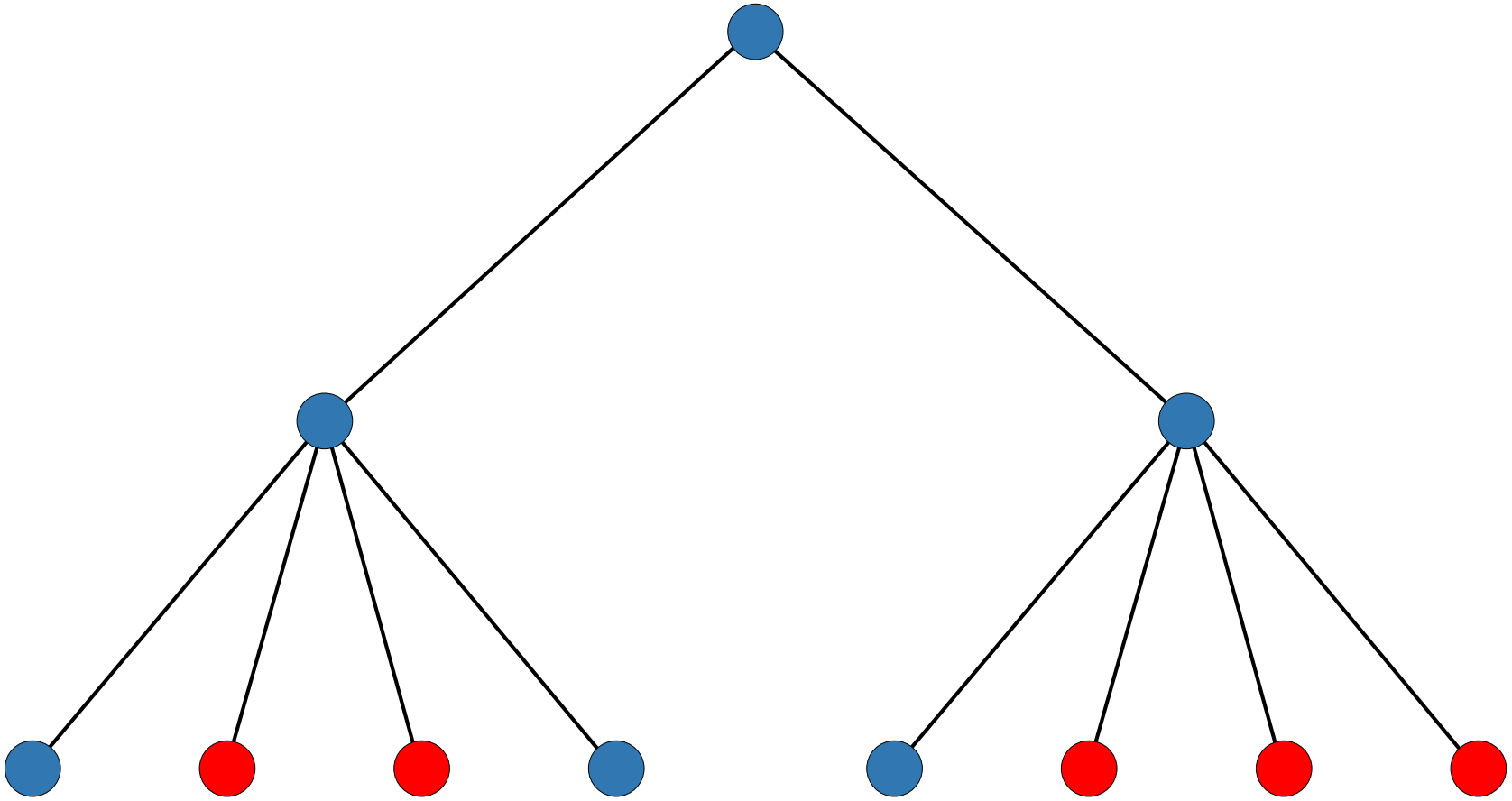
Multi Greedy Method



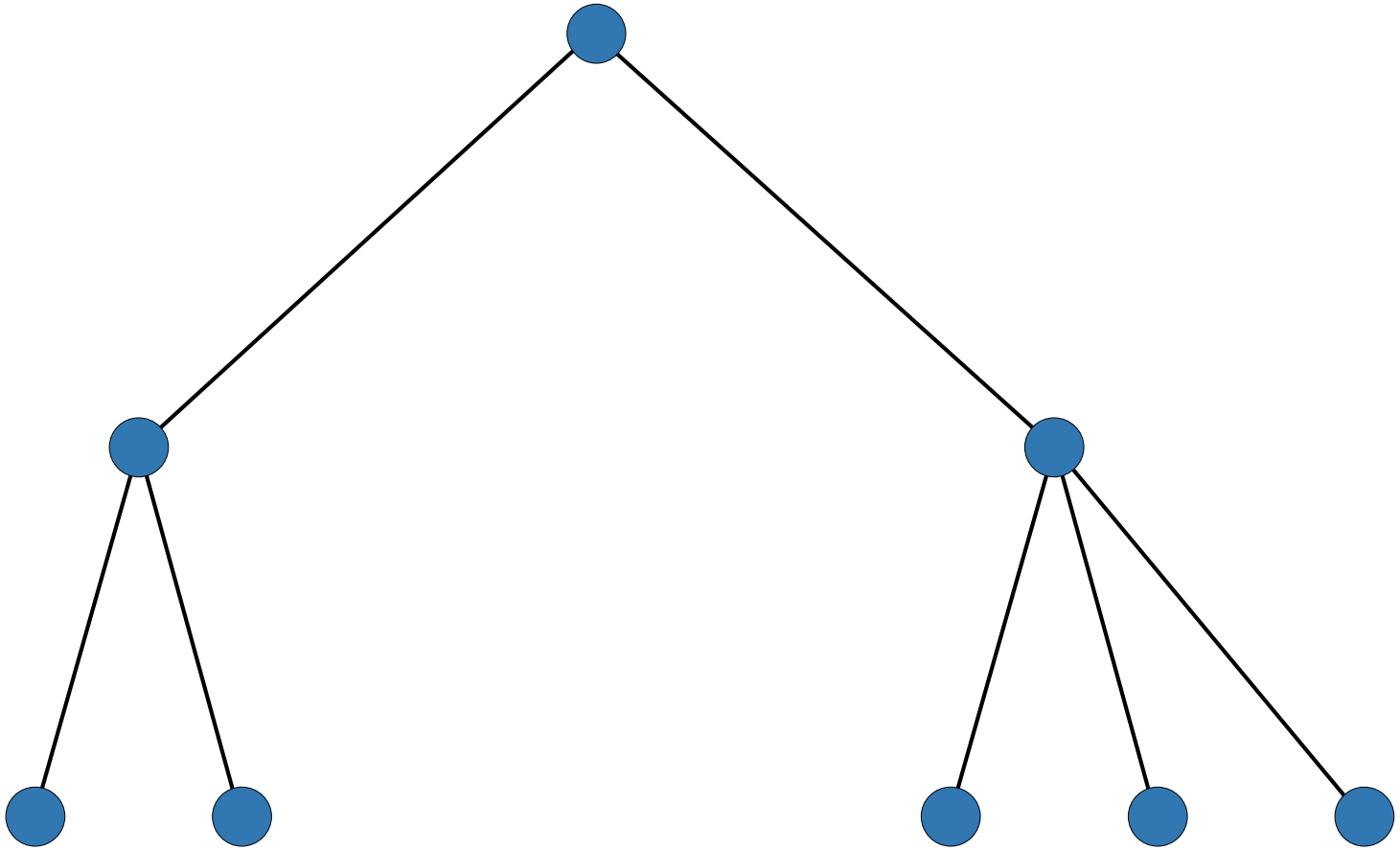
Multi Greedy Method



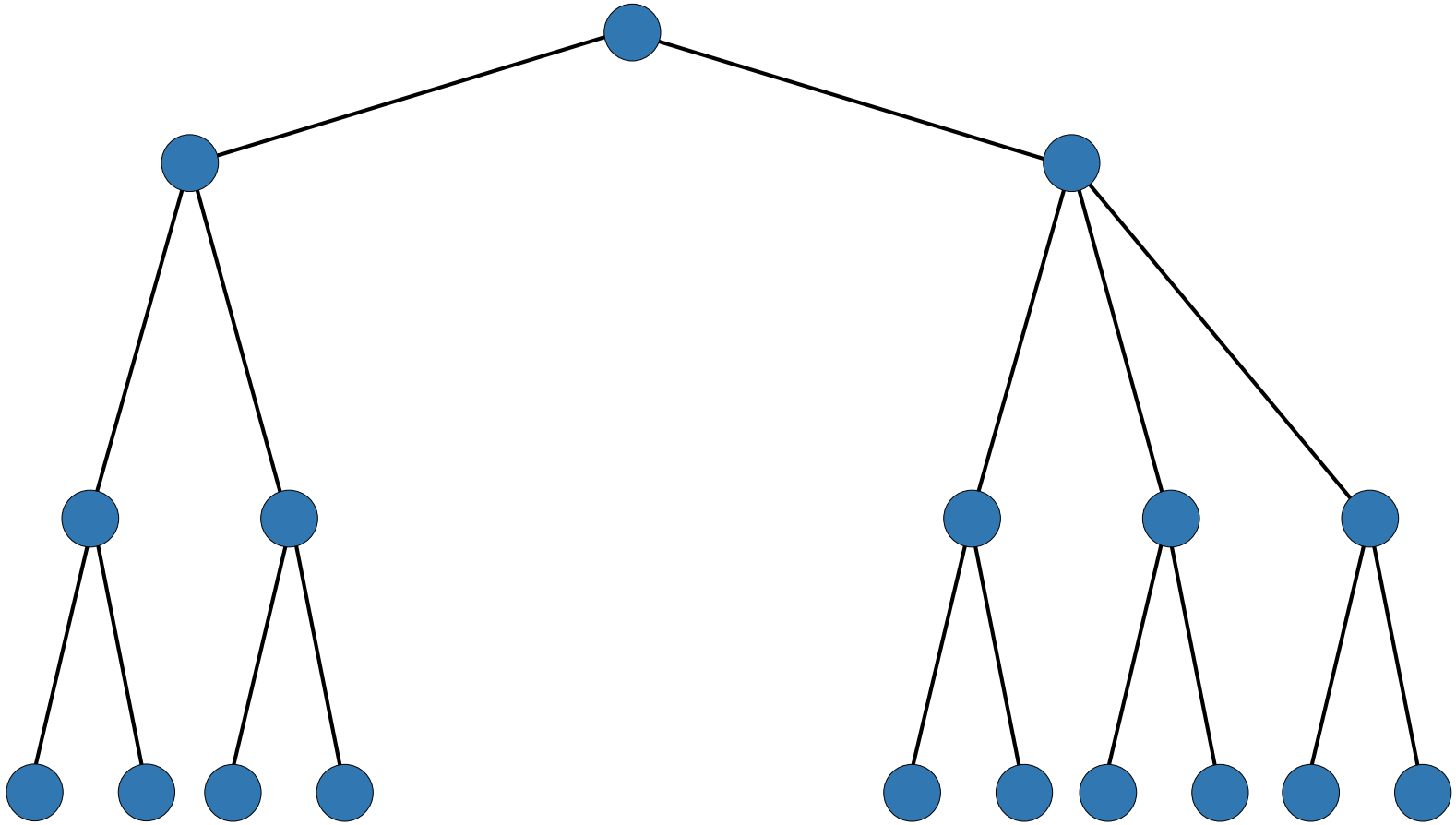
Multi Greedy Method



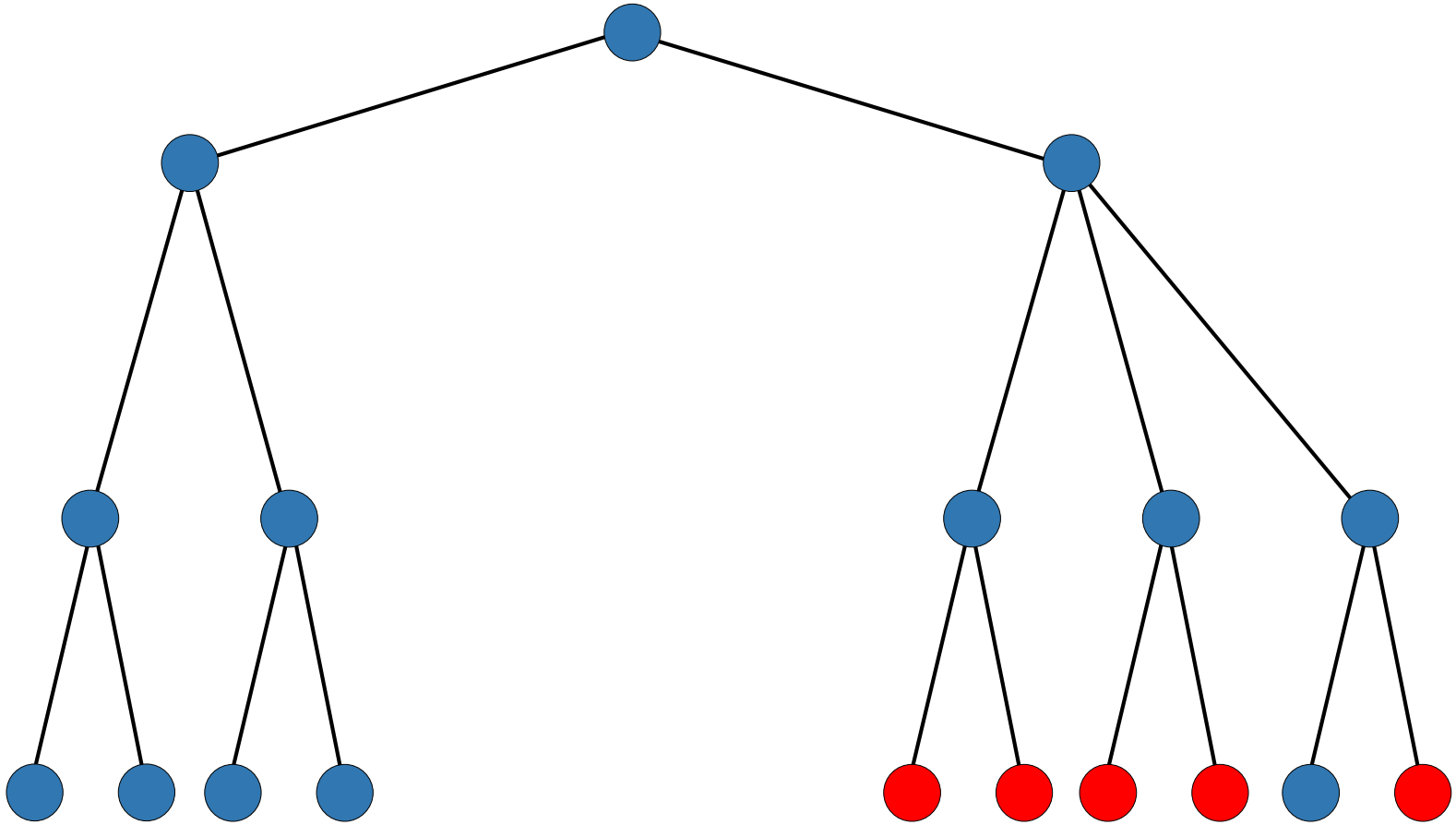
Multi Greedy Method



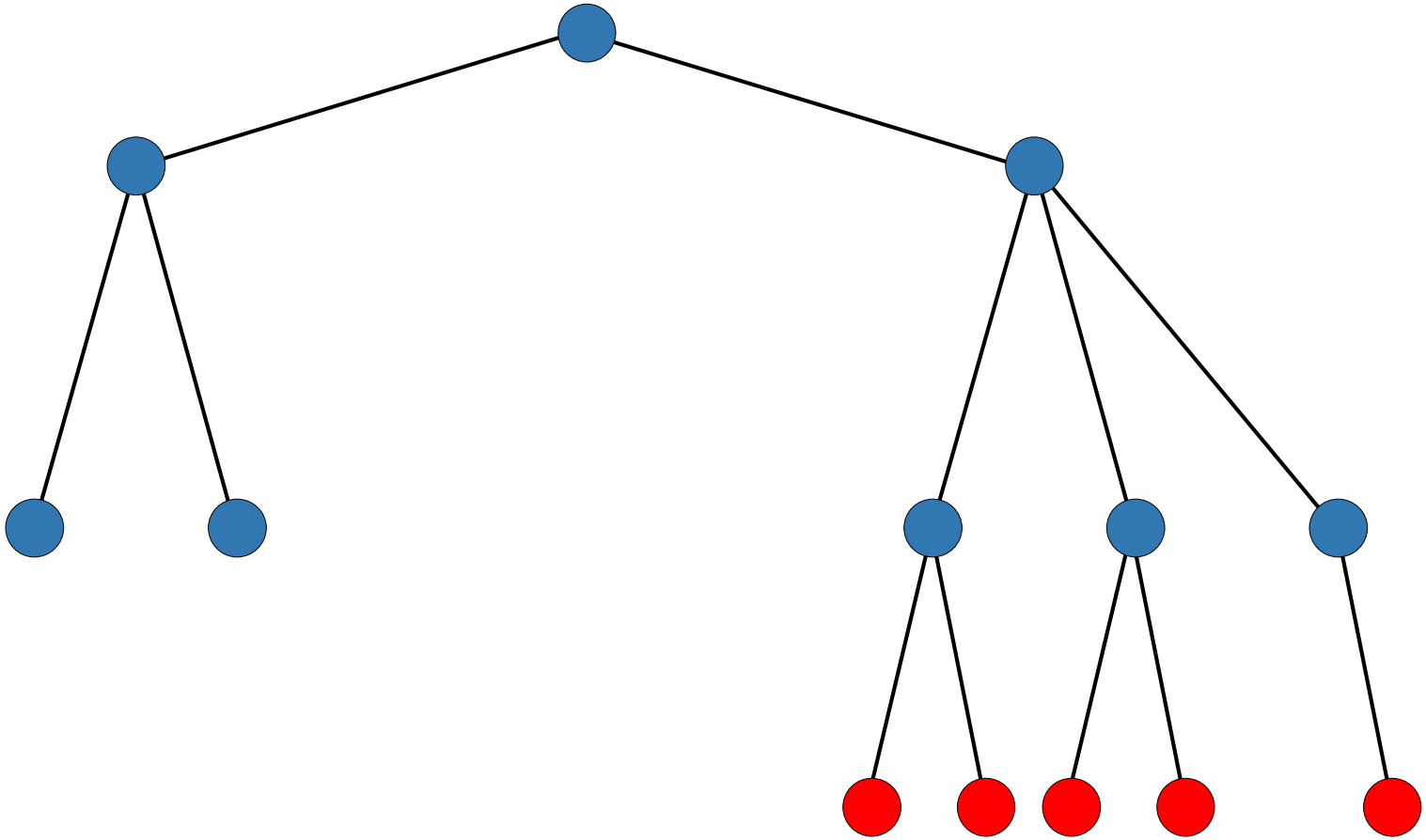
Multi Greedy Method



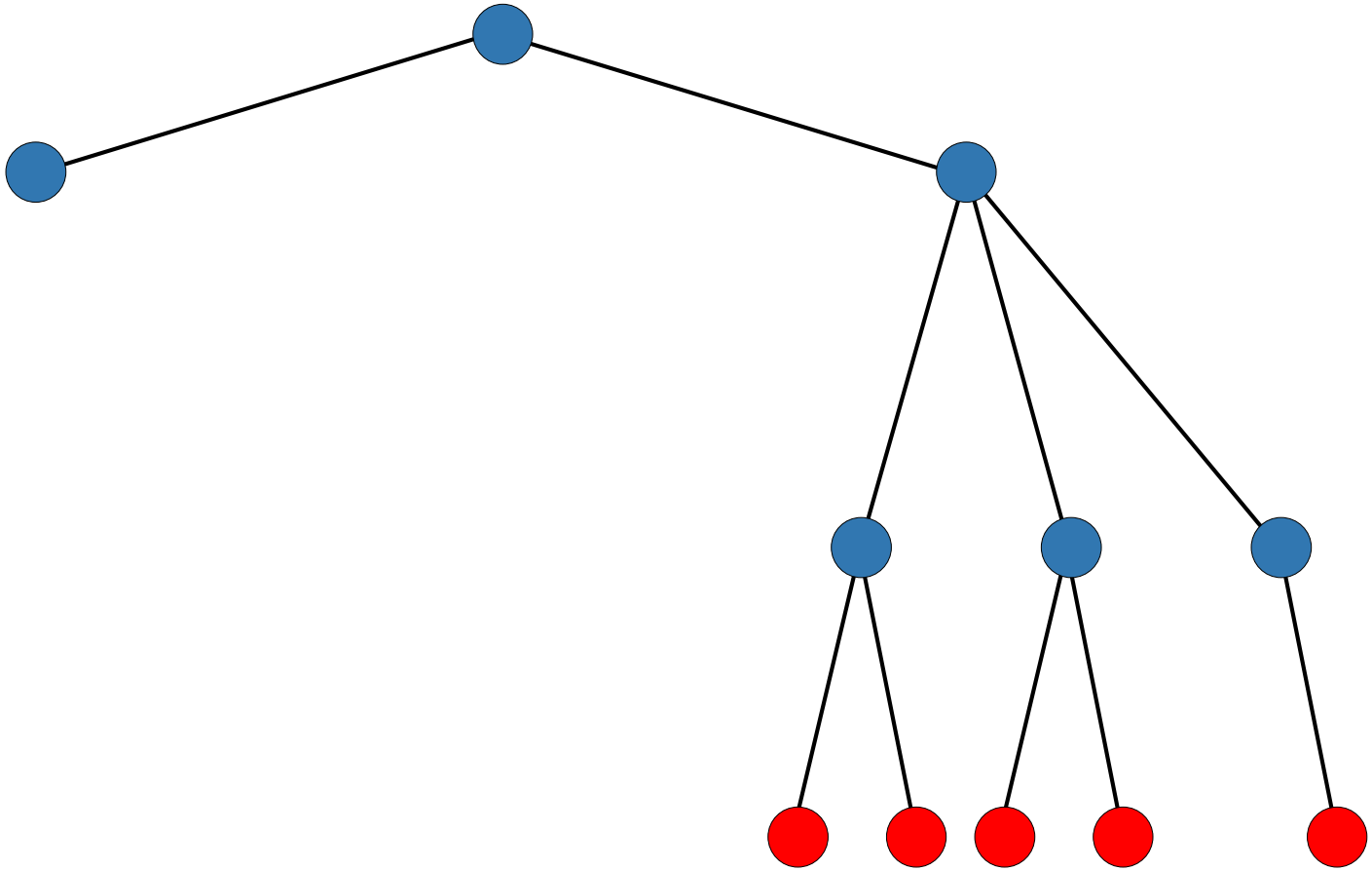
Multi Greedy Method



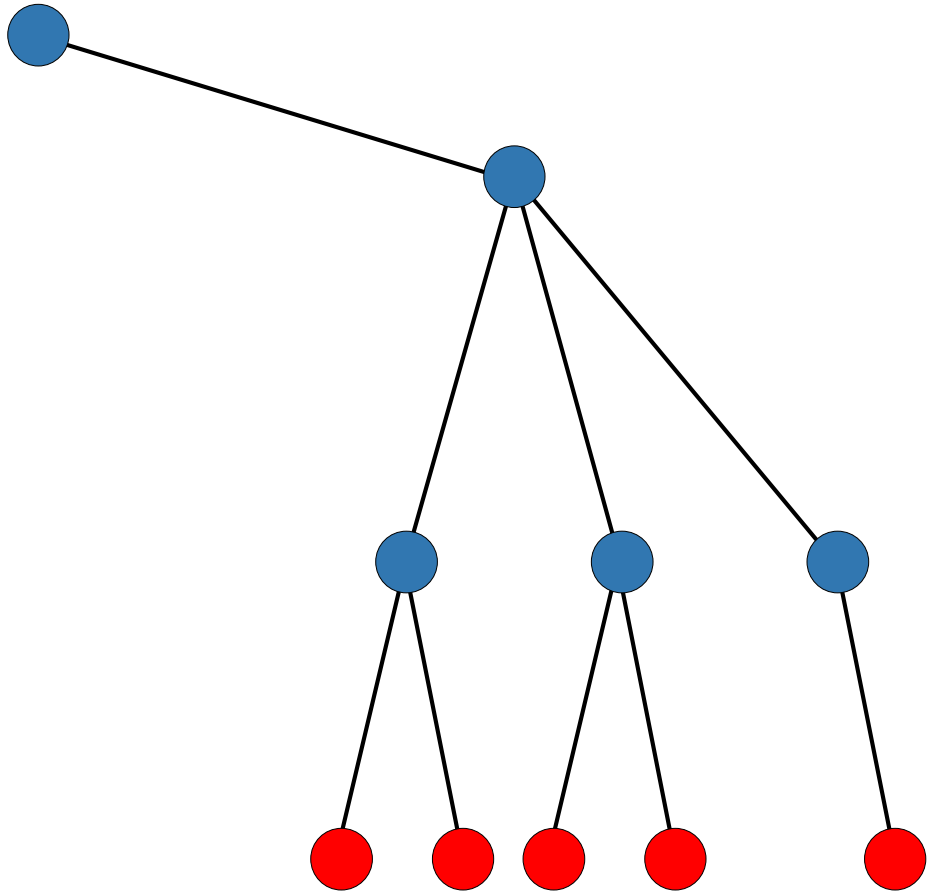
Multi Greedy Method



Multi Greedy Method



Multi Greedy Method



Multi Greedy Approach

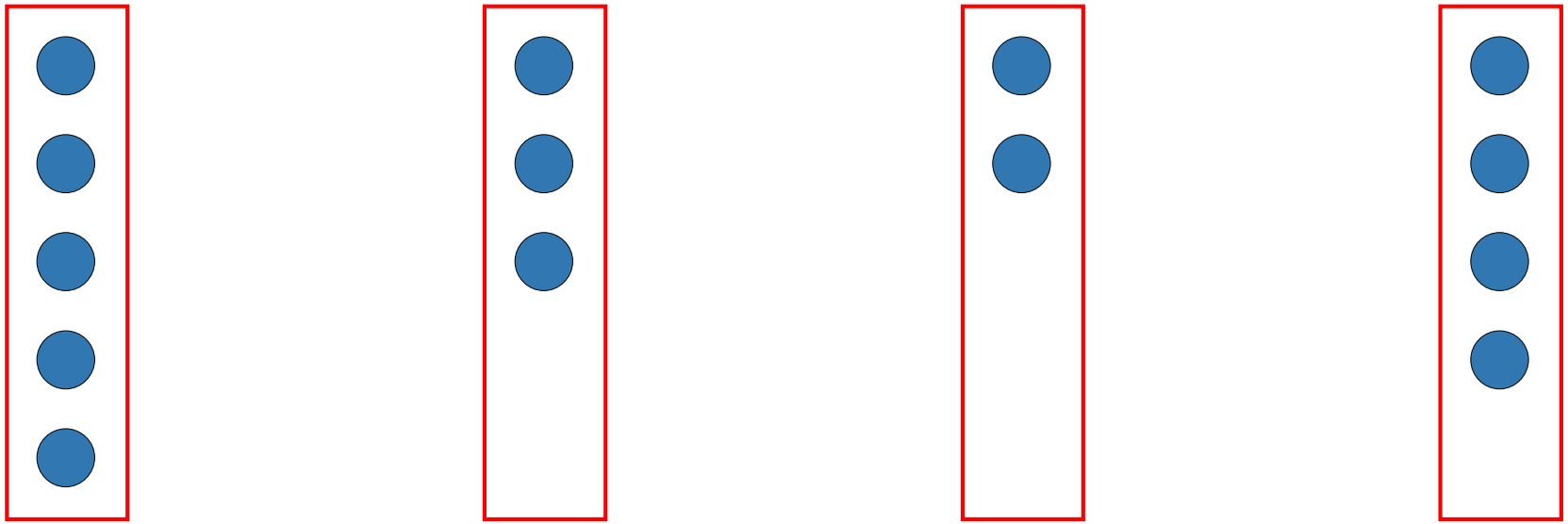
- Fast
- Correctly demangles side chains for test set
- May yield suboptimal solutions
- How to determine the quality of the solution?
- Optimal algorithm: based on polyhedral optimization, ILP formulation

Graph Problem

For each rotamer *r* of side-chain *i* create a node v_{i_r} with weight

$$E(v_{i_r}) = E_{i_r}^{tpl}$$

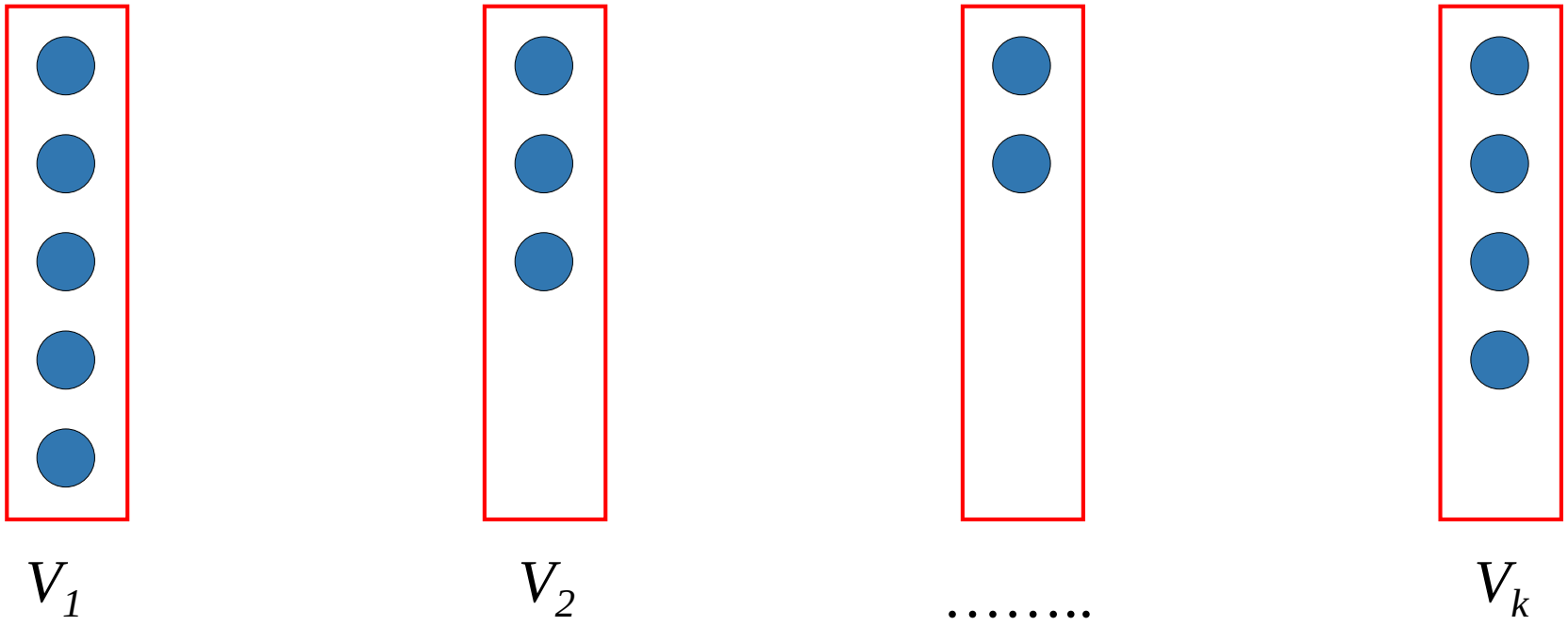
Graph Problem



For each rotamer *r* of side-chain *i* create a node v_{i_r} with weight

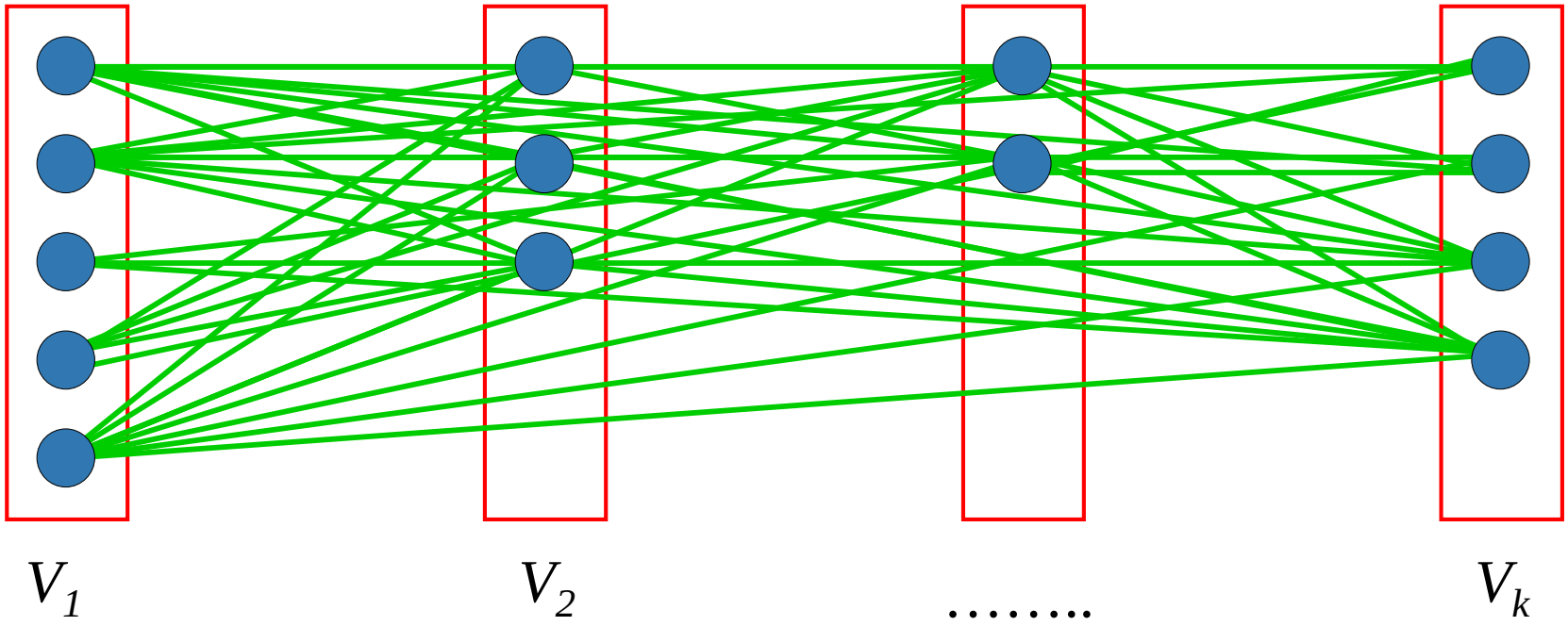
$$E(v_{i_r}) = E_{i_r}^{tpl}$$

Graph Problem



A k-partite graph with partitions $V_1 \ldots V_k$ is constructed

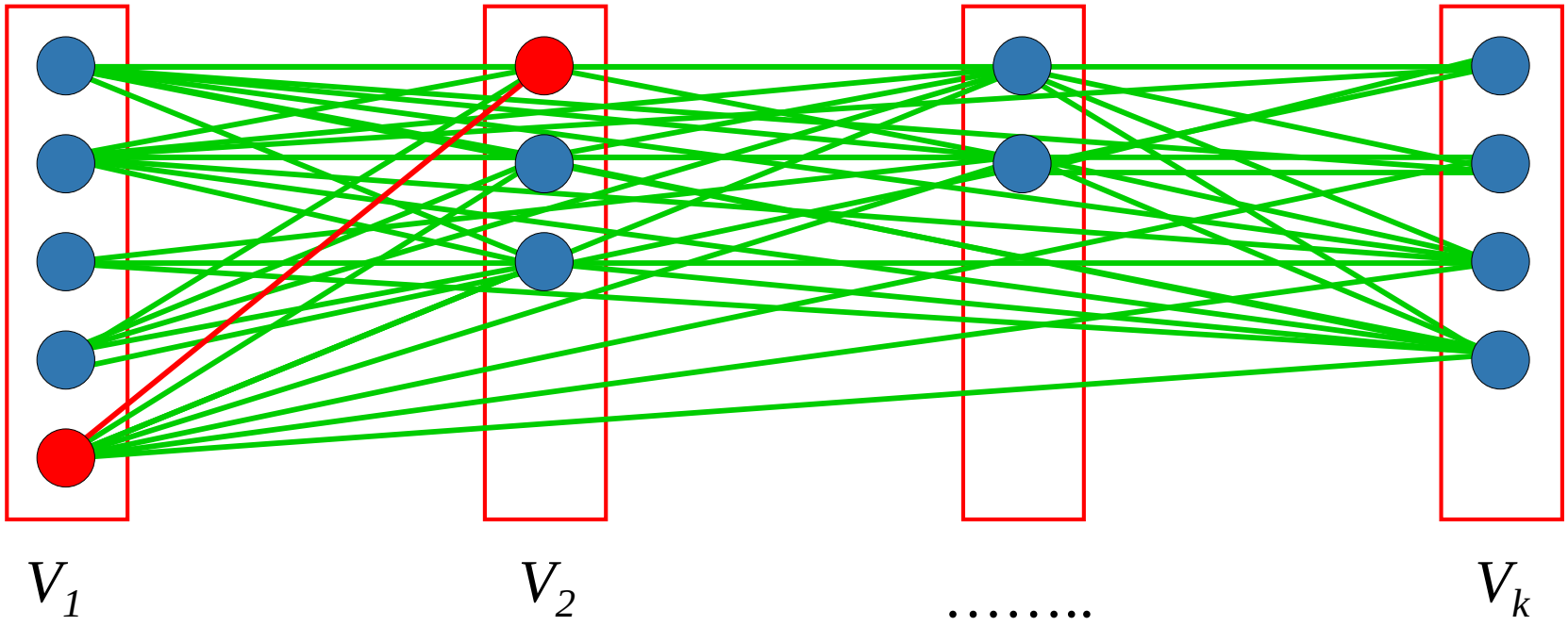
Graph Problem



Pairwise interactions are represented by edges *uv* with weight

$$E(uv) = E^{pw}_{i_r, j_s}$$

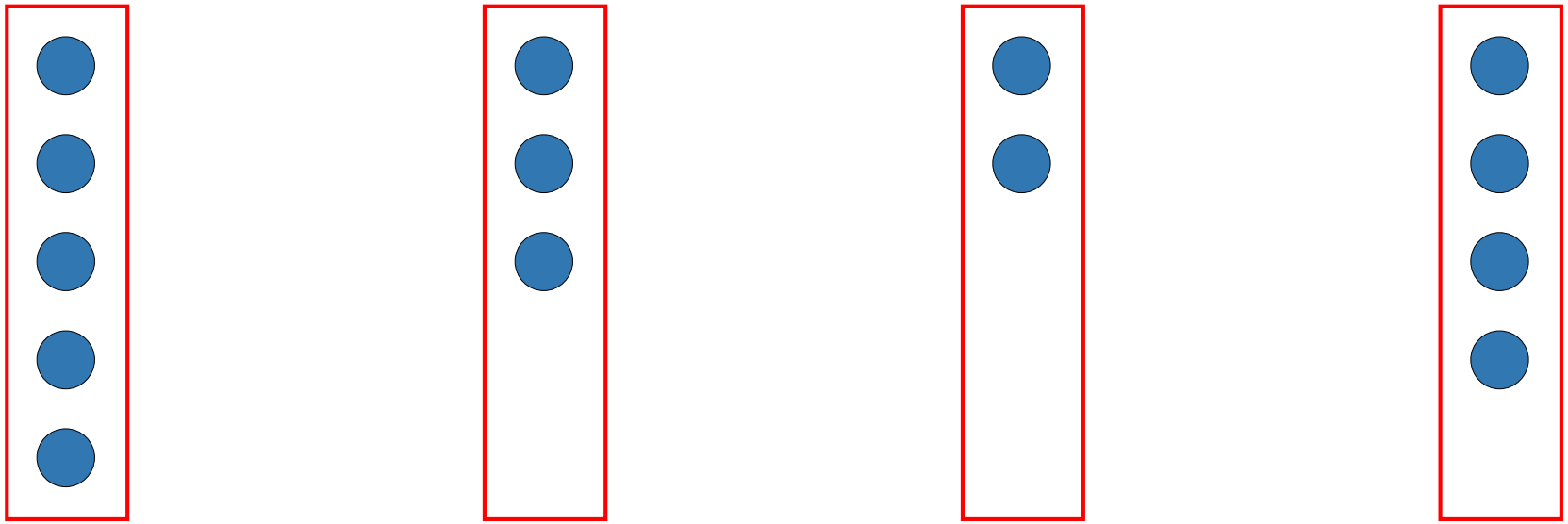
Graph Problem



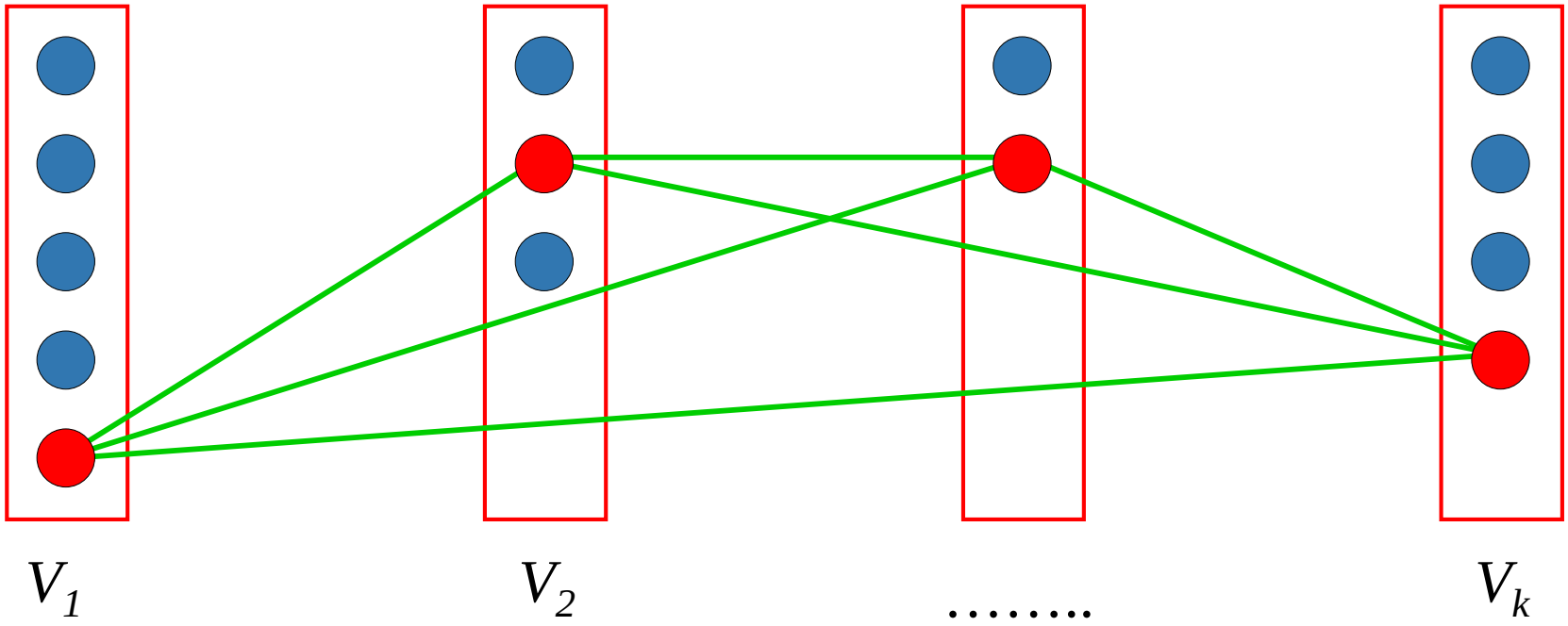
Pairwise interactions are represented by edges *uv* with weight

$$E(uv) = E^{pw}_{i_r, j_s}$$

Graph Problem

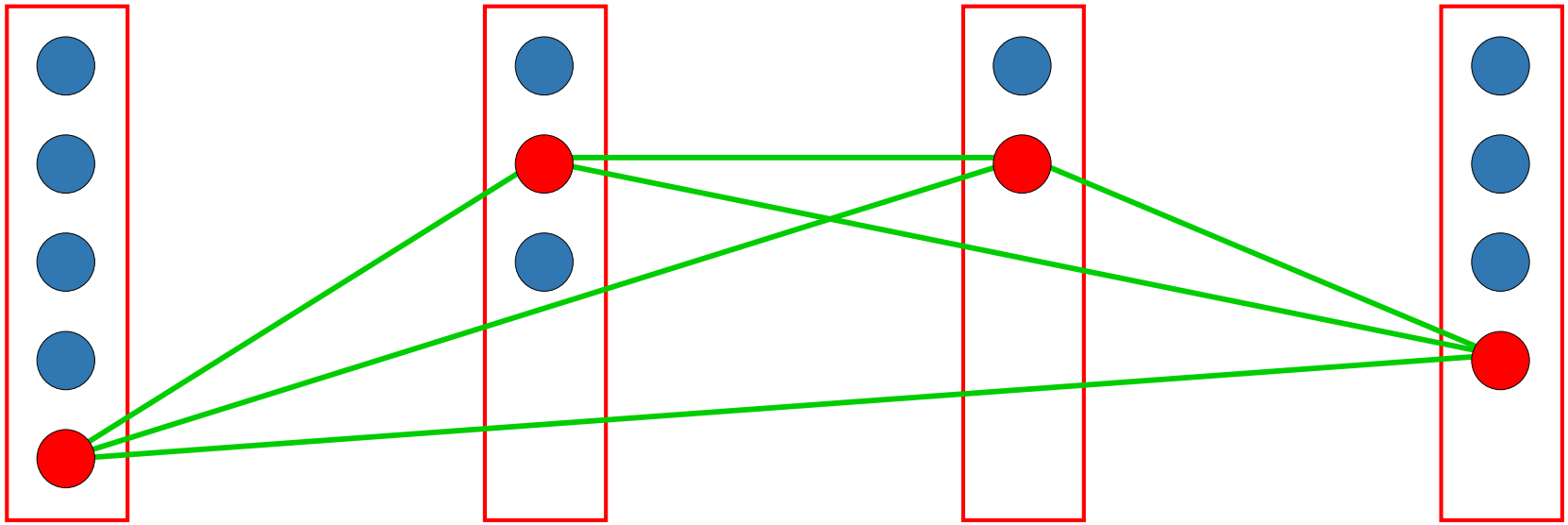


Graph Problem



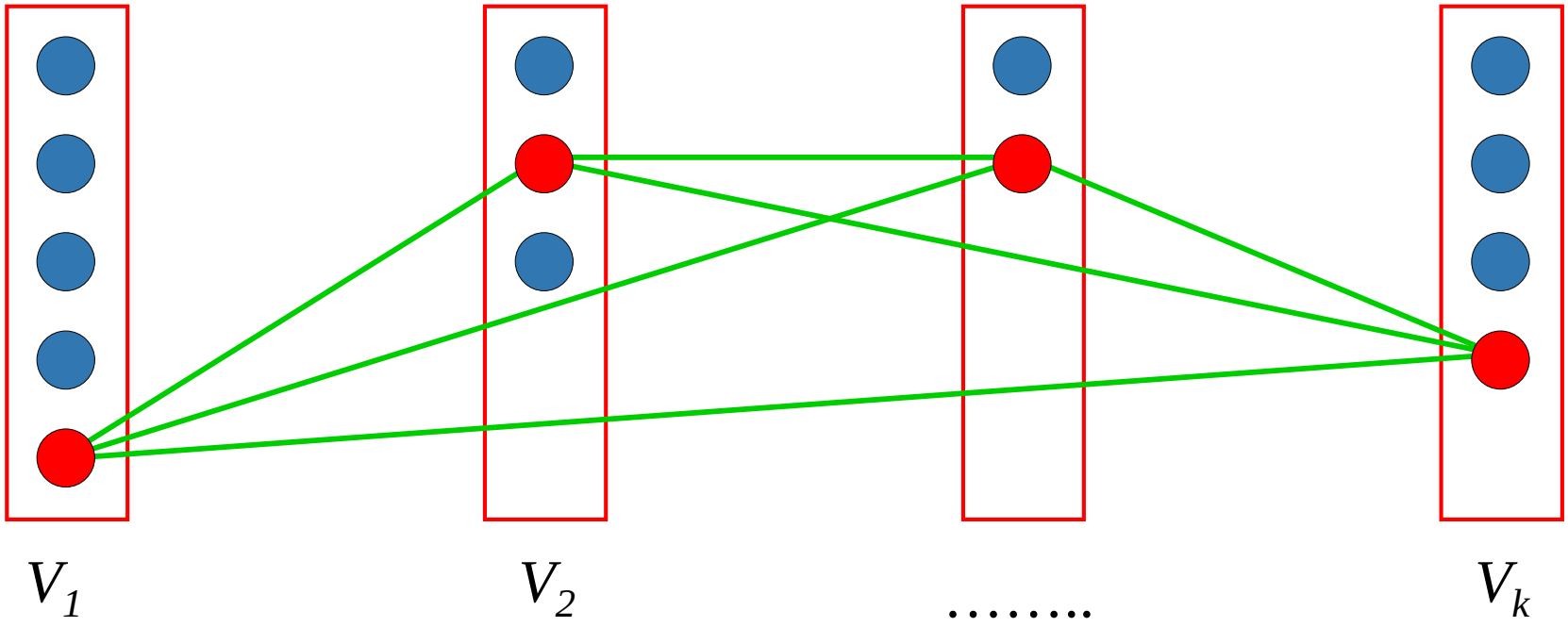
rotamer graph

Graph Problem



$$E^{total} = \sum_i E^{tpl}_{i_r} + \sum_i \sum_{j<i} E^{pw}_{i_r, j_s}$$

Graph Problem



$$E(RG) = \sum_{v \in RG} E(v) + \sum_{uv \in RG} E(uv)$$

Graph Problem

- Binary decision variables for nodes and edges
  - x_v for each node v
  - x_{uv} for each edge uv
- Node is *selected* if $x_v = 1$
- Edge is *selected* if $x_{uv} = 1$

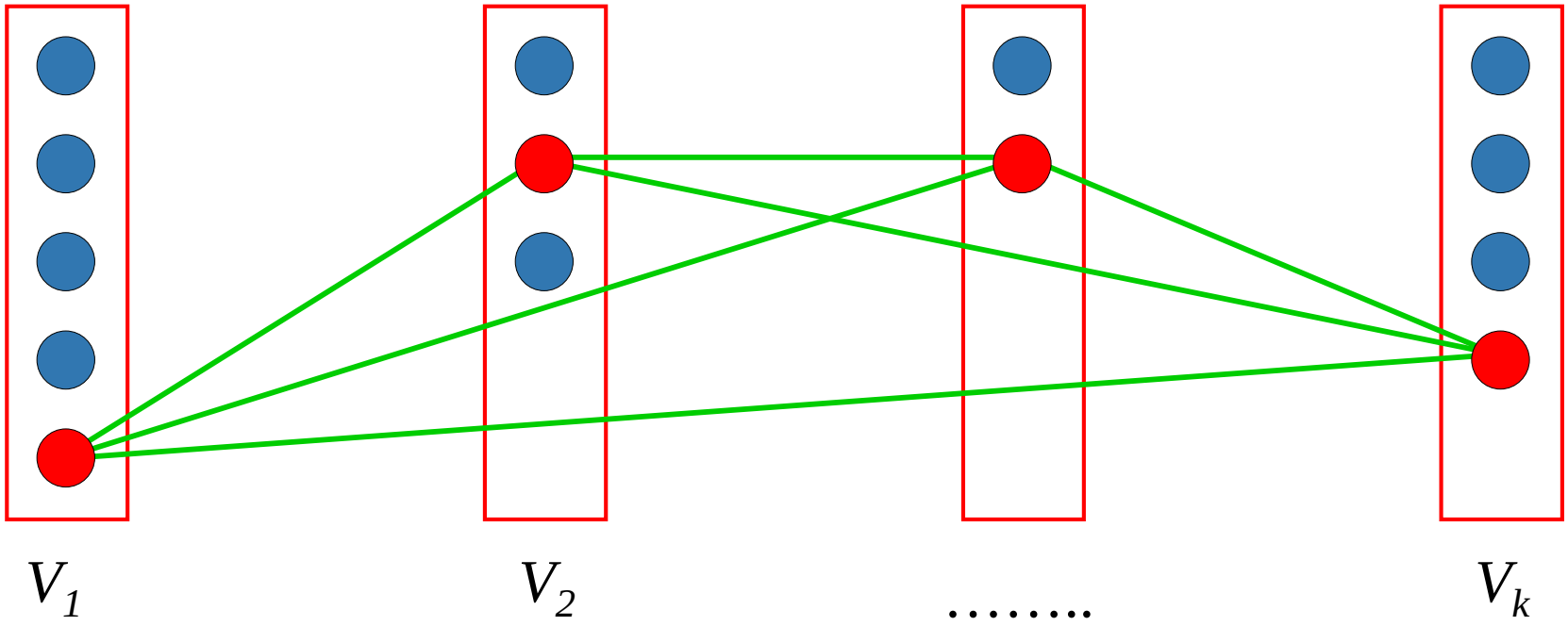
$$E(RG) = \sum_{v \in RG} E(v) + \sum_{uv \in RG} E(uv)$$

Graph Problem

- Binary decision variables for nodes and edges
  - x_v for each node v
  - x_{uv} for each edge uv
- Node is *selected* if $x_v = 1$
- Edge is *selected* if $x_{uv} = 1$

$$E = \sum_{v \in V} x_v E(v) + \sum_{uv \in E} x_{uv} E(uv)$$

Graph Problem



$$E = \sum_{v \in V} x_v E(v) + \sum_{uv \in E} x_{uv} E(uv)$$

Integer Linear Program

$$E = \sum_{v \in V} x_v E(v) + \sum_{uv \in E} x_{uv} E(uv)$$

Determine x_v, x_{uv} that minimize E

ILP: Basic Constraint System

$$\min\left[\sum_{v\in V} x_v E(v) + \sum_{uv\in E} x_{uv} E(uv)\right]$$

s.t. $$\sum_{v\in V_i} x_v = 1 \quad \text{for all } i \in \{1...k\}$$

$$x_{uv} \le x_u \quad \text{for all } uv \in E$$

$$x_{uv} \le x_v \quad \text{for all } uv \in E$$

ILP: Basic Constraint System

$$\min \left[\sum_{v \in V} x_v \left(E(v) - E_{\max} \right) + \sum_{uv \in E} x_{uv} \left(E(uv) - E_{\max} \right) \right]$$

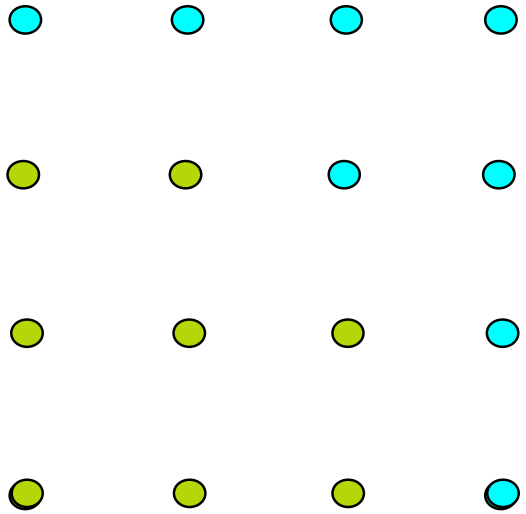
s.t. $$\sum_{v \in V_i} x_v = 1 \quad \text{for all } i \in \{1 \ldots k\}$$

$$x_{uv} \leq x_u \quad \text{for all } uv \in E$$

$$x_{uv} \leq x_v \quad \text{for all } uv \in E$$

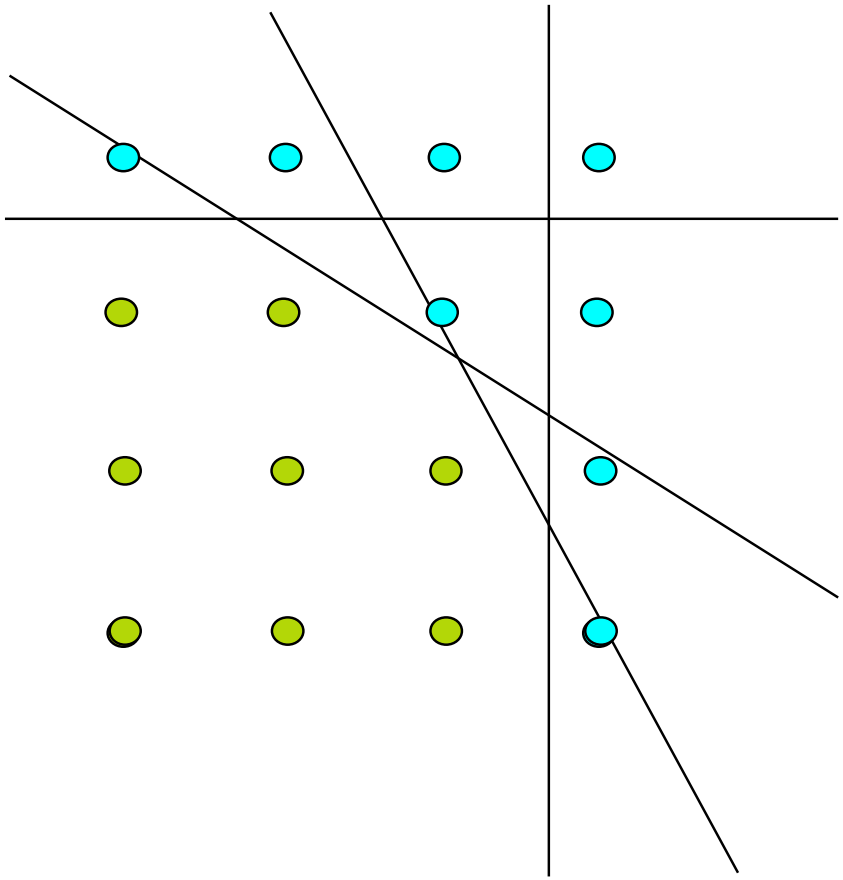
Branch & Cut

Solve ILP



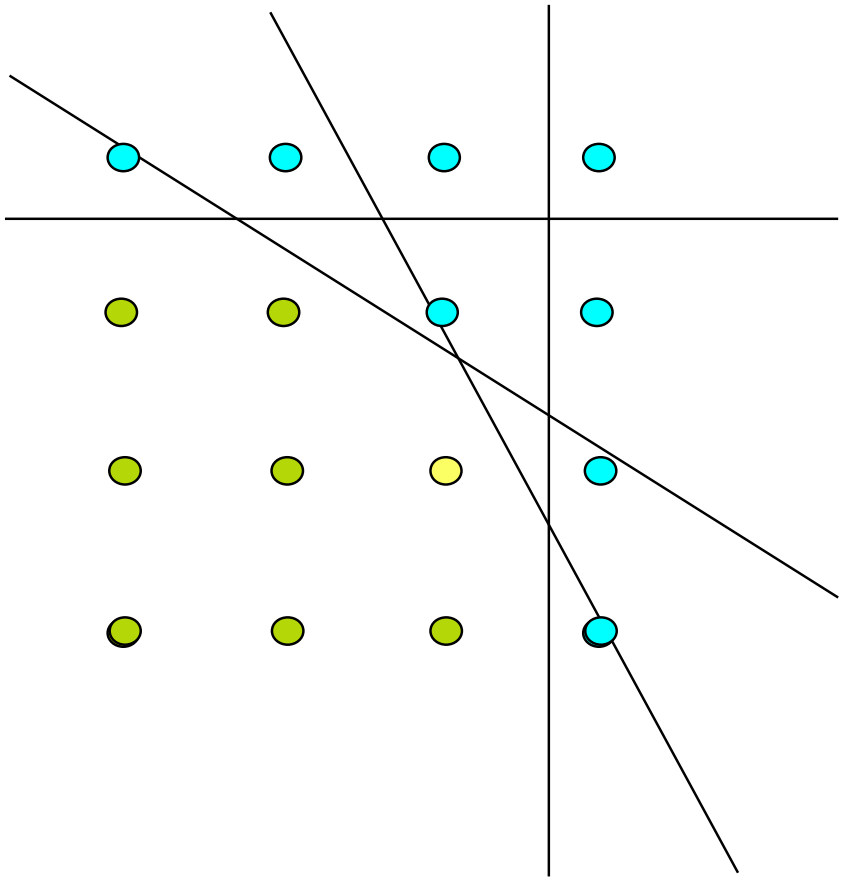
Branch & Cut

Solve ILP

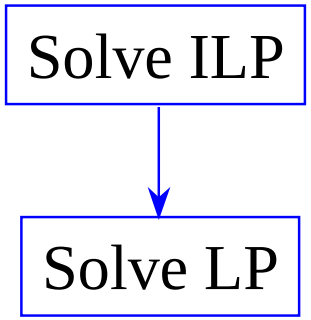


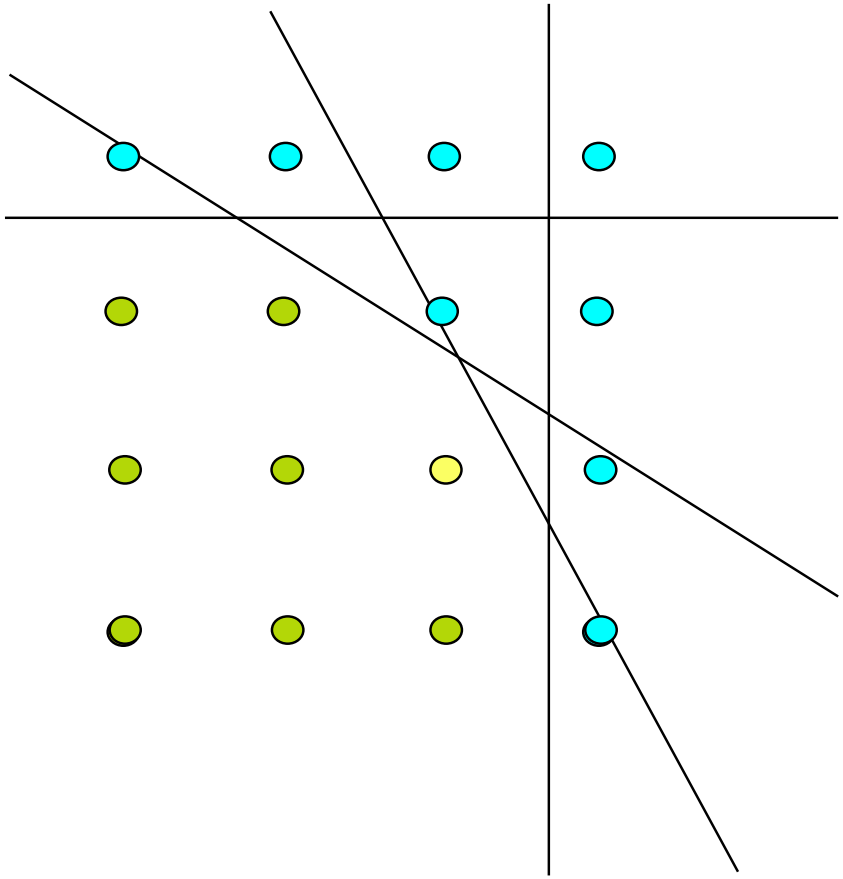
Branch & Cut

Solve ILP

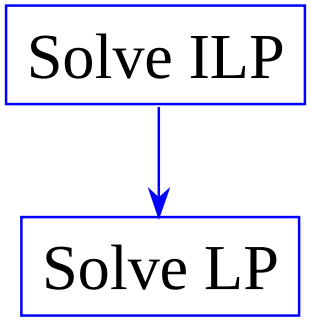


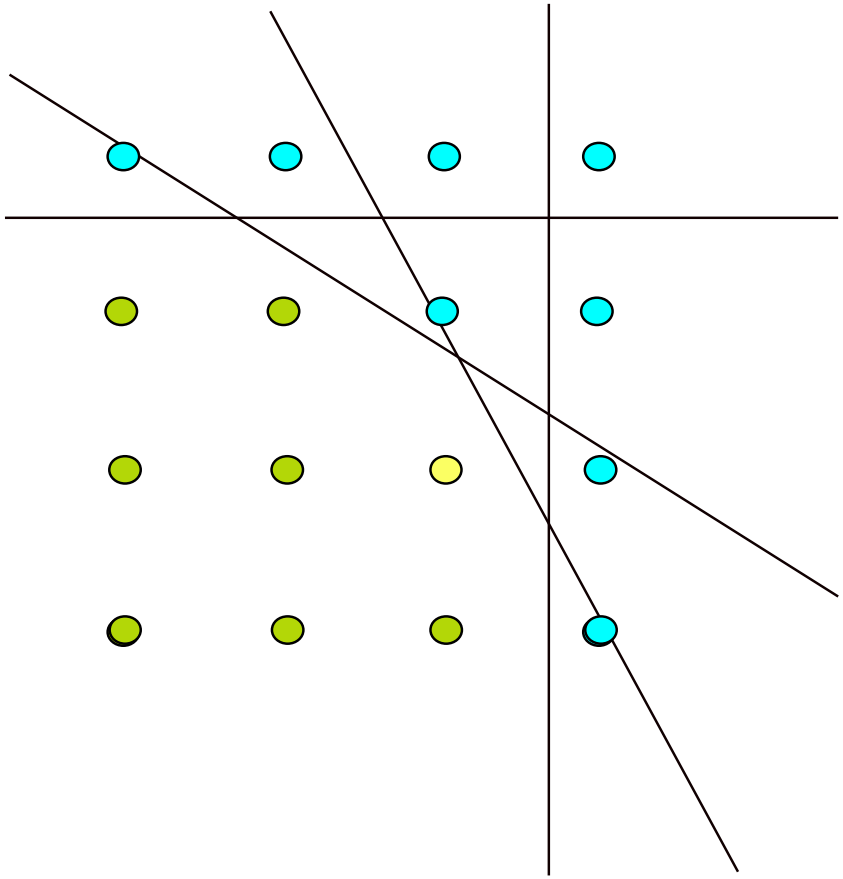
Branch & Cut



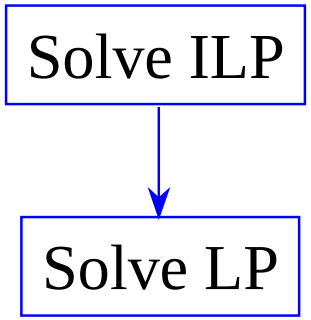


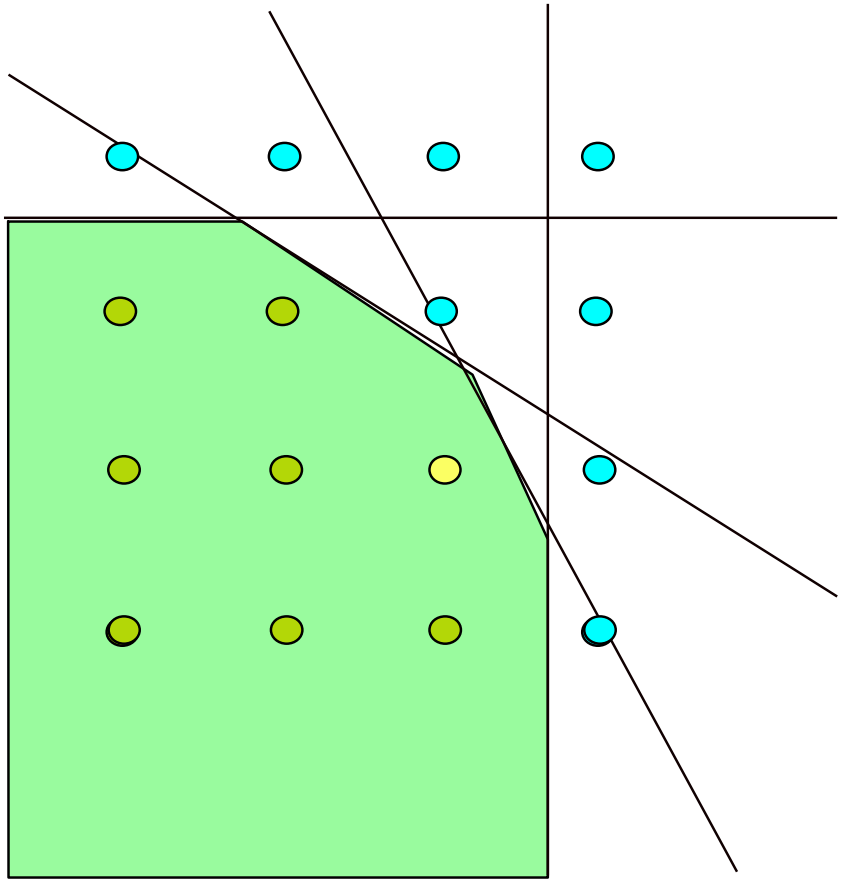
Branch & Cut



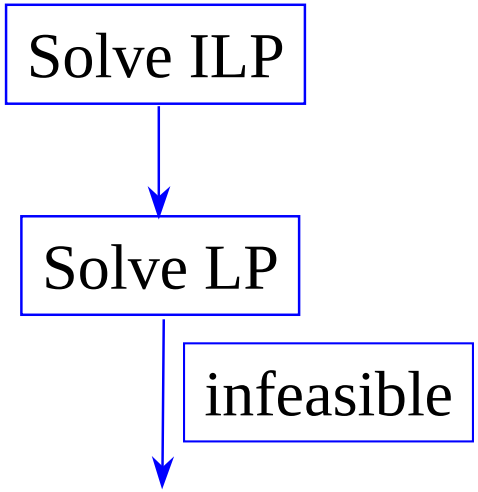


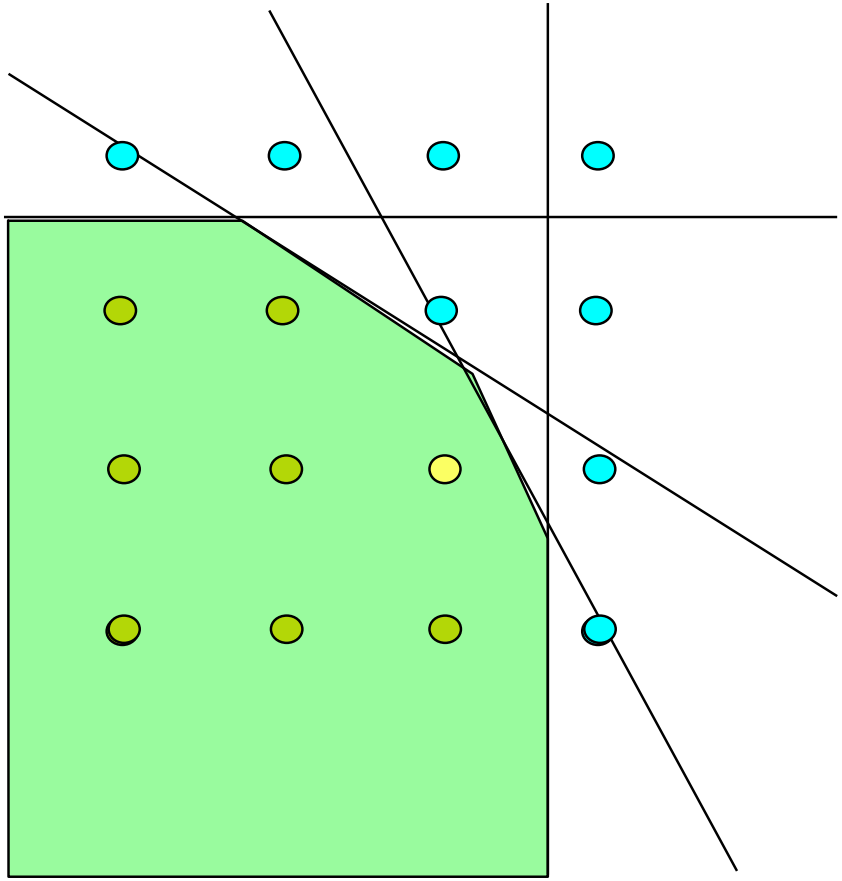
Branch & Cut



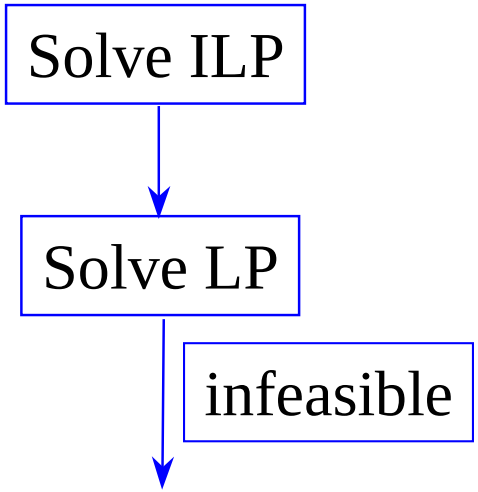


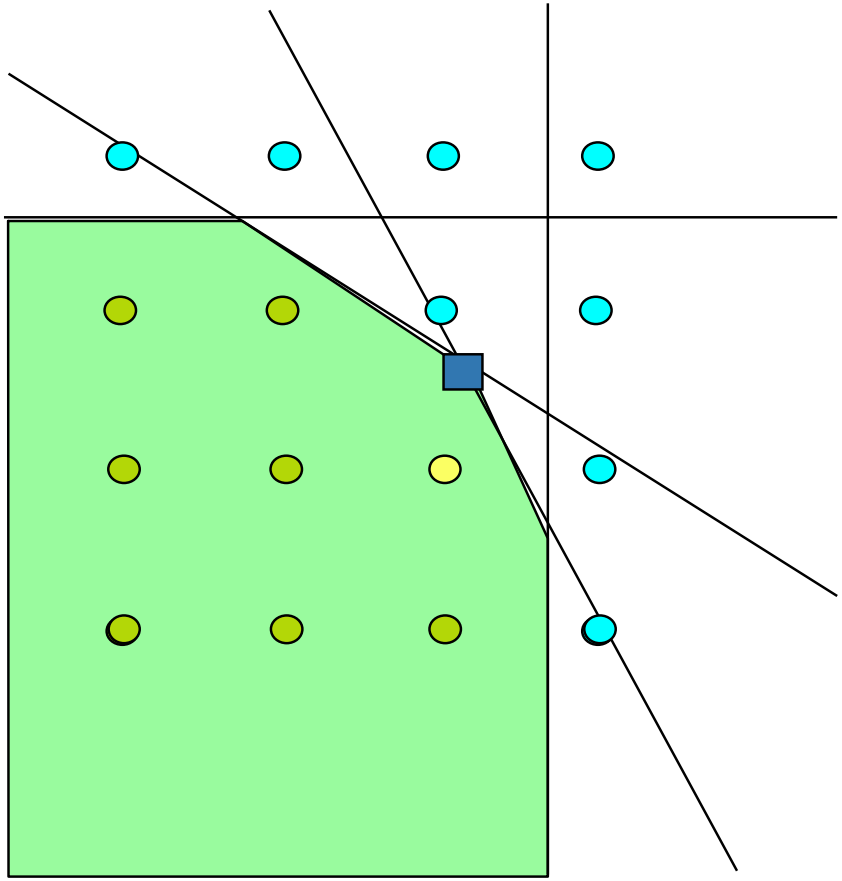
Branch & Cut



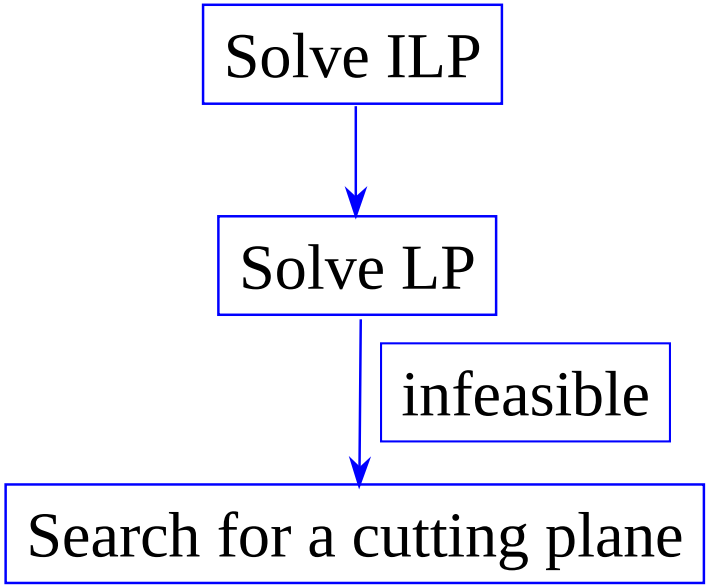


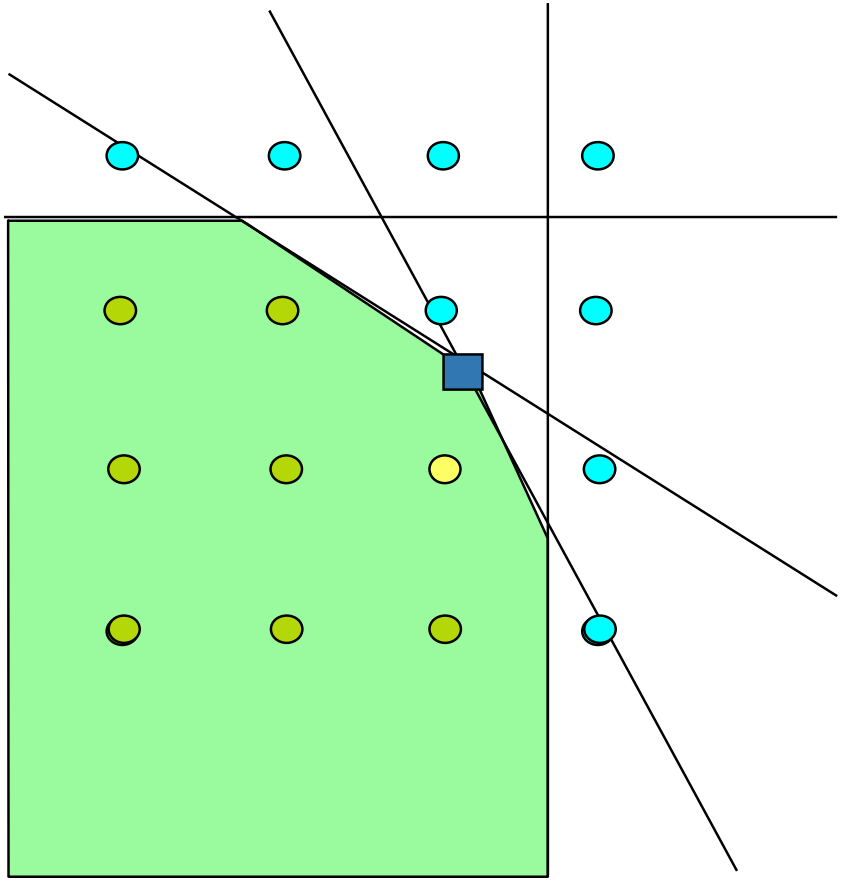
Branch & Cut



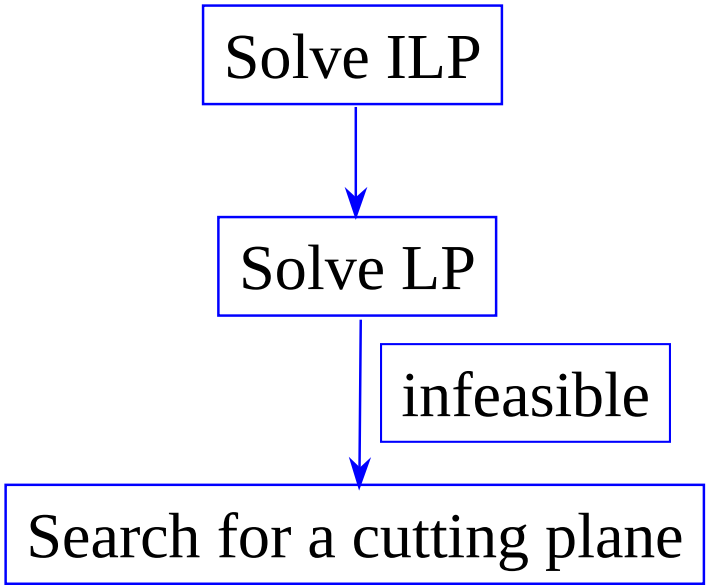


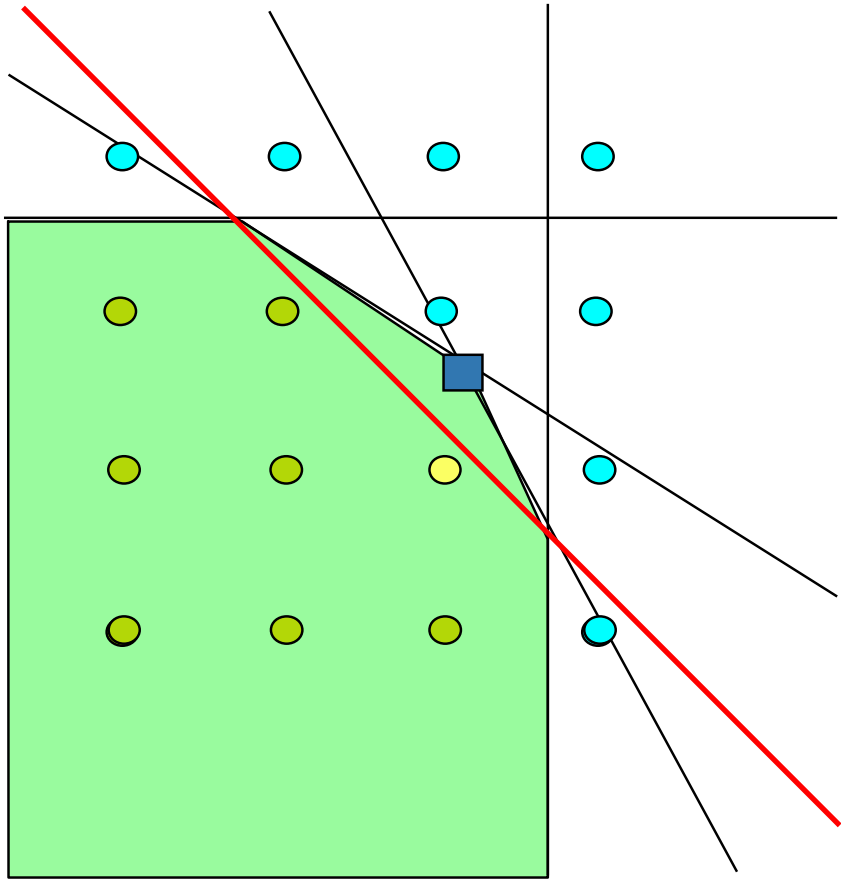
Branch & Cut



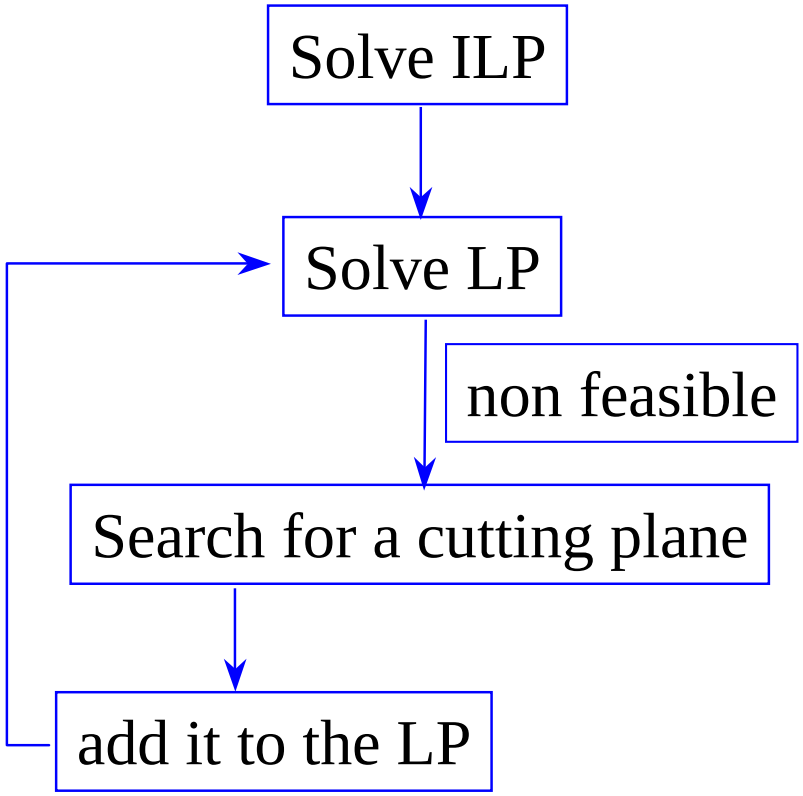


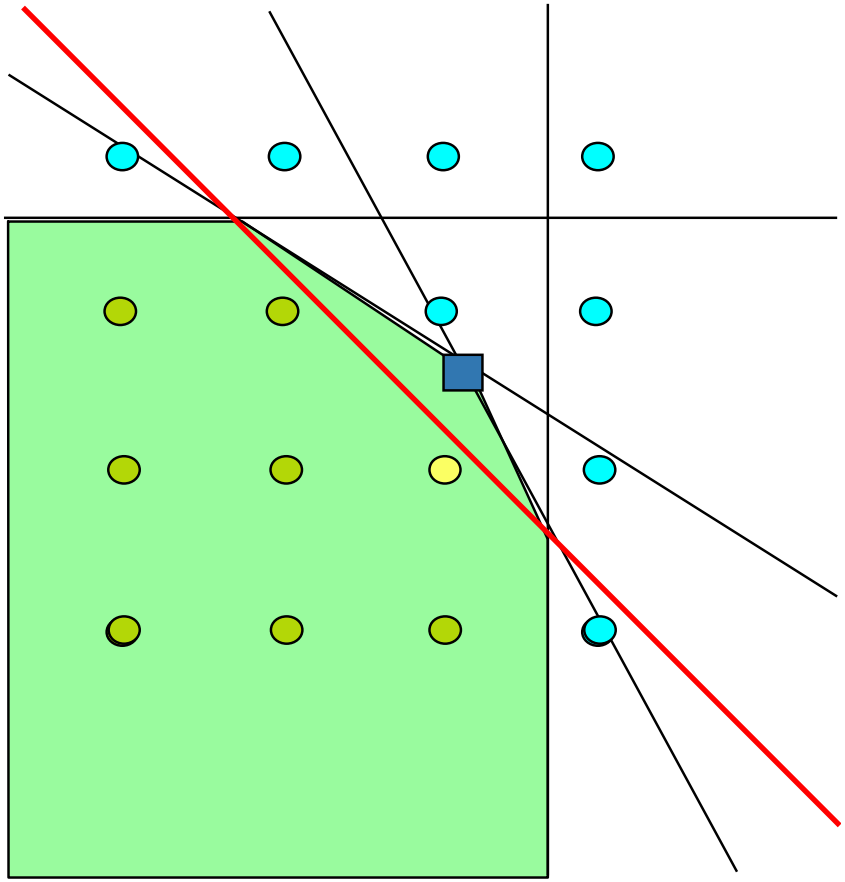
Branch & Cut



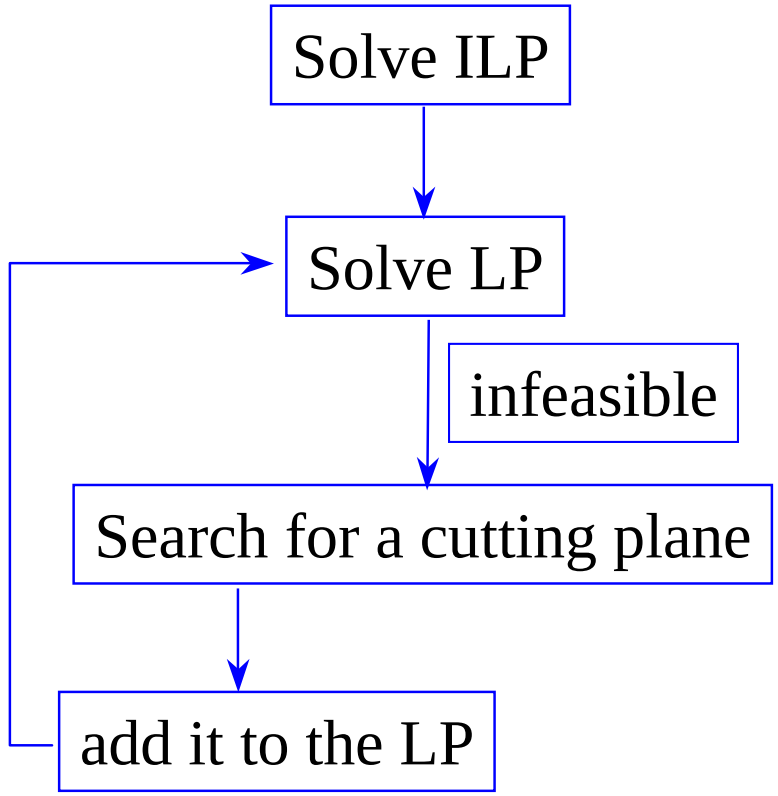


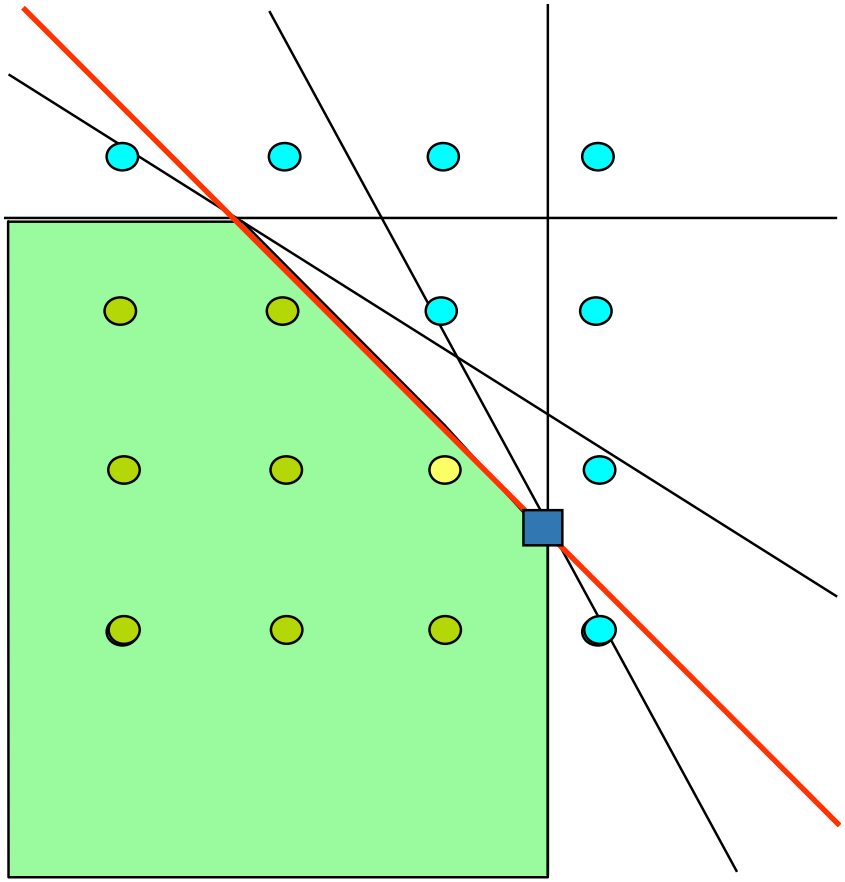
Branch & Cut


Solve ILP
Solve LP
non feasible
Search for a cutting plane
add it to the LP

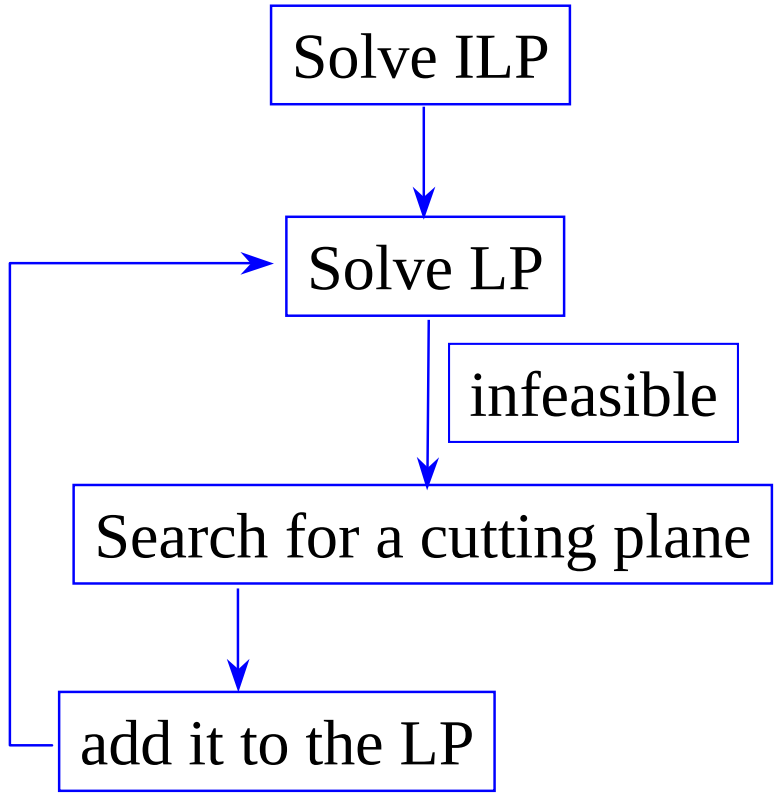


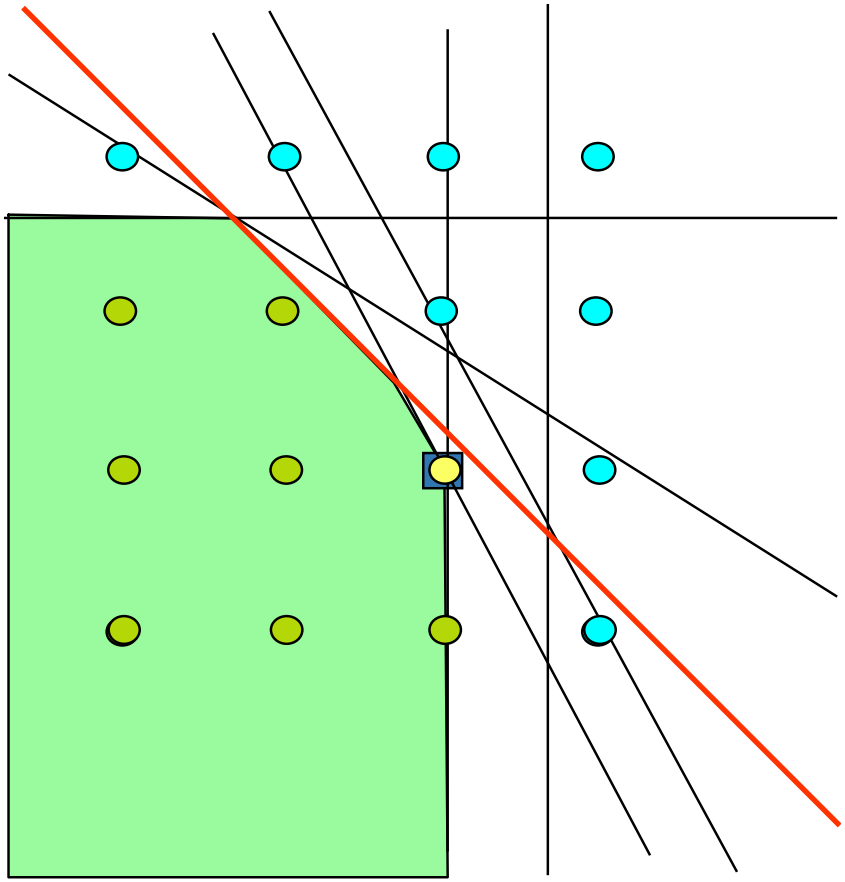
Branch & Cut


Solve ILP
Solve LP
infeasible
Search for a cutting plane
add it to the LP

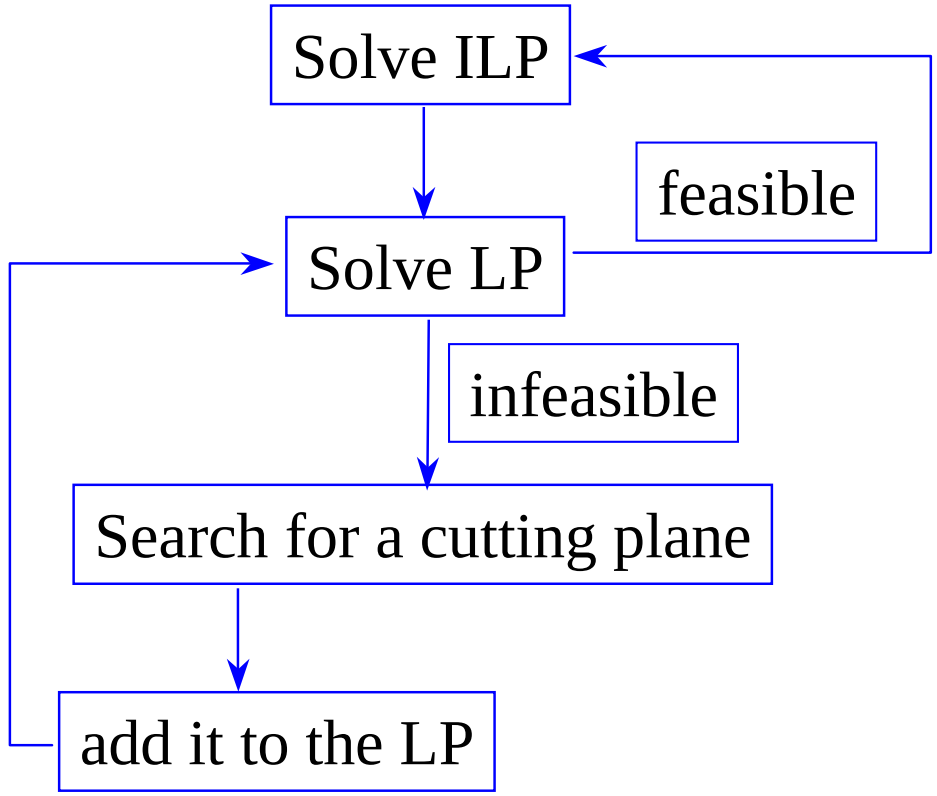


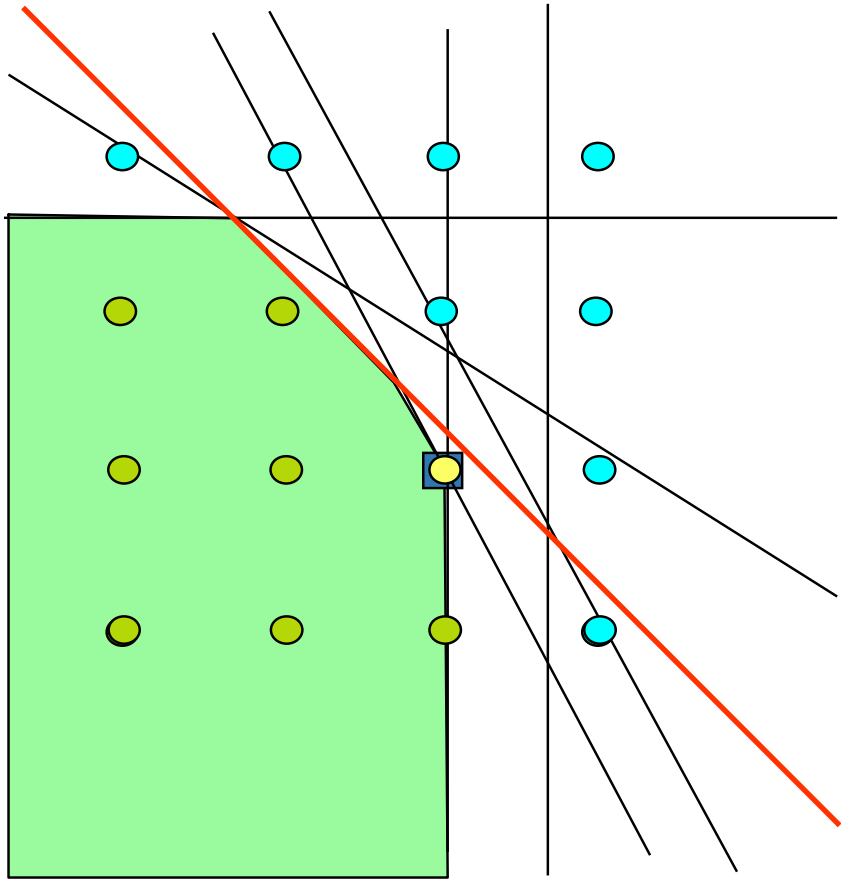
Branch & Cut


Solve ILP
Solve LP
infeasible
Search for a cutting plane
add it to the LP

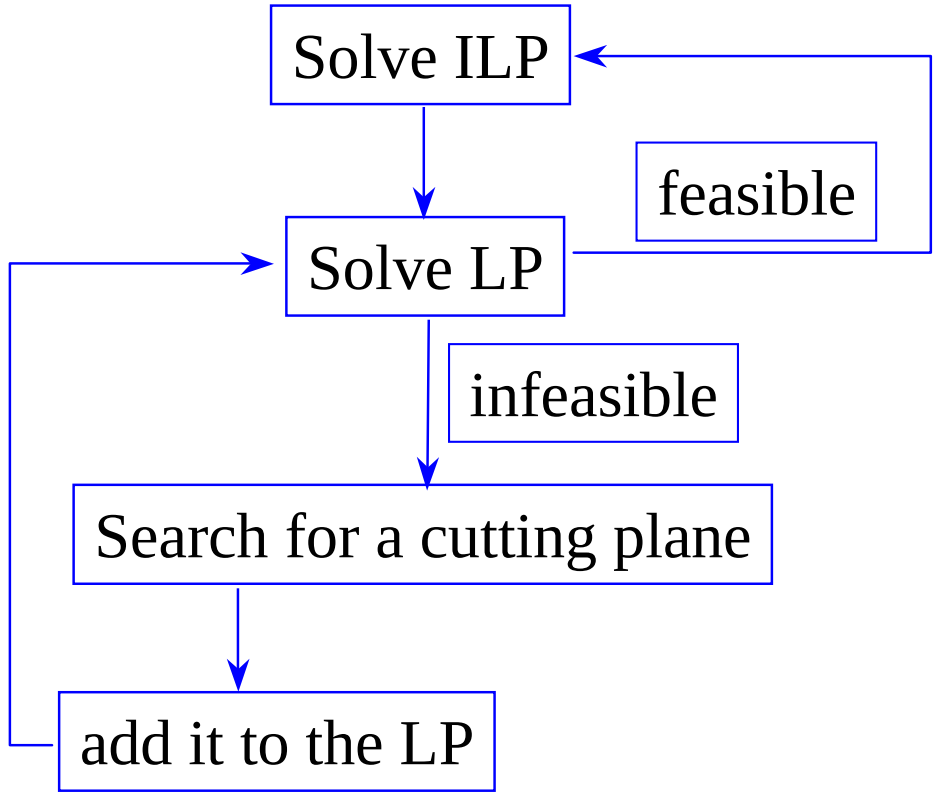


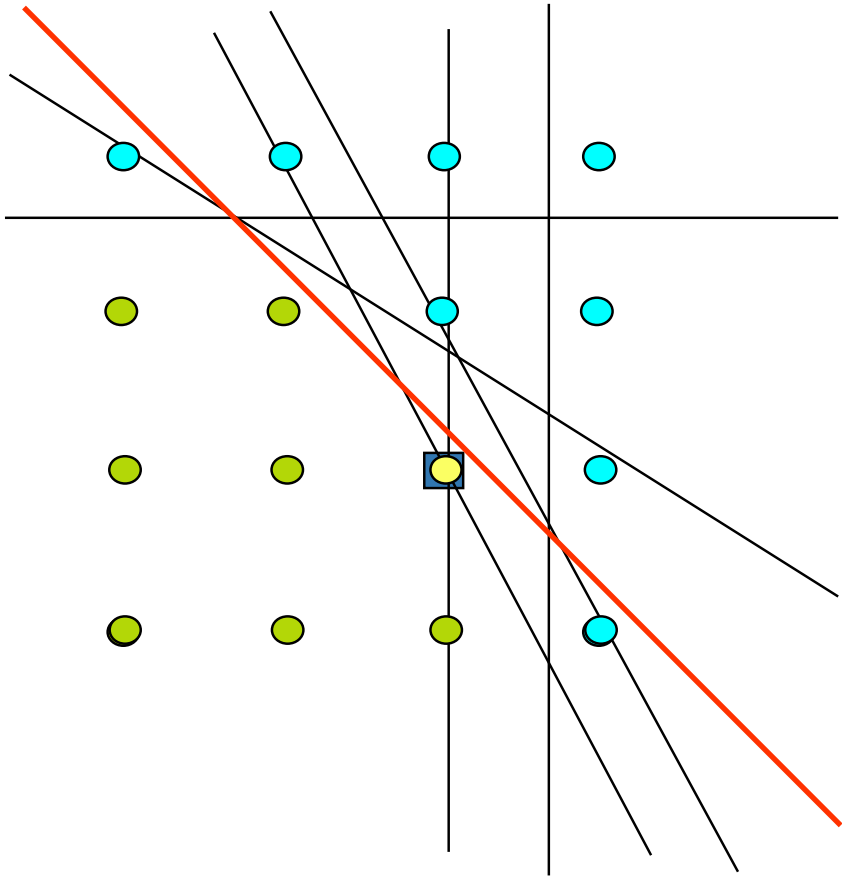
Branch & Cut


Solve ILP
feasible
Solve LP
infeasible
Search for a cutting plane
add it to the LP

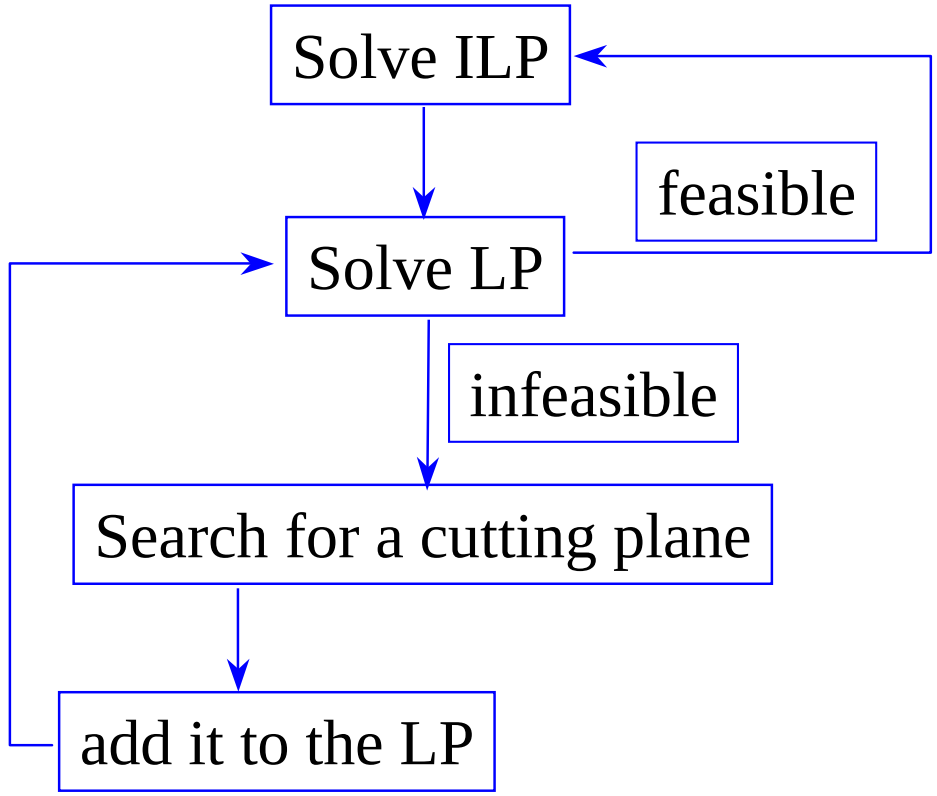


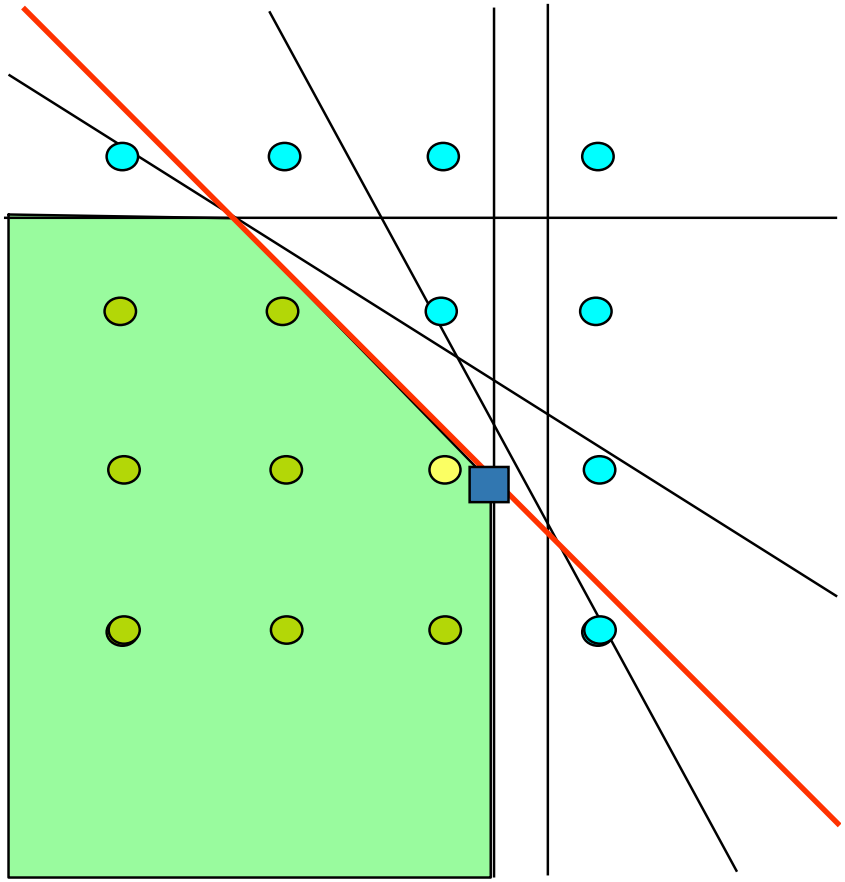
Branch & Cut


Solve ILP
feasible
Solve LP
infeasible
Search for a cutting plane
add it to the LP

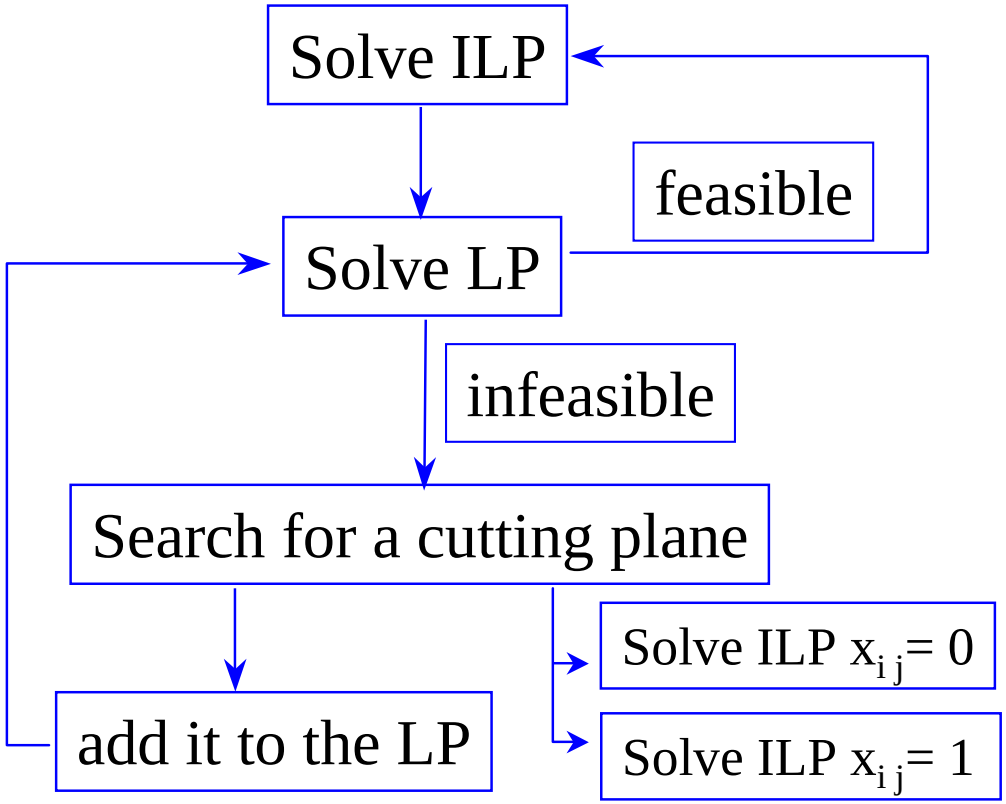


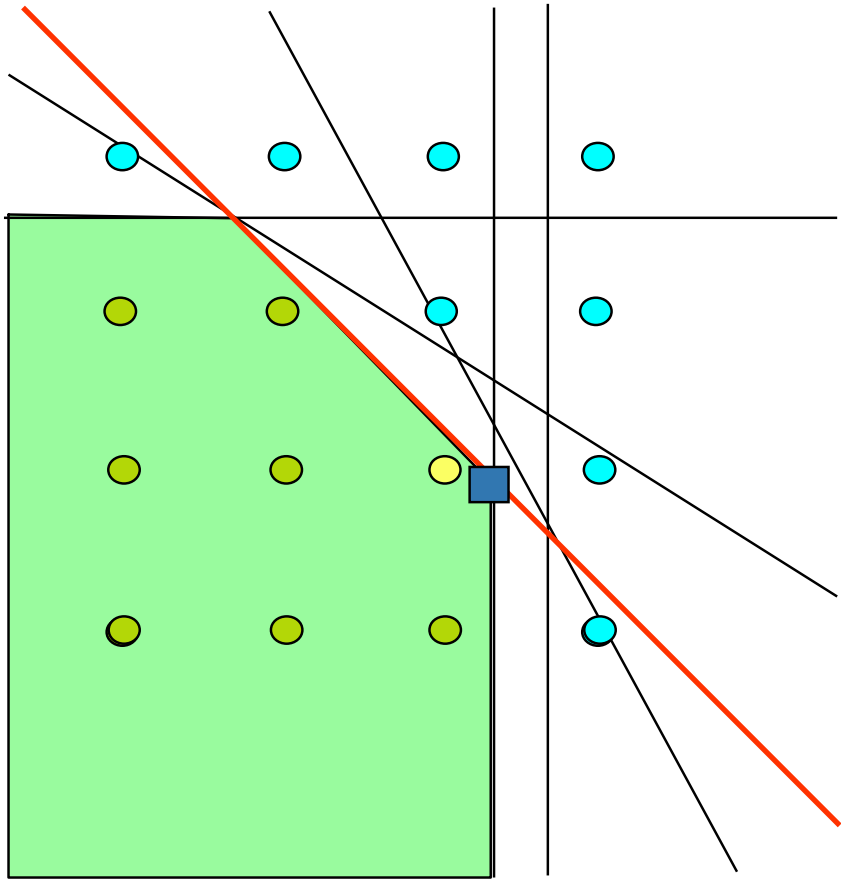
Branch & Cut


Solve ILP
feasible
Solve LP
infeasible
Search for a cutting plane
add it to the LP

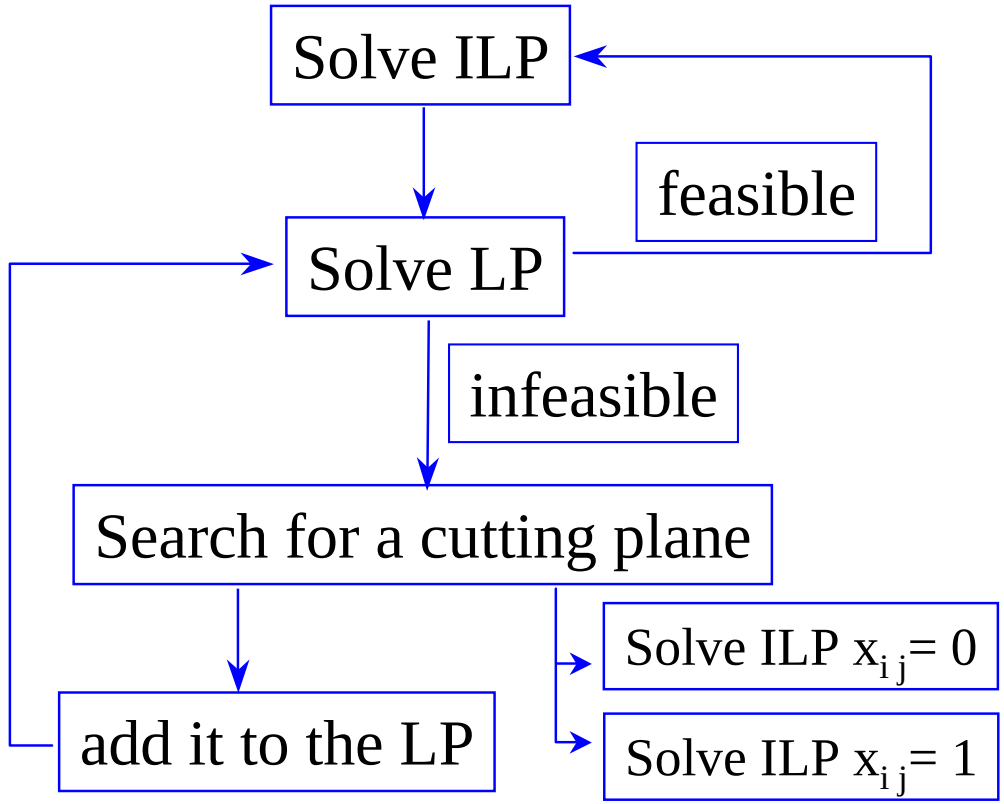


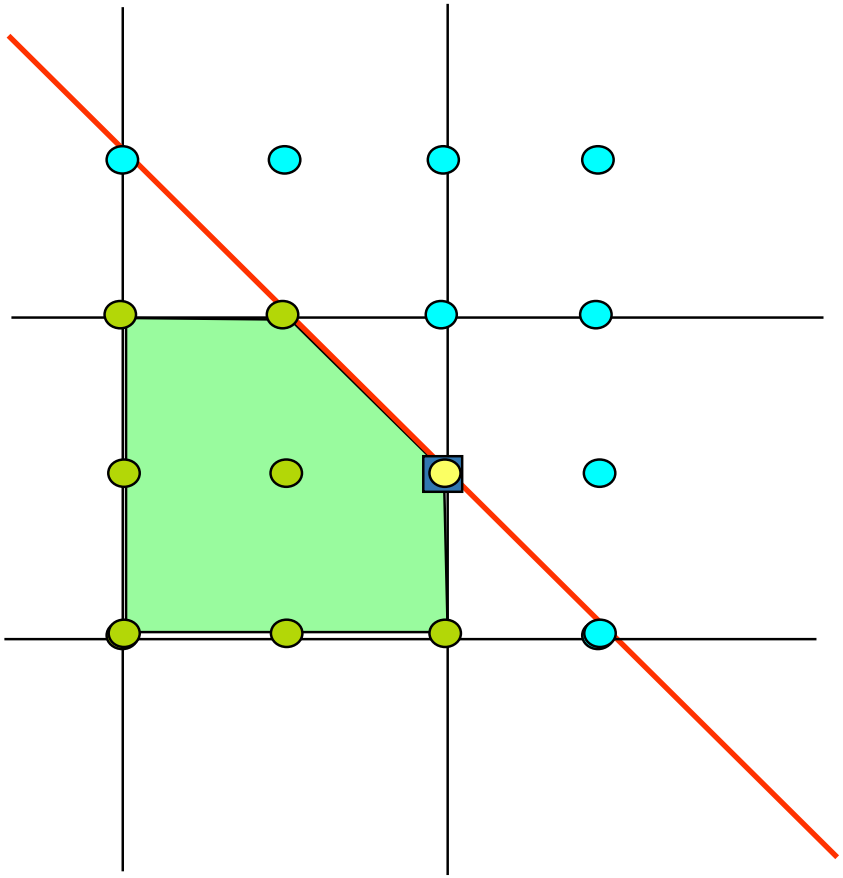
Branch & Cut


Solve ILP
feasible
Solve LP
infeasible
Search for a cutting plane
Solve ILP $x_{ij}= 0$
Solve ILP $x_{ij}= 1$
add it to the LP



Branch & Cut


Solve ILP
feasible
Solve LP
infeasible
Search for a cutting plane
Solve ILP $x_{ij}= 0$
add it to the LP
Solve ILP $x_{ij}= 1$



Energetic Evaluation

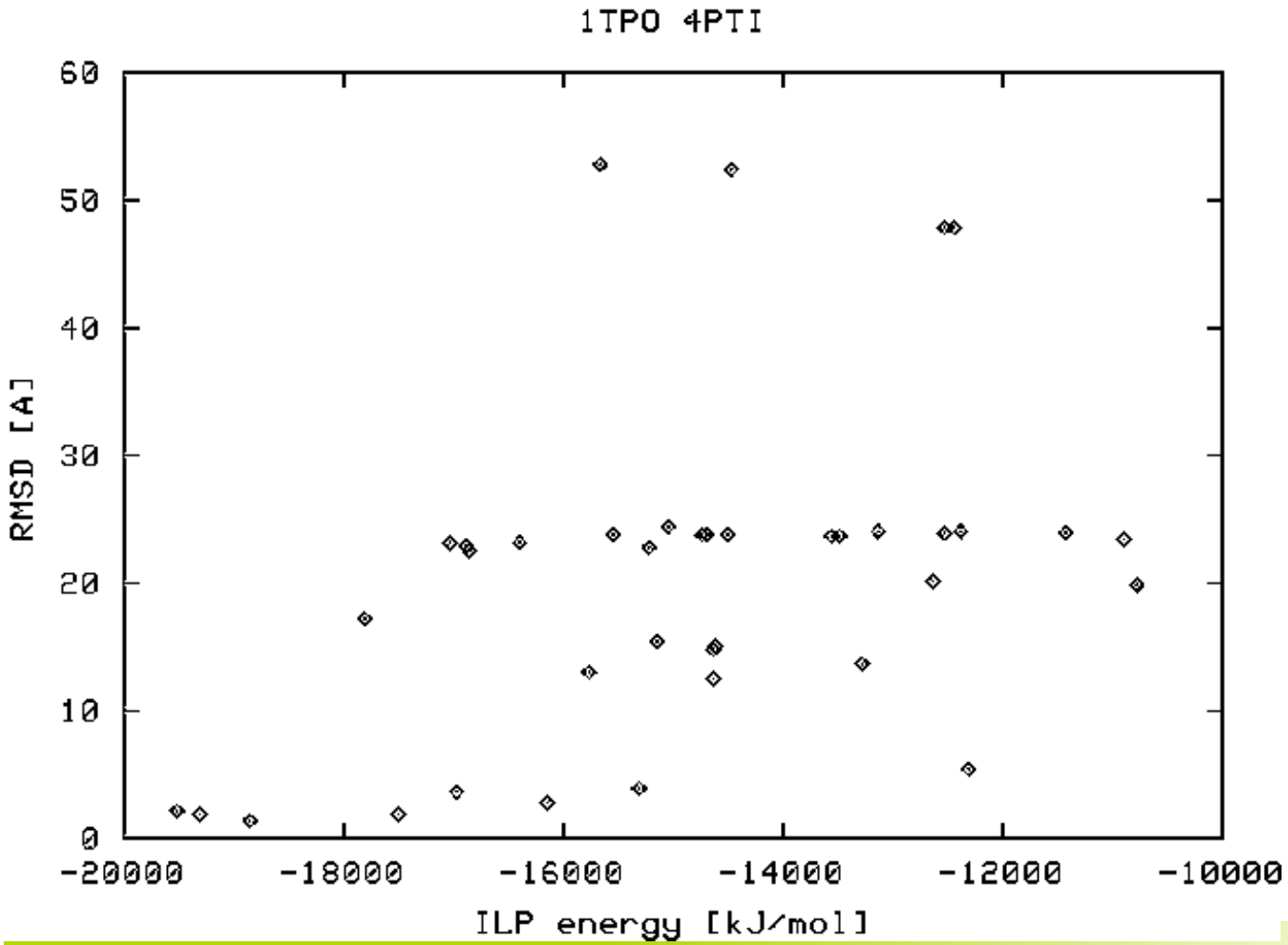
- Based on Jackson & Sternberg
- Energetic evaluation has to be extended:
  - inclusion of internal energies
- AMBER torsion energies:

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{cav} + \Delta G_{tors}$$

Implementation

- Molecular DS, rotamers, energies: BALL
- Branch-&-Cut: LEDA and ABACUS
- LP solver: CPLEX and SOPLEX

Docking Trypsin/BPTI - Results



Running Times

(UltraSparc II, 333 MHz, per candidate)

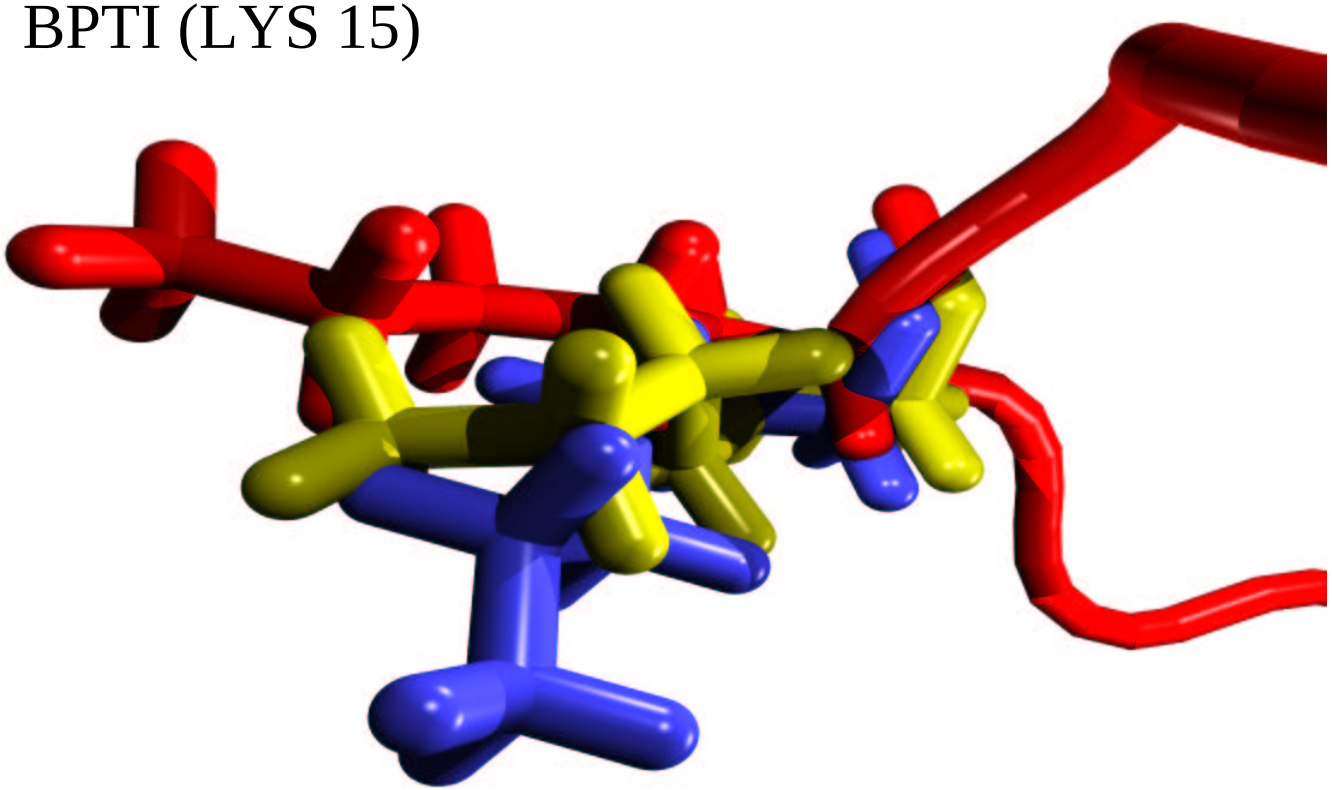
Stage	Avg. time [min]	
	Multi Greedy	ILP
Calculation of energies	30	
Combinatorial problem	3	14
Side-chain optimization	5	
Energetic evaluation	70	
Total	108	119

Results I

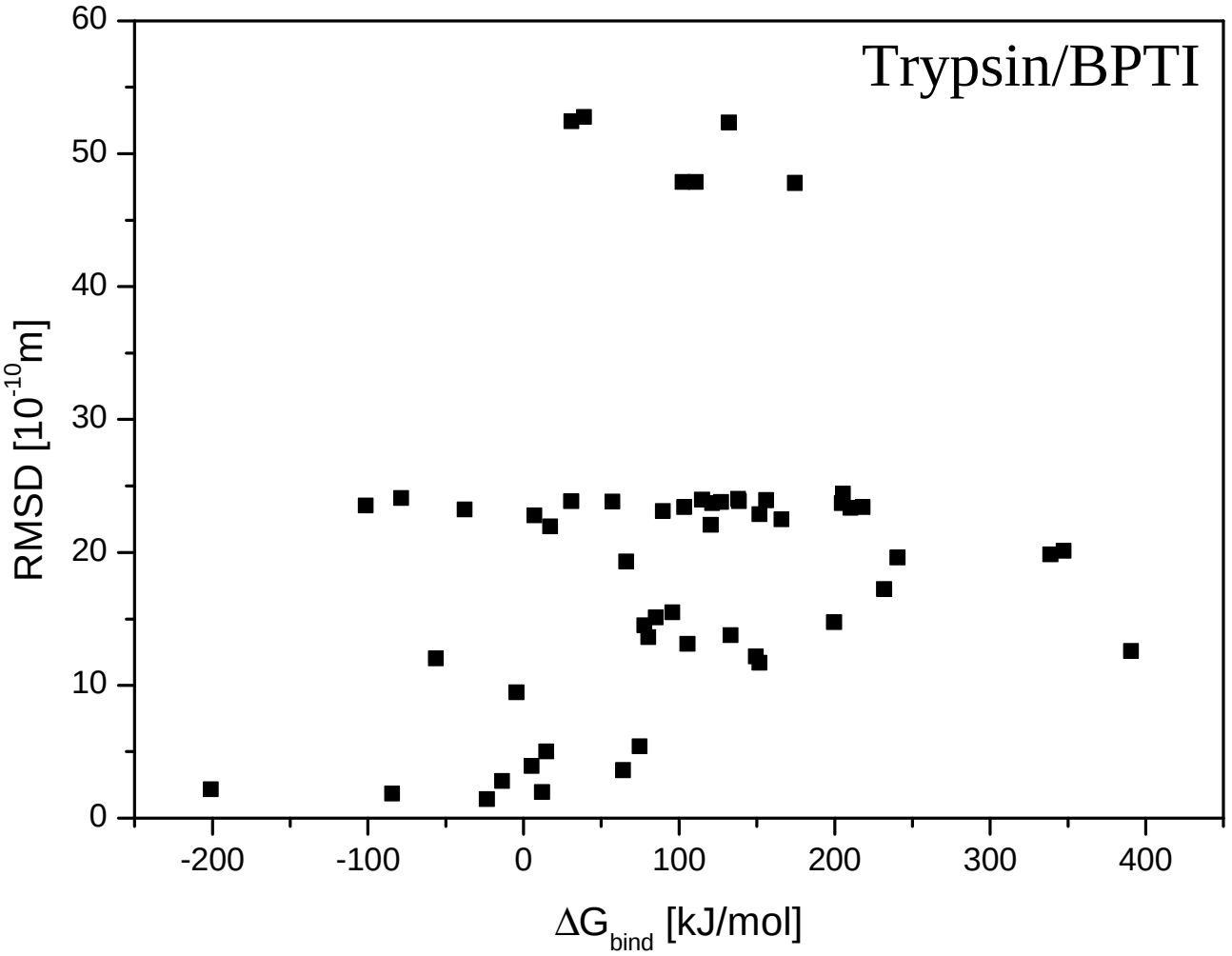
example	RMSD/Å
1TPO/4PTI	1.4
1SBC/2CI2	1.8
5CHA/2OVO	3.2

Results II

BPTI (LYS 15)



Results III


Trypsin/BPTI
RMSD [10^-10 m]
ΔG bind [kJ/mol]
60
50
40
30
20
10
0
-200
-100
0
100
200
300
400

Holm&Sander	1992	Monte-Carlo
Bruccoleri&Novotny	1992	Exhaustive Search
Desmet et al.	1992	Dead-End-Elimination
Totrov&Abagyan	1994a	Monte-Carlo
Totrov&Abagyan	1994b	Monte-Carlo
Laughton	1994	Local Homology Modeling
Leach	1994	A*-Algorithm
Weng et al.	1996	Exhaustive Search
Leach&Lemon	1998	A*-Algorithm

Integration of Experimental Data

- Main problem in Docking: Energy Function
- Avoid the energy calculation!
- Include experimental data
  - Containing geometric information
  - Derived from simple and inexpensive experiment

 Nuclear Magnetic Resonance (NMR)

Overview

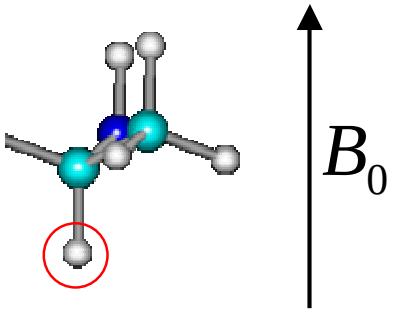
- What is NMR spectroscopy?
- Defining a NMR-based scoring function
  - Predicting chemical shifts
  - Synthesizing spectra
  - Comparing NMR spectra
- Results

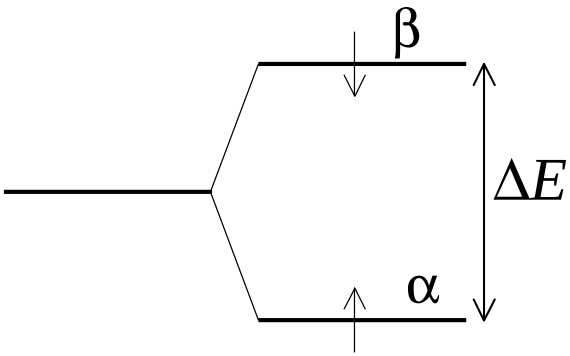
NMR - Basics

- ^{1}H nuclei possess *spin angular momentum*

NMR - Basics

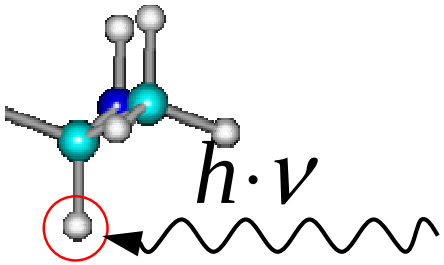
- ^{1}H nuclei possess *spin angular momentum*
- In an external magnetic field B_0, each nucleus can assume one of two spin states: α or β

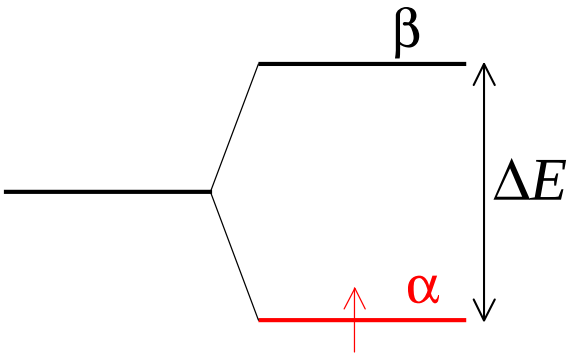




NMR - Basics

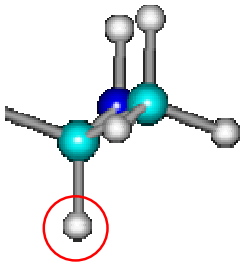
- ^{1}H nuclei possess *spin angular momentum*
- In an external magnetic field B_0, each nucleus can assume one of two spin states: α or β
- Supplying energy ΔE enables transition

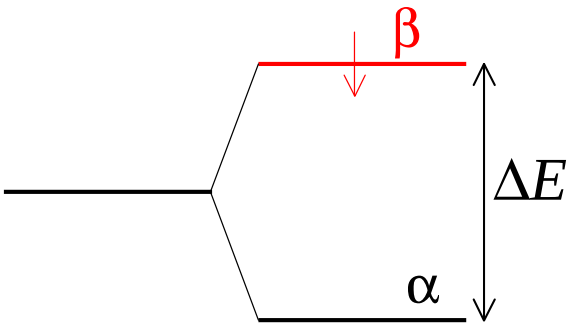


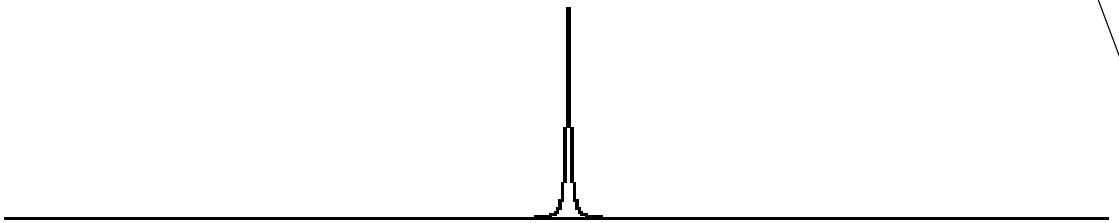


NMR - Basics

- ^{1}H nuclei possess *spin angular momentum*
- In an external magnetic field B_0, each nucleus can assume one of two spin states: α or β
- Supplying energy ΔE enables transition

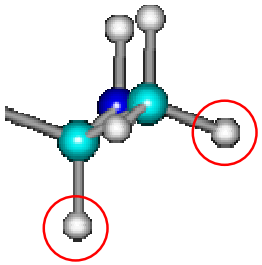


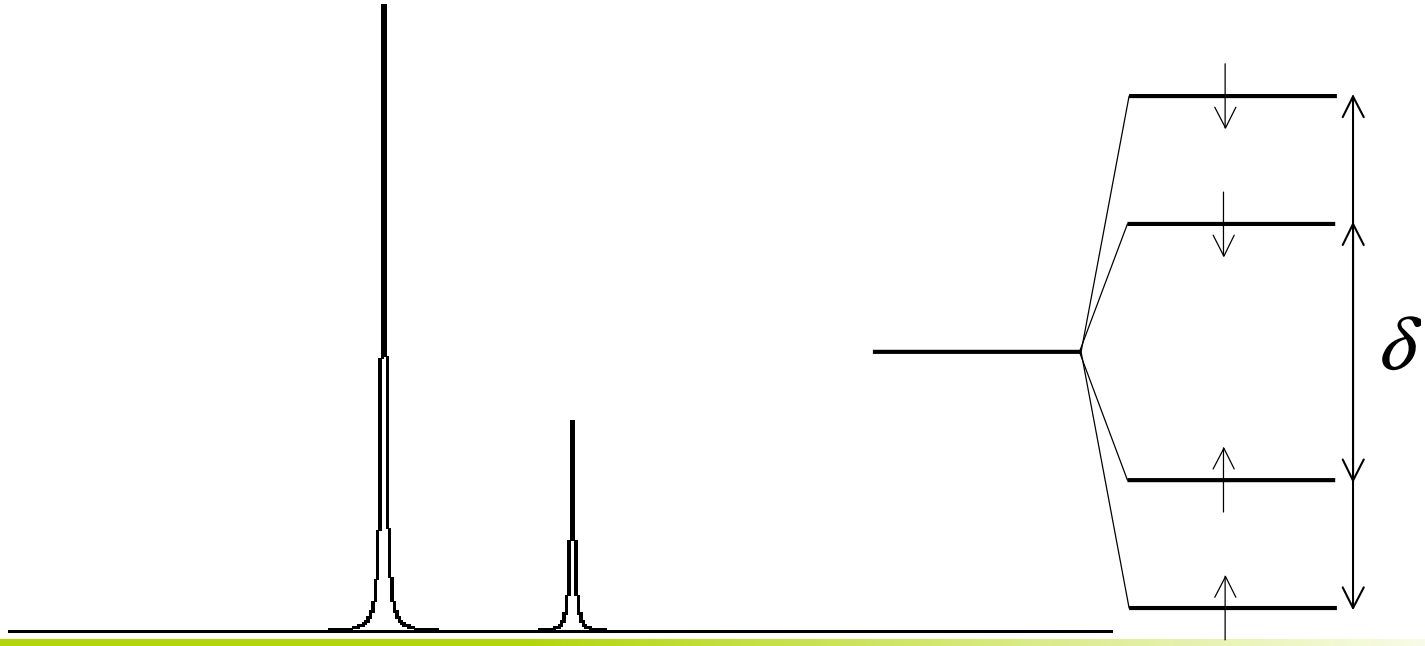




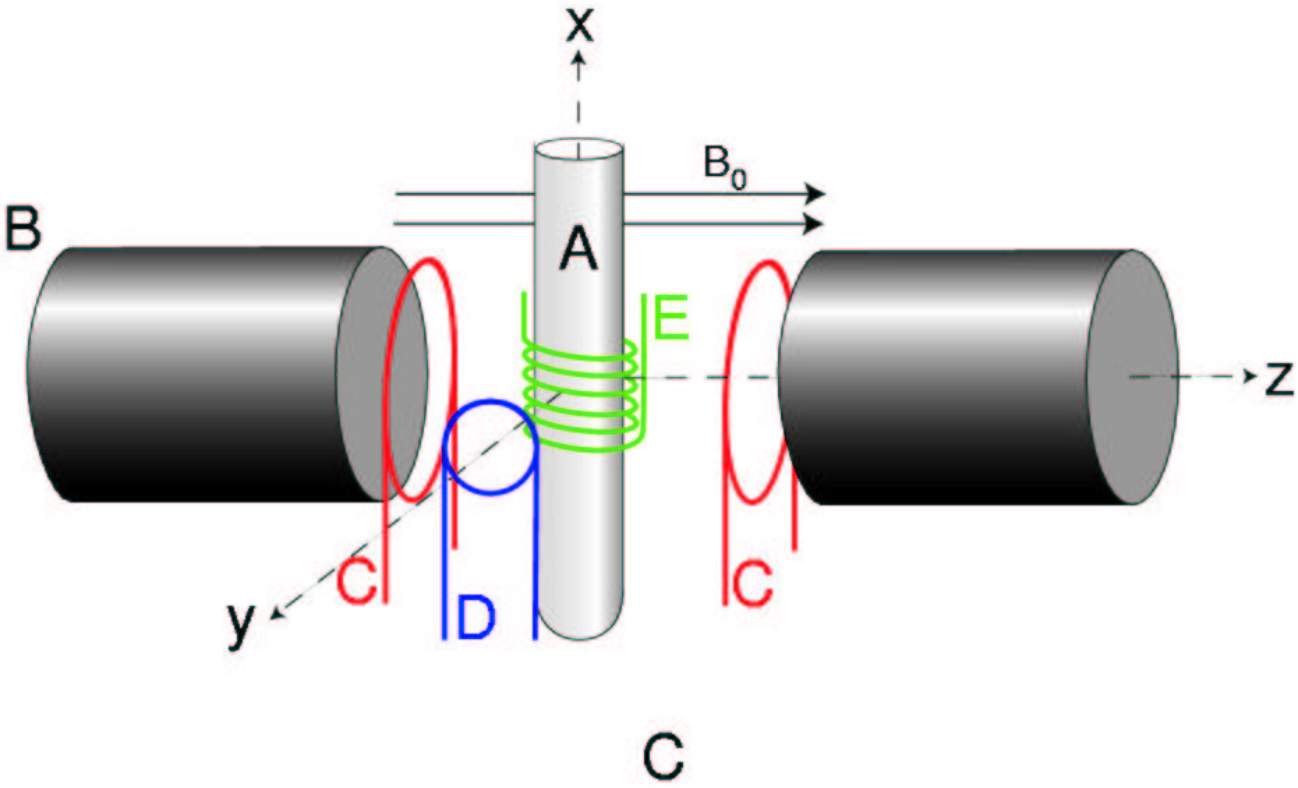
NMR - Basics

- ΔE depends on
  - magnitude of external field
  - electronic structure surrounding the nucleus

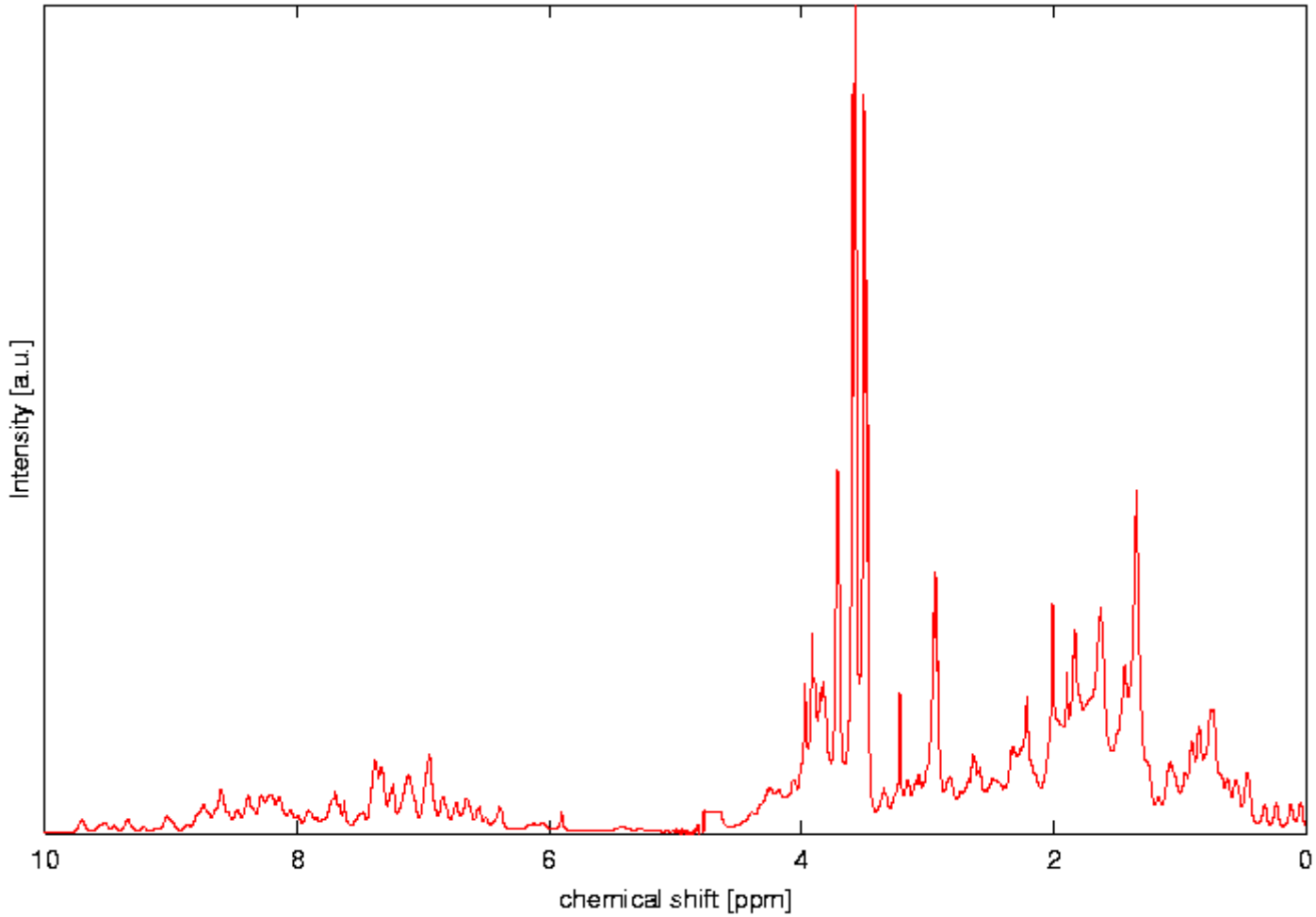




NMR: The Hardware



1D ^{1}H-NMR spectra

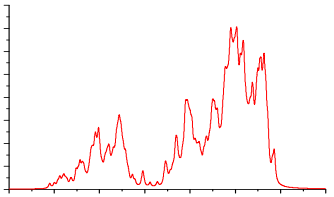


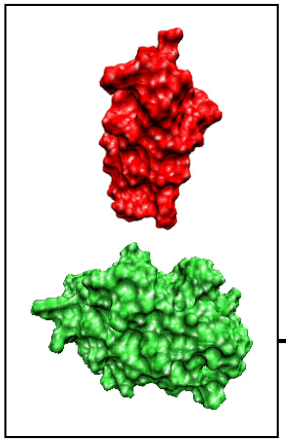
Structural Information in Spectra

- Chemical shift in proteins depends on
  - Topology (chemical environment)
  - Geometry (conformation)
- Certain experiments can
  - identify which proton causes which peak (assignment)
  - yield distance information (e.g., NOE constraints)

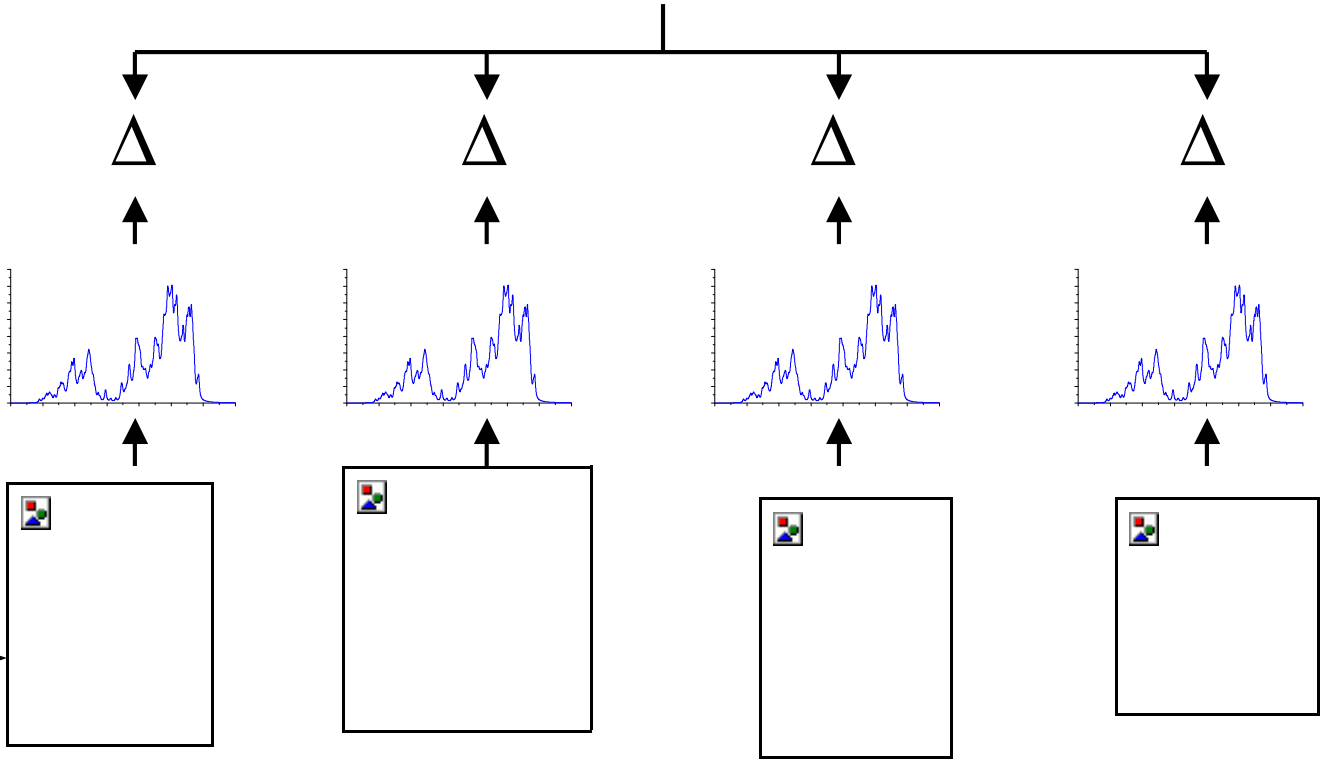
NMR-based Protein Docking

Experiment



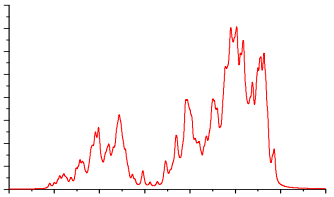


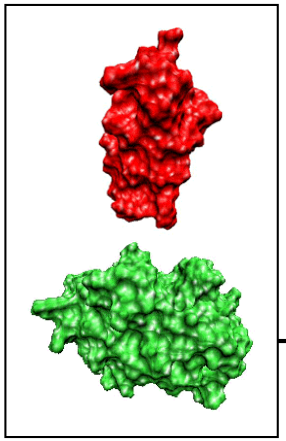
Docking



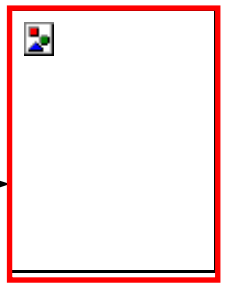
NMR-based Protein Docking

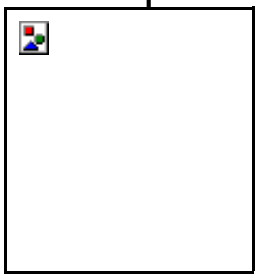
Experiment

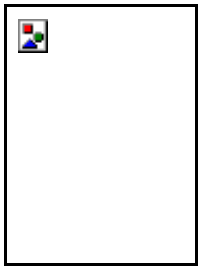


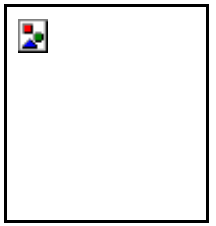


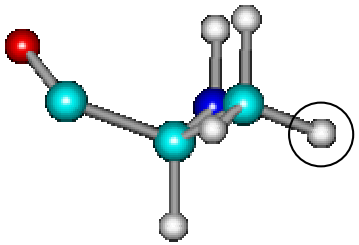
Docking





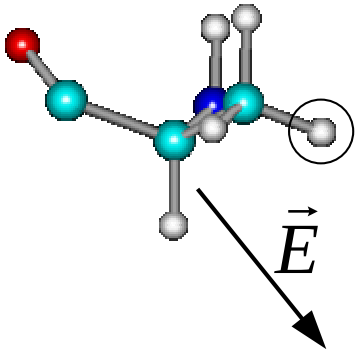


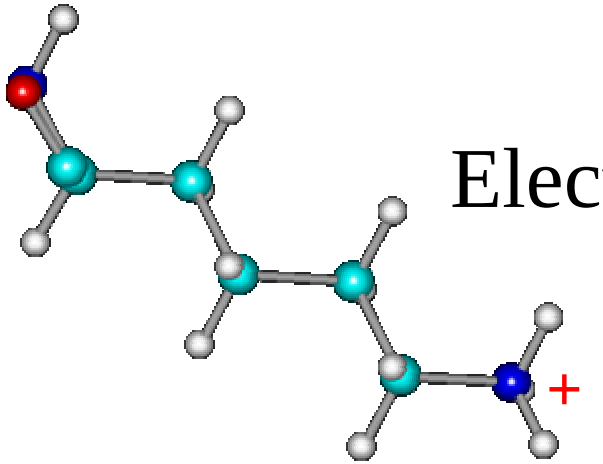




Random coil shift

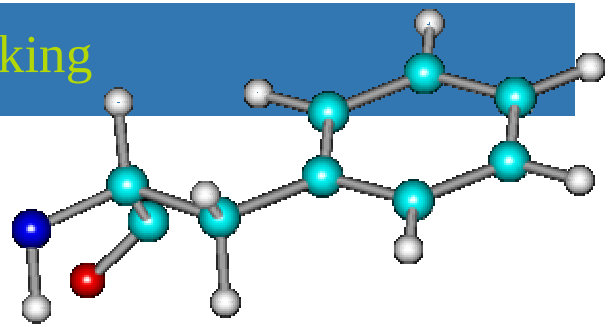
$$\delta = \delta_{RC}$$

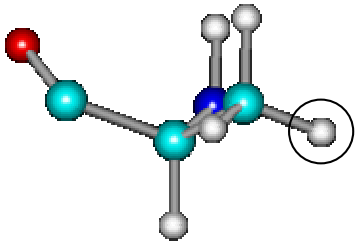


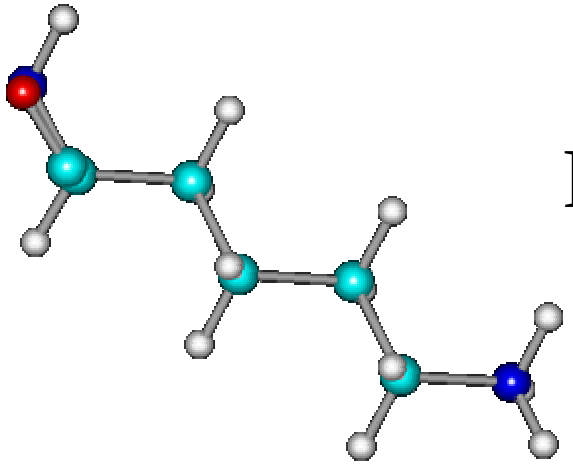


Electrostatic field

$$\delta = \delta_{RC} + \delta_{ES}$$

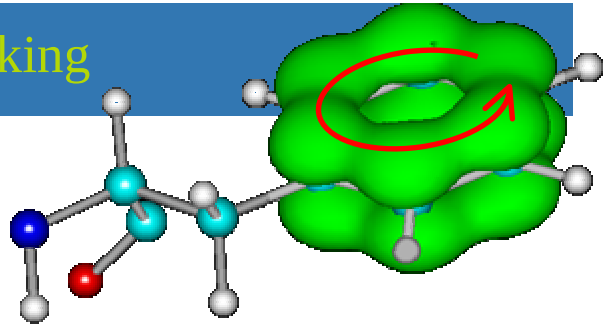


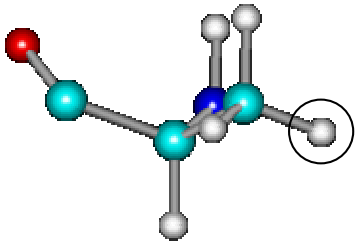




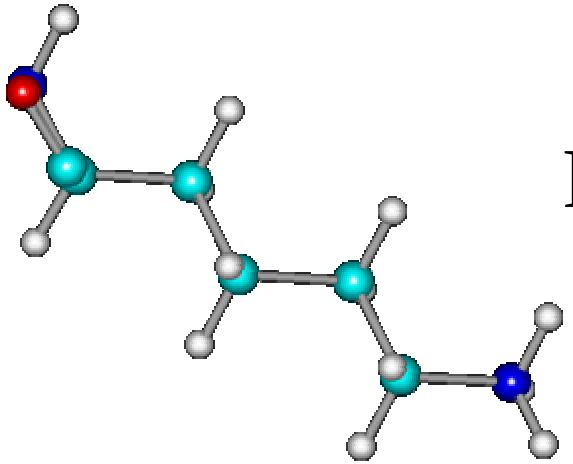
Ring current effect

$$\delta = \delta_{RC} + \delta_{ES}$$



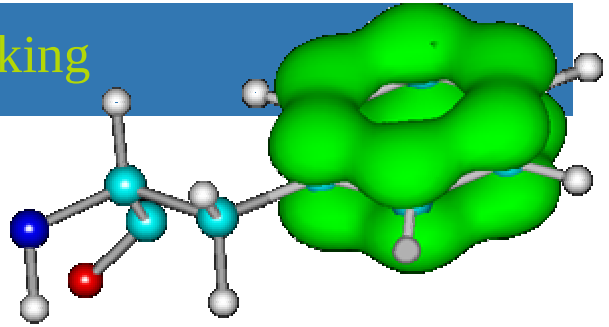


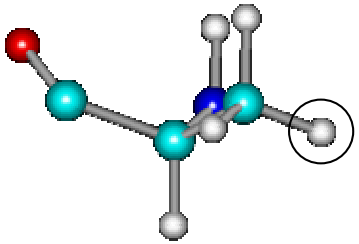
$\vec{B}_{RS}$

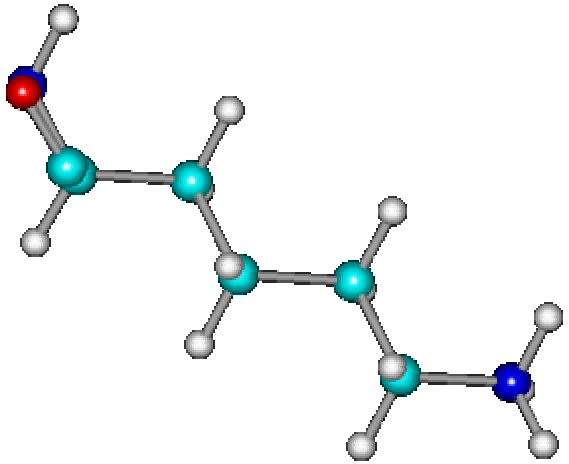


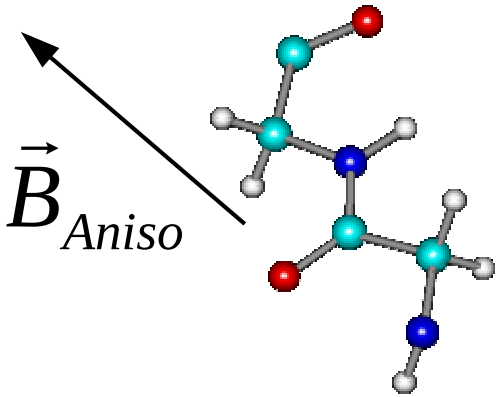
Ring current effect

$$\delta = \delta_{RC} + \delta_{ES} + \delta_{RS}$$









Magnetic Anisotropy

$$\delta = \delta_{RC} + \delta_{ES} + \delta_{RS} + \delta_{Aniso}$$

Spectrum Synthesis/Comparison

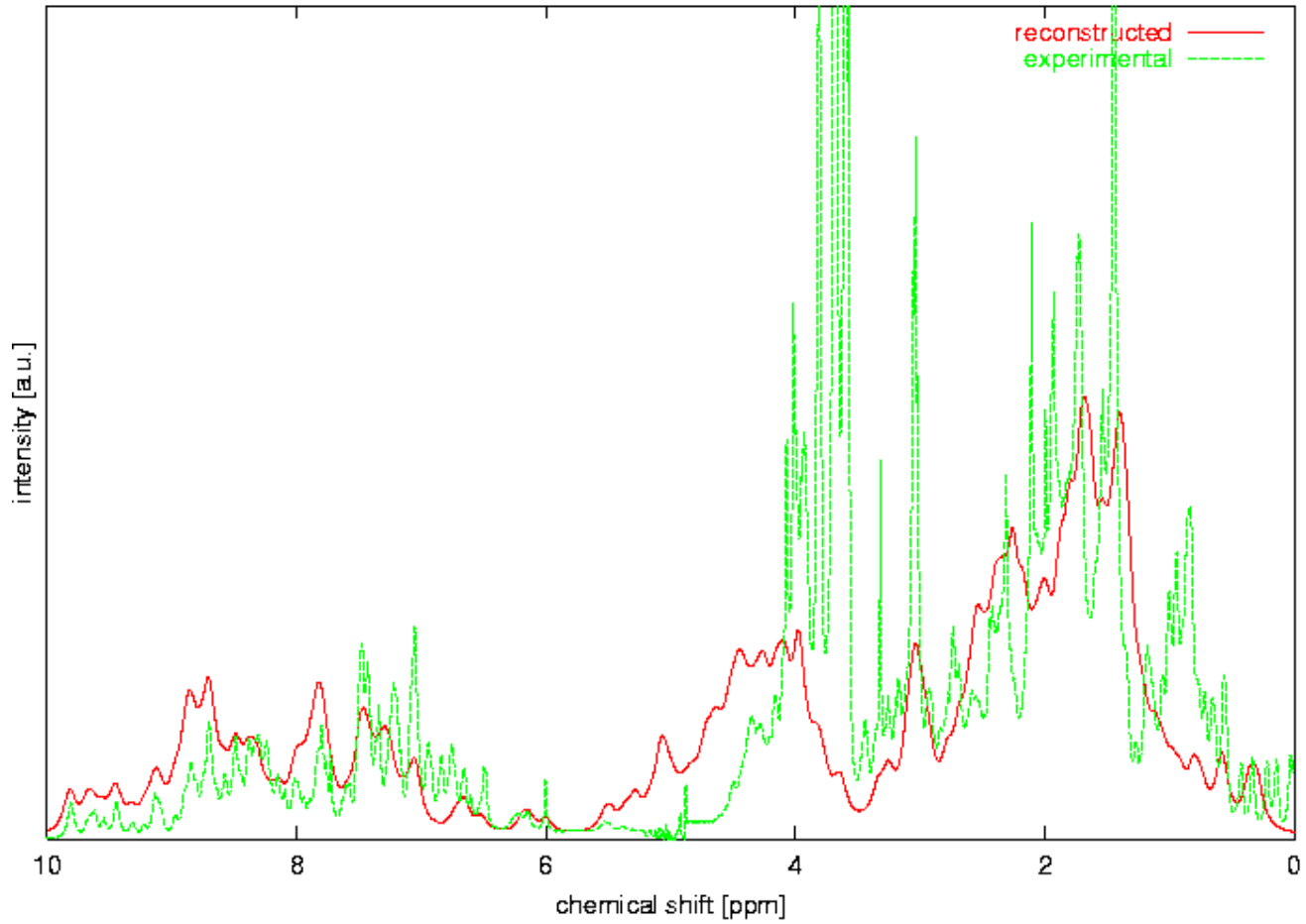
- Removal of rapidly exchanging protons
- Lorentzian peaks of equal width

$$S(\delta) = \sum_i \frac{1}{1 + \frac{(\delta - \delta_i)^2}{W}}$$

- Spectrum comparison via *difference area* Δ

$$\Delta(S_{calc.} - S_{exp.}) = \int | S_{calc.}(\delta) - S_{exp.}(\delta) | \, d\delta$$

Synthesized Spectra



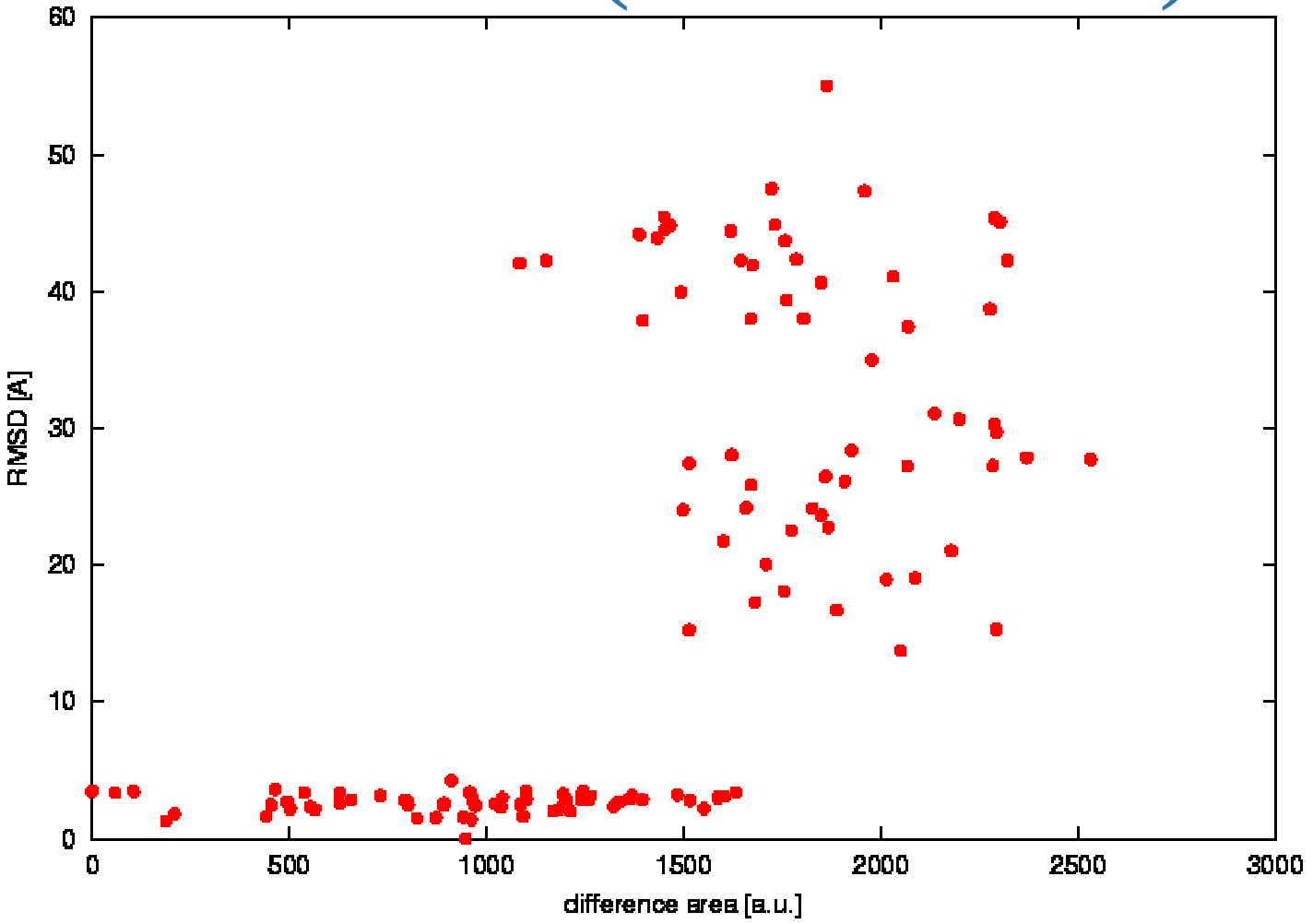
Results – Test Set

- Bound complex structures (PDB)
  - 1DT7 A/B
  - 1DT7 A/X
  - 1CFF
  - 1CKK
- NMR assignments (BMRB)
  - Reconstruction of exp. spectra

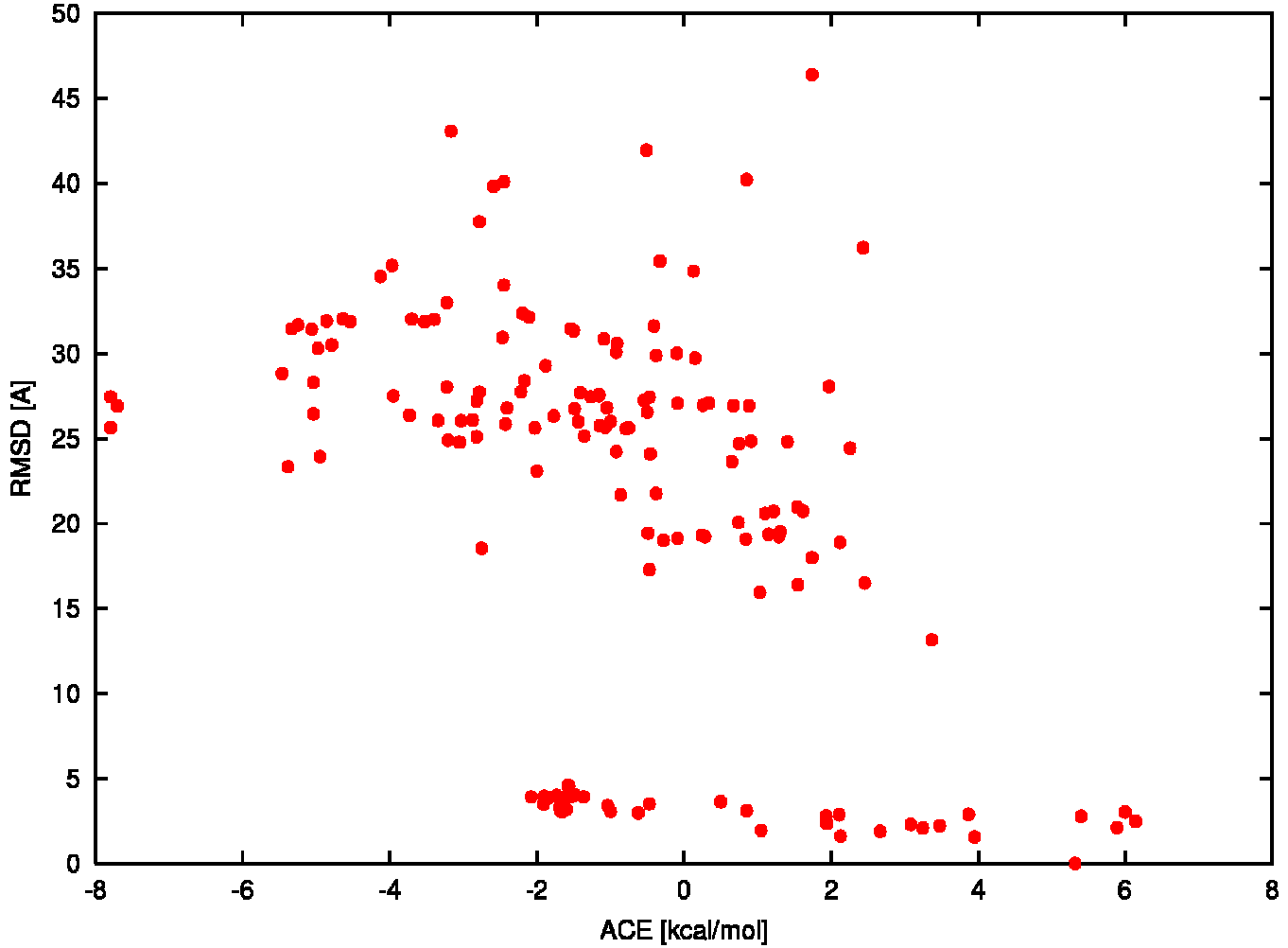
Results - Overview

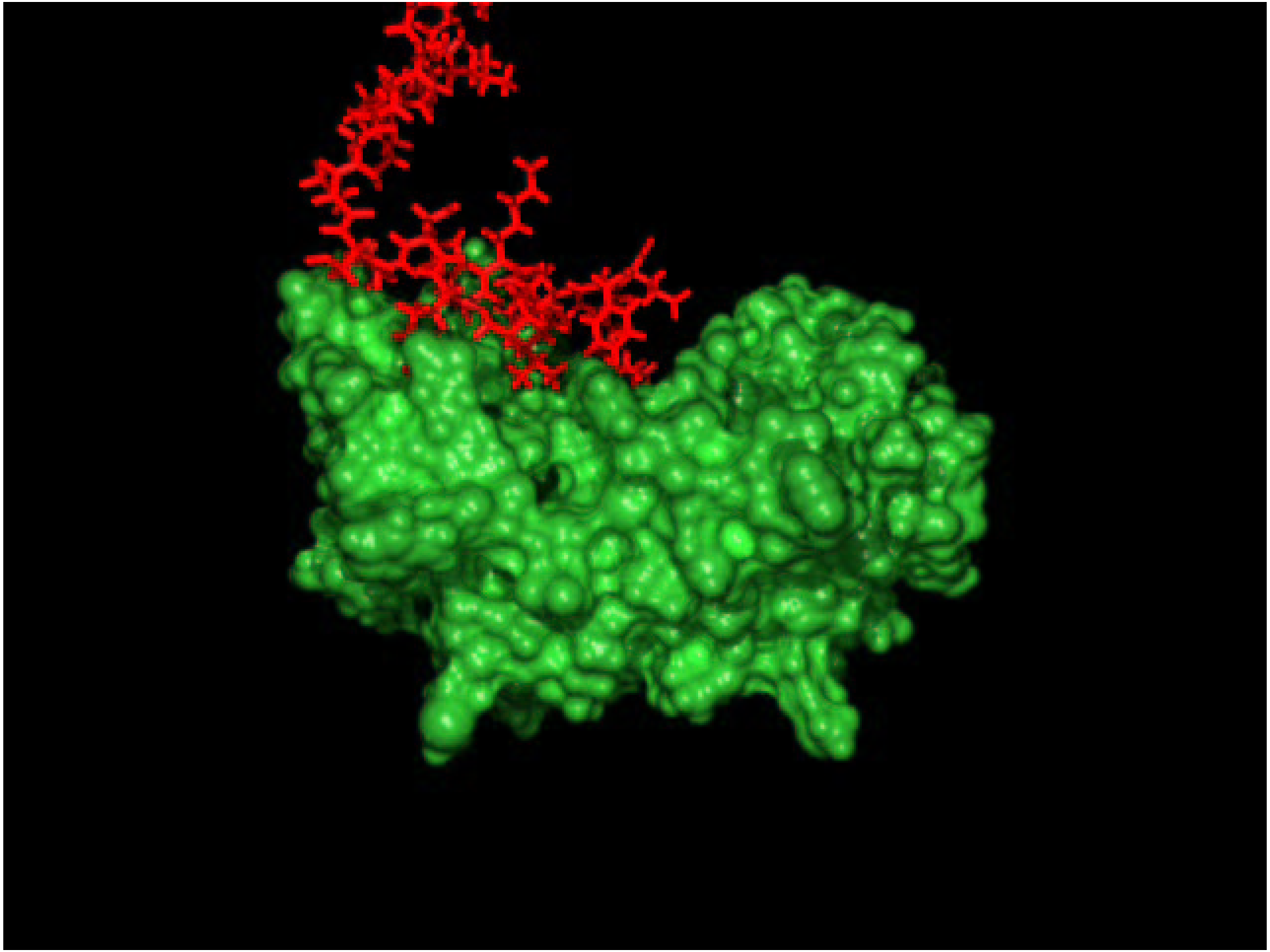
Complex	# false positive among top 10	Rank of 1st true positive
1DT7 A/B	2	1
1DT7 A/X	0	1
1CFF	4	2
1CKK	2	1

1DT7 A/X (NMR-based)



1DT7 A/X (ACE)





Open Problems

- Improving the shift model
  - H-bonds
  - Solvent effects
- Modeling peak width, couplings?
- Blind predictions of protein complexes
- Validation with larger data set and exp. data

Summary

- Protein Docking is an interesting problem!
- Key difficulties
  - Energetic evaluation
  - Protein flexibility
- Integration of experimental evidence improves scoring

Free Energy Functions

- Contact-based
  - Atomic contact potential
  - Residue contact potentials
- Molecular Mechanics
- Surface-based potentials
- …..

Contributions to Free Energy I

- Hydrogen bonds
- Salt bridges
- Entropic contributions
  - Solvent effects
  - Loss of side-chain entropy
  - Loss of DOF on binding

Contributions to Free Energy II

- Electrostatic interactions
  - Coulomb-based
  - Continuum models (solvent effects)
- Hydrophobic interactions
- Van-der-Waals interactions
- ….

Energetic Evaluation

Jackson, Sternberg (1994):

- continuum electrostatics
  - solution of the Poisson-Boltzmann equation
  - includes electrostatic solvent effects
- cavitation free energy
  - measure for the hydrophobic interaction

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{cav} + \Delta G_{conf} + \Delta G_{vdW}$$

Energetic Evaluation

Jackson, Sternberg (1994):

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{cav}$$

Energetic Evaluation

Jackson, Sternberg (1994):

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{cav}$$

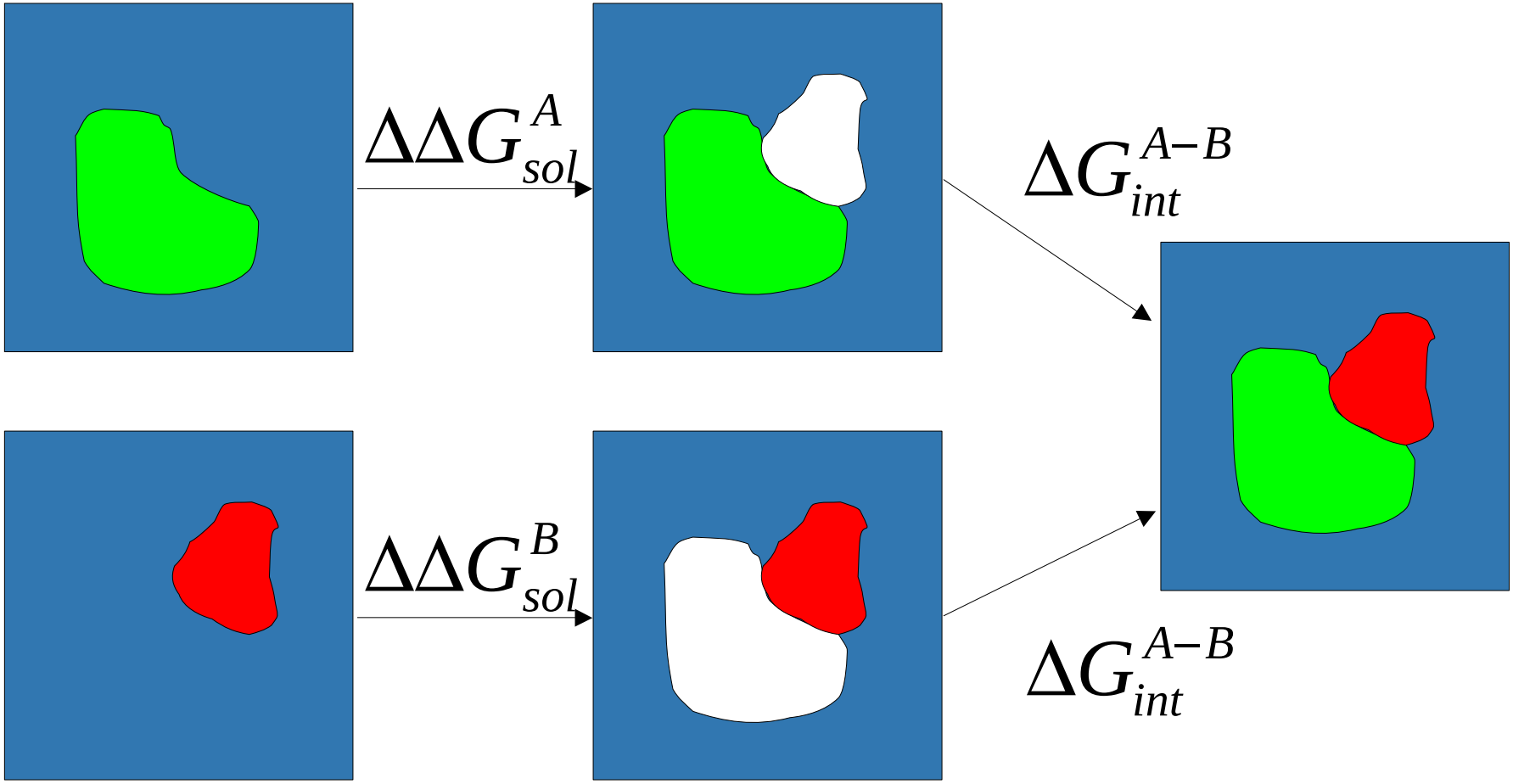
$$\Delta G_{ele} = \Delta\Delta G_{sol}^{A} + \Delta\Delta G_{sol}^{B} + \Delta G_{int}^{A-B}$$

Poisson Boltzmann Equation

$$\nabla\left(\varepsilon(\vec{r})\nabla\varphi(\vec{r})\right) - \bar{\kappa}^2(\vec{r})\,\sinh\left[\frac{e^0\varphi(\vec{r})}{k\,T}\right] = -\frac{\rho(\vec{r})}{\varepsilon_0}$$

- Spatially dependent dielectric constant ε(r)
  - ~2-4 inside protein
  - 78 outside (water)
- Charge distribution ρ potential ϕ
- Solve differential equation on grid
- Yields total electrostatic free energy ΔG_{ES}

Continuum Electrostatics



Energetic Evaluation

Jackson, Sternberg (1994):

$$\Delta G_{bind} = \Delta G_{ele} + \Delta G_{cav}$$

$$\Delta G_{ele} = \Delta\Delta G_{sol}^{A} + \Delta\Delta G_{sol}^{B} + \Delta G_{int}^{A-B}$$

$$\Delta G_{cav} = \gamma_{MS}\Delta A_{MS} = \gamma_{MS}(A_{AB} - A_A - A_B)$$

Docking of Unbound Structures

- All previous examples are “easy”
  - structures **A** and **B** derived from true complex structure (bound structures)
  - Geometric complementarity guaranteed!
- More interesting problem:
  - Use unbound structures
  - Problem: few test sets - **A**, **B**, and **AB** have to be known!

Why Docking?

Decomposition of A/B into Disks

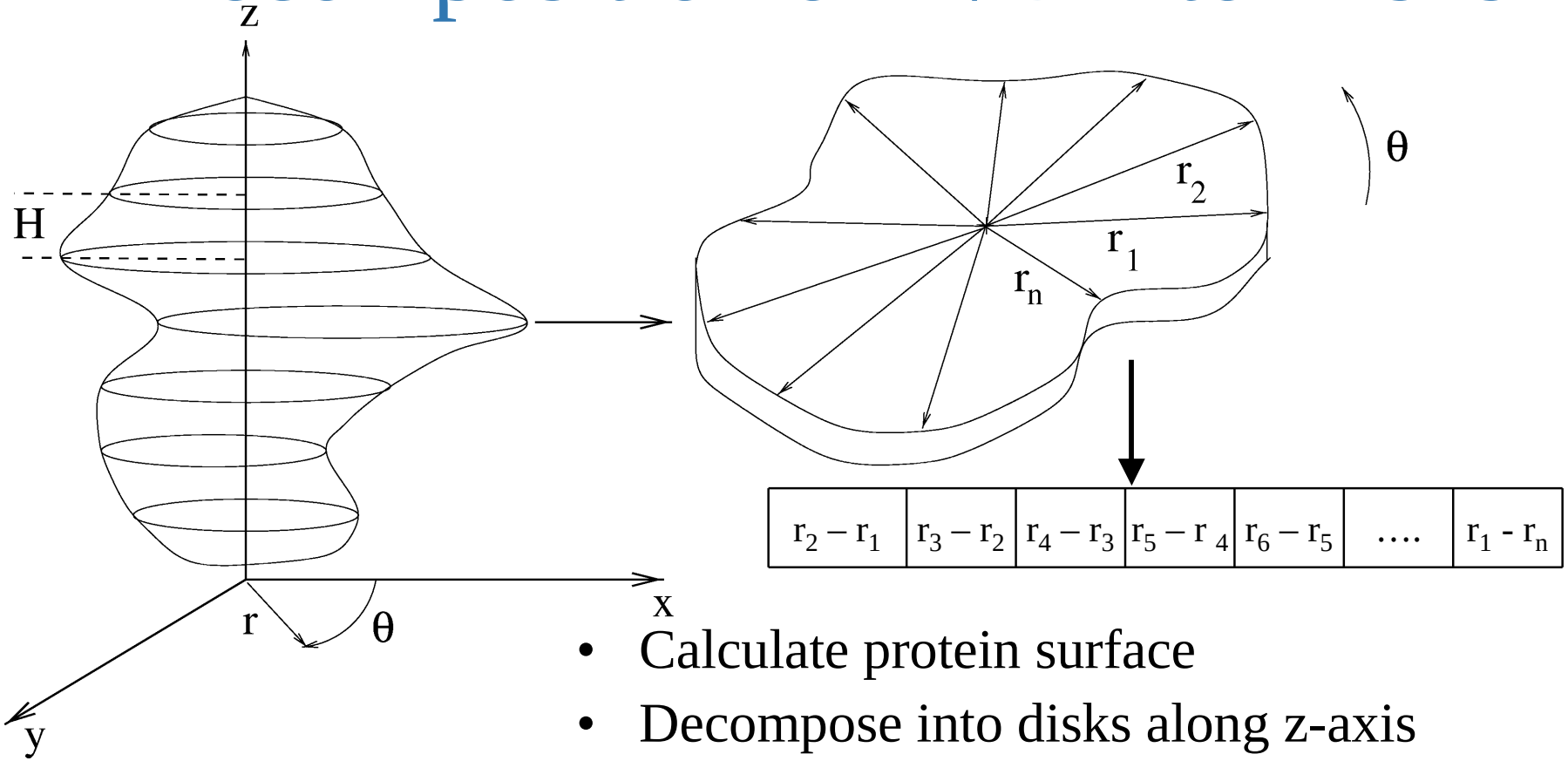
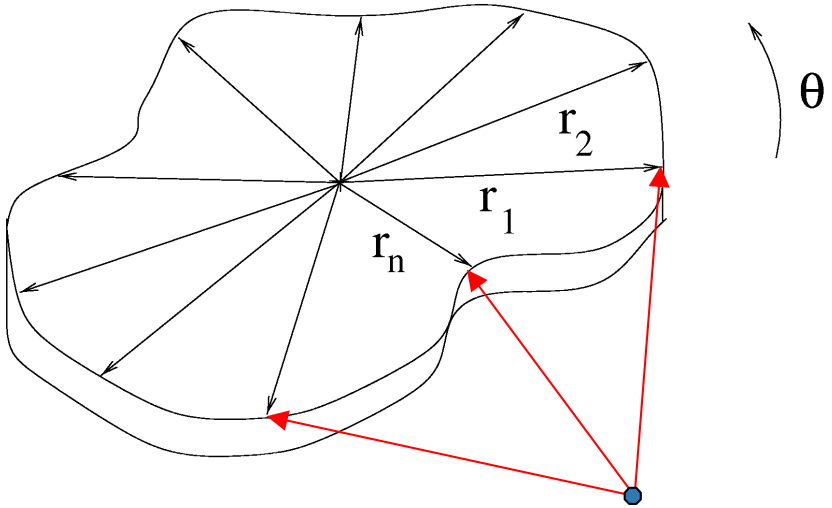


Fig. after Helmer-Citterich, Tramontano, JMB 1994

- Calculate protein surface
- Decompose into disks along z-axis
- Decompose disk into segments of equal area
- Differential surface: $(r_2 - r_1)$, $(r_3 - r_2)$,…

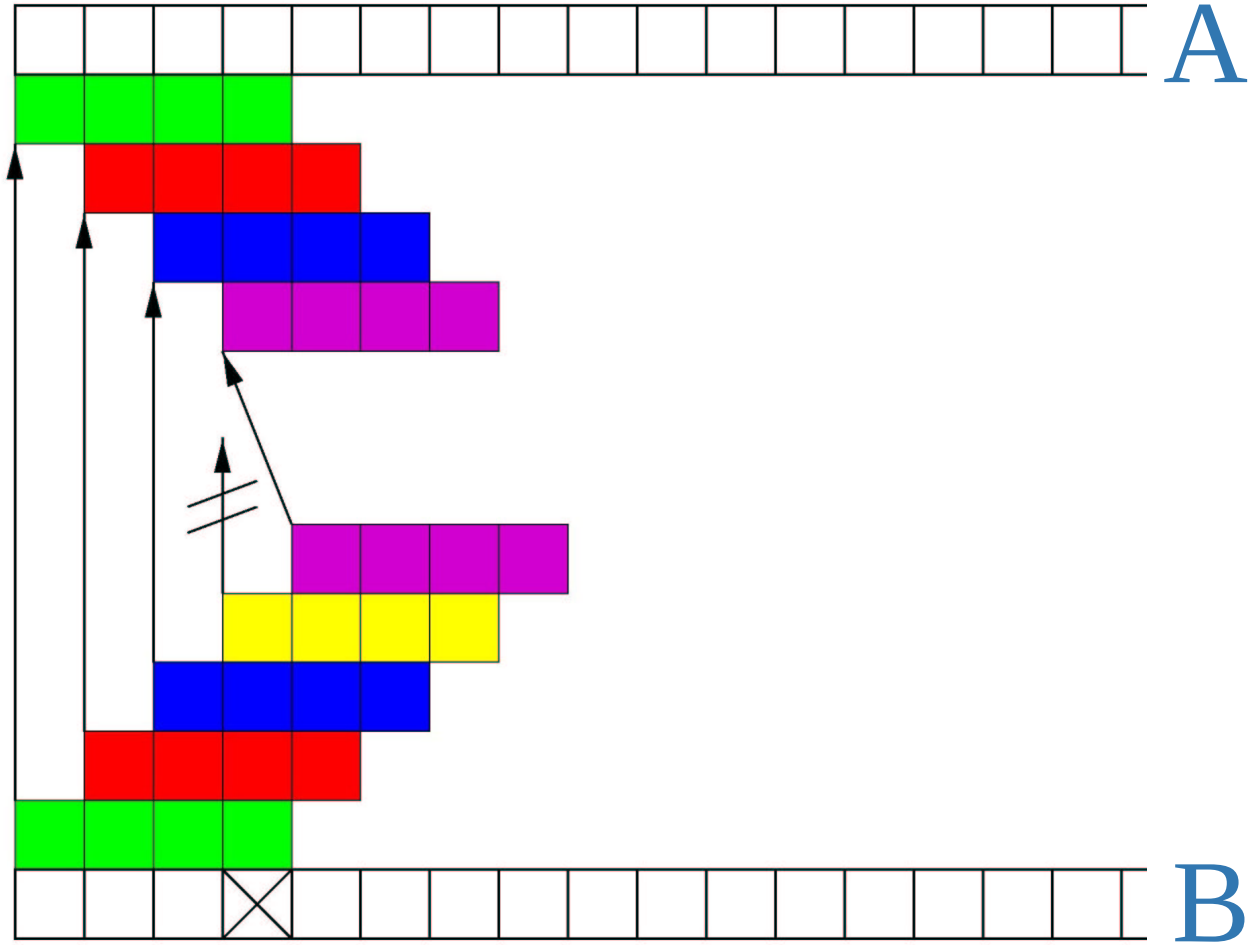
CENTER FOR BIOINFORMATICS

Matching of Disc Segments



- Try to match *windows*, i.e. stretches of successive radii in a disk, from **A** and **B**
- Matching windows define tentative transformations
- Aggregate overlapping windows to *regions*

Horizontal Windows Growing



Overall Algorithm

- Determine surface of **A** and **B**
- Decompose **A** into disks along z-axis
- For all rotations of **B**:
  - Decompose **B** into disks
  - Find matching windows on differential surface
  - Grow windows horizontally
  - Grow windows vertically (→ regions)
  - Each region defines a transformation