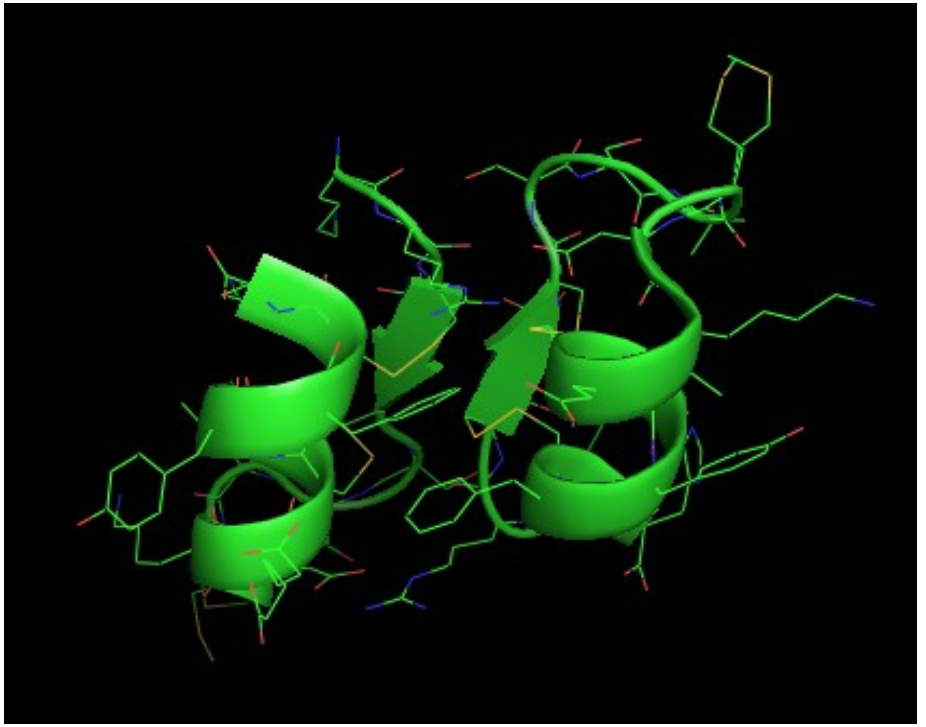
EFFICIENT CONFORMER LIBRARIES TO IMPROVE SIDECHAIN OPTIMIZATION



CIBM Seminar - 20th September 2011

Sabareesh Subramaniam
Senes Lab, UW Biochemistry

PROTEINS

Peptide = chain of amino acids

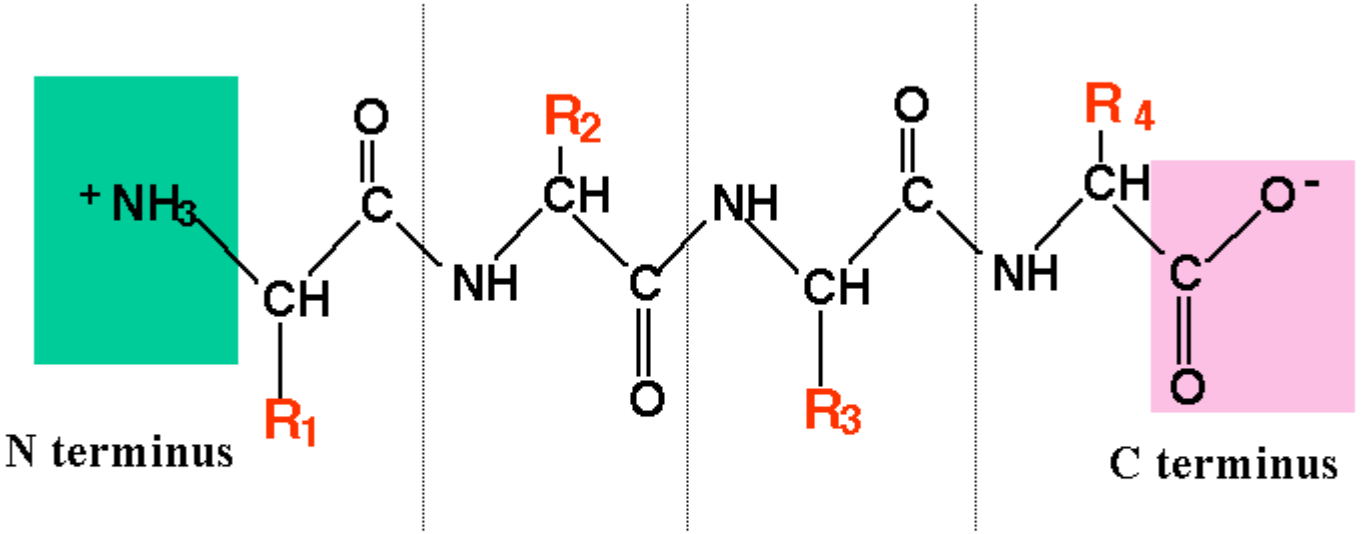
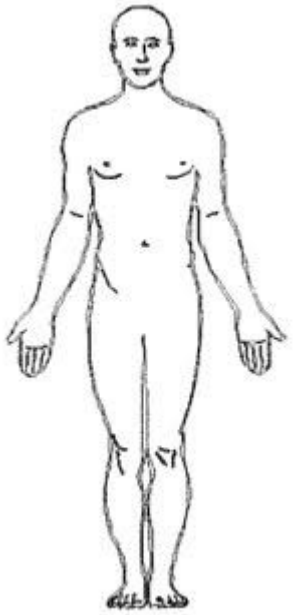


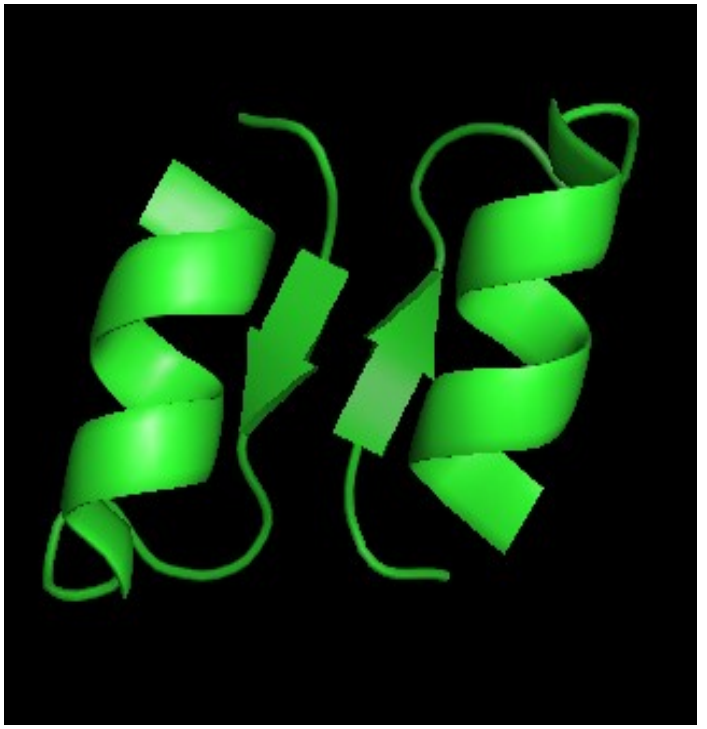
Image from www.molecularstation.com

PROTEINS

- Perform important biological functions
- Structure and function closely related
- Structures VERY helpful to study proteins
- X-Ray Crystallography, NMR to determine structure
- Computational modelling when structure not available

Structure Prediction via Homology Modeling

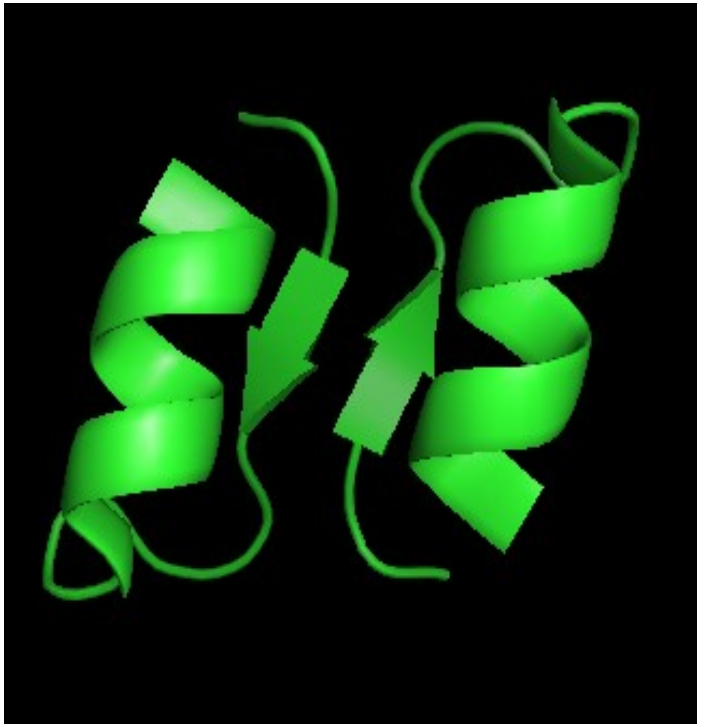


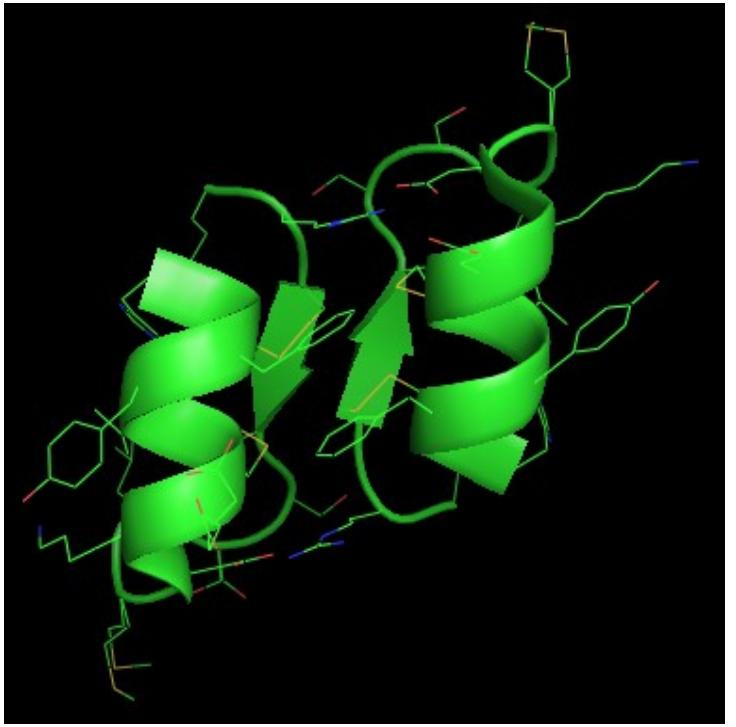


SIDECHAIN OPTIMIZATION

Backbone

Add sidechains to achieve minimum energy configuration



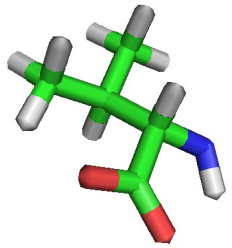


COMPUTATIONAL MODELING

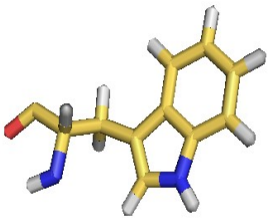
- Structure Prediction
  - Sequence → backbone(from homolog) → **Sidechain optimization**
- Protein Design
- Docking

SIDECHAIN OPTIMIZATION

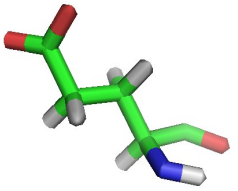
GLU

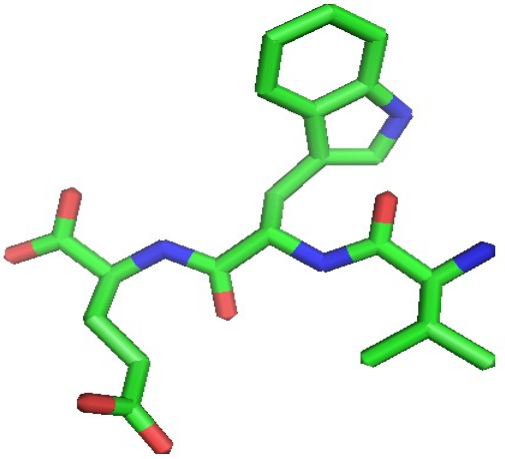


TRP

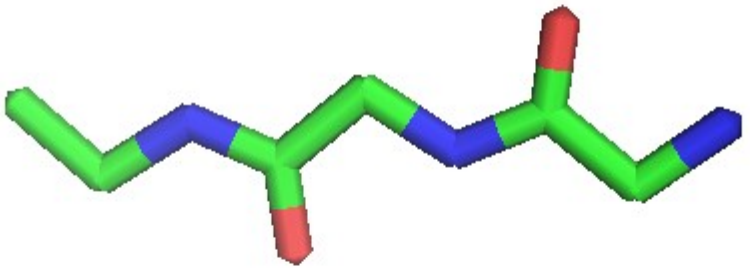


VAL

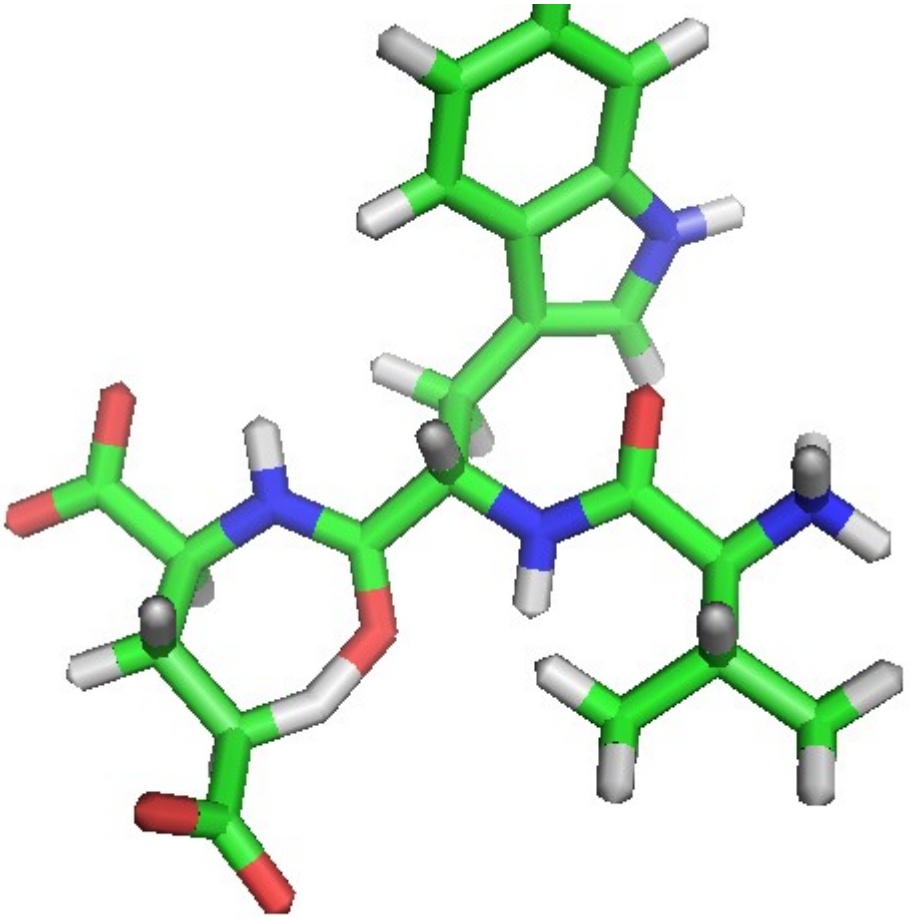




SIDECHAIN OPTIMIZATION

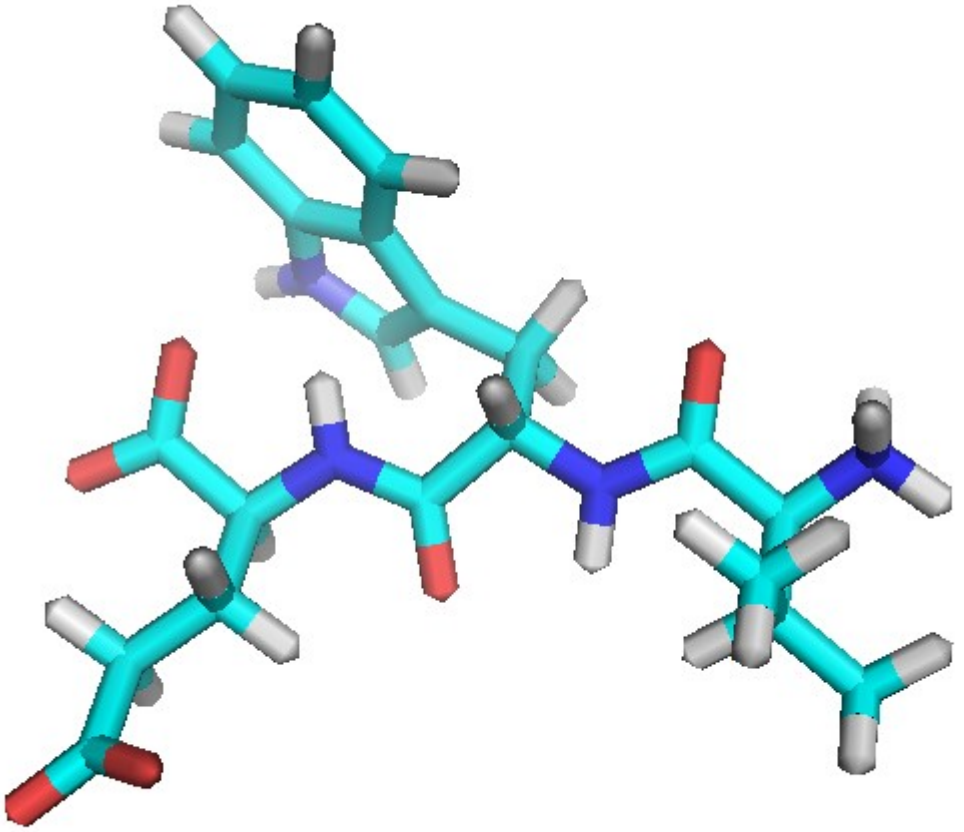


SIDECHAIN OPTIMIZATION



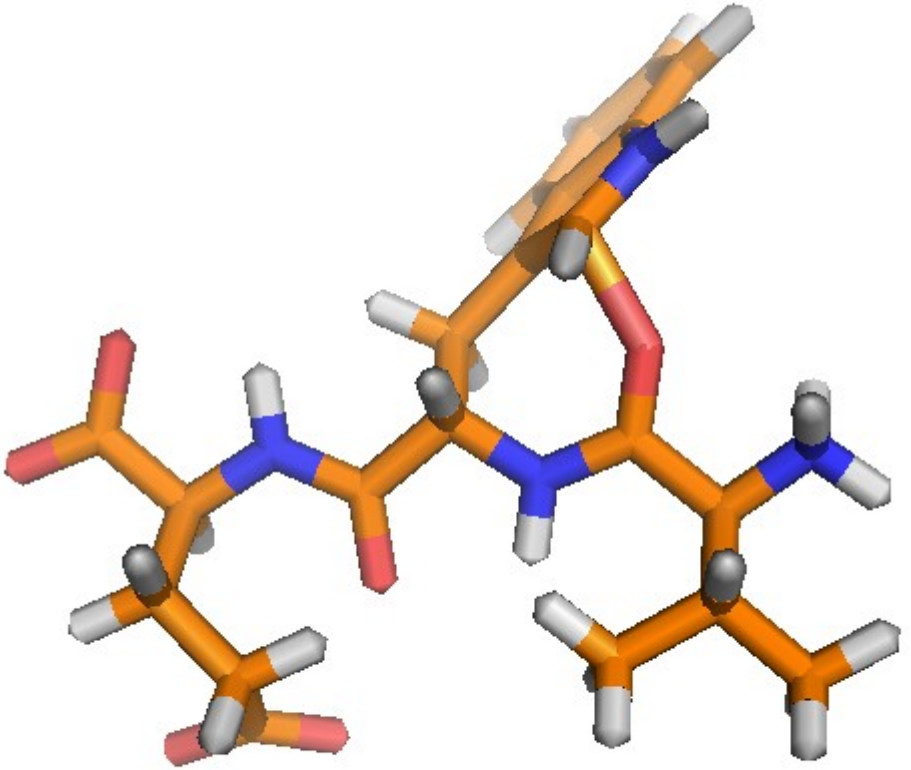
Energy = 40000 Kcal / Mol

SIDECHAIN OPTIMIZATION



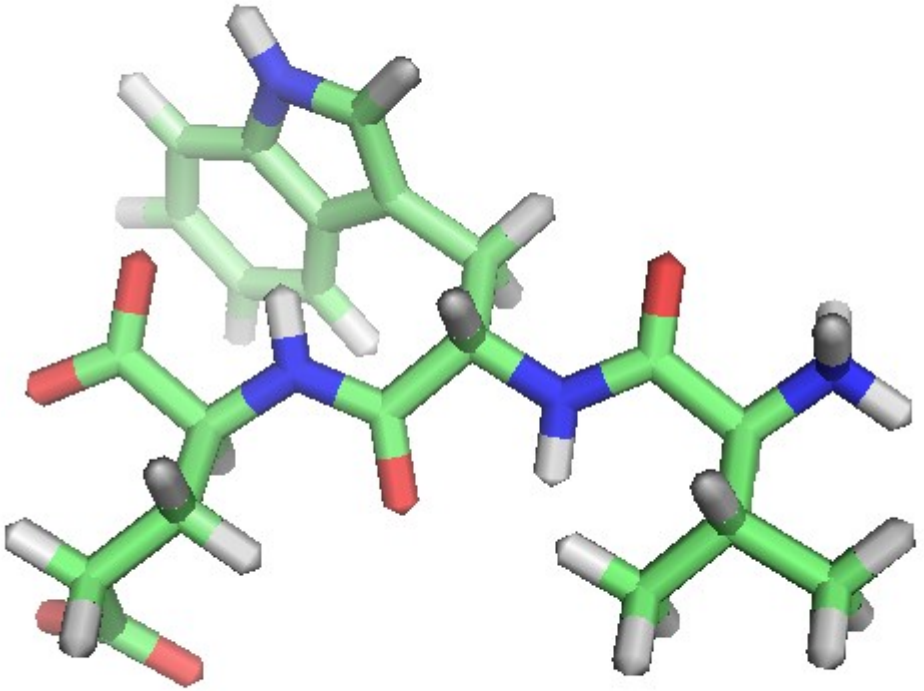
Energy = 60 Kcal / Mol

SIDECHAIN OPTIMIZATION



Energy = 30000 Kcal / Mol

SIDECHAIN OPTIMIZATION



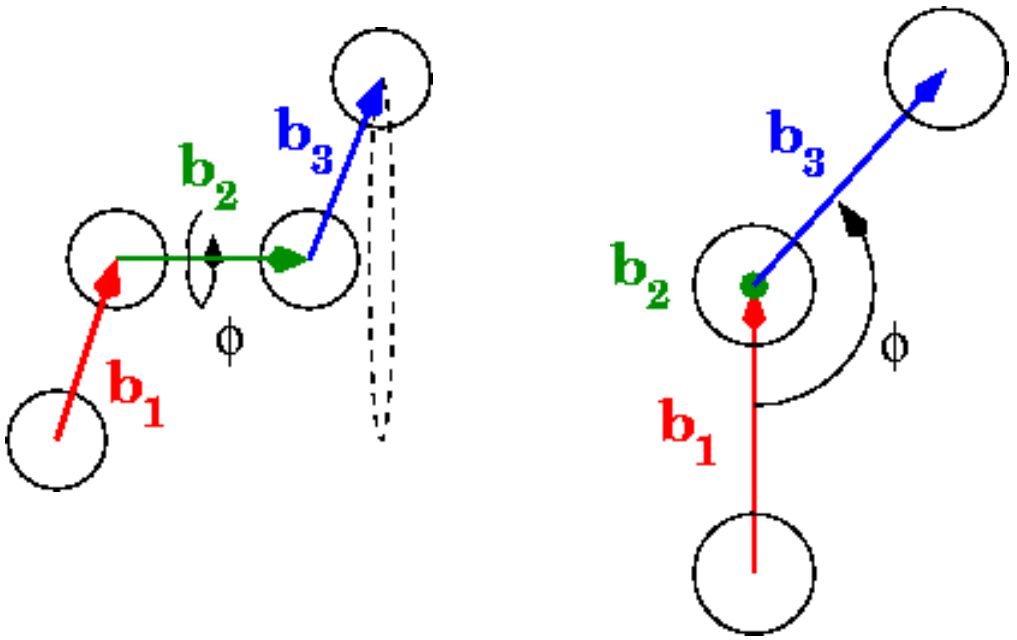
Energy = 70 Kcal / Mol

DEGREES OF FREEDOM

Bond distances

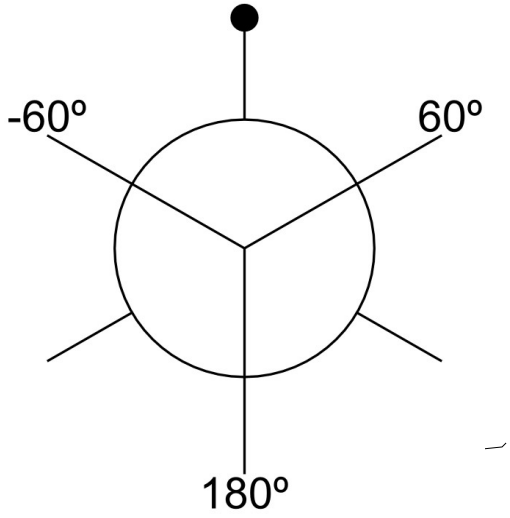
Bond Angles

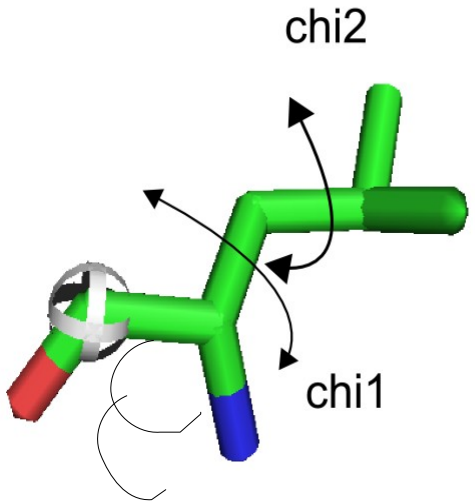
Dihedral or torsional angles

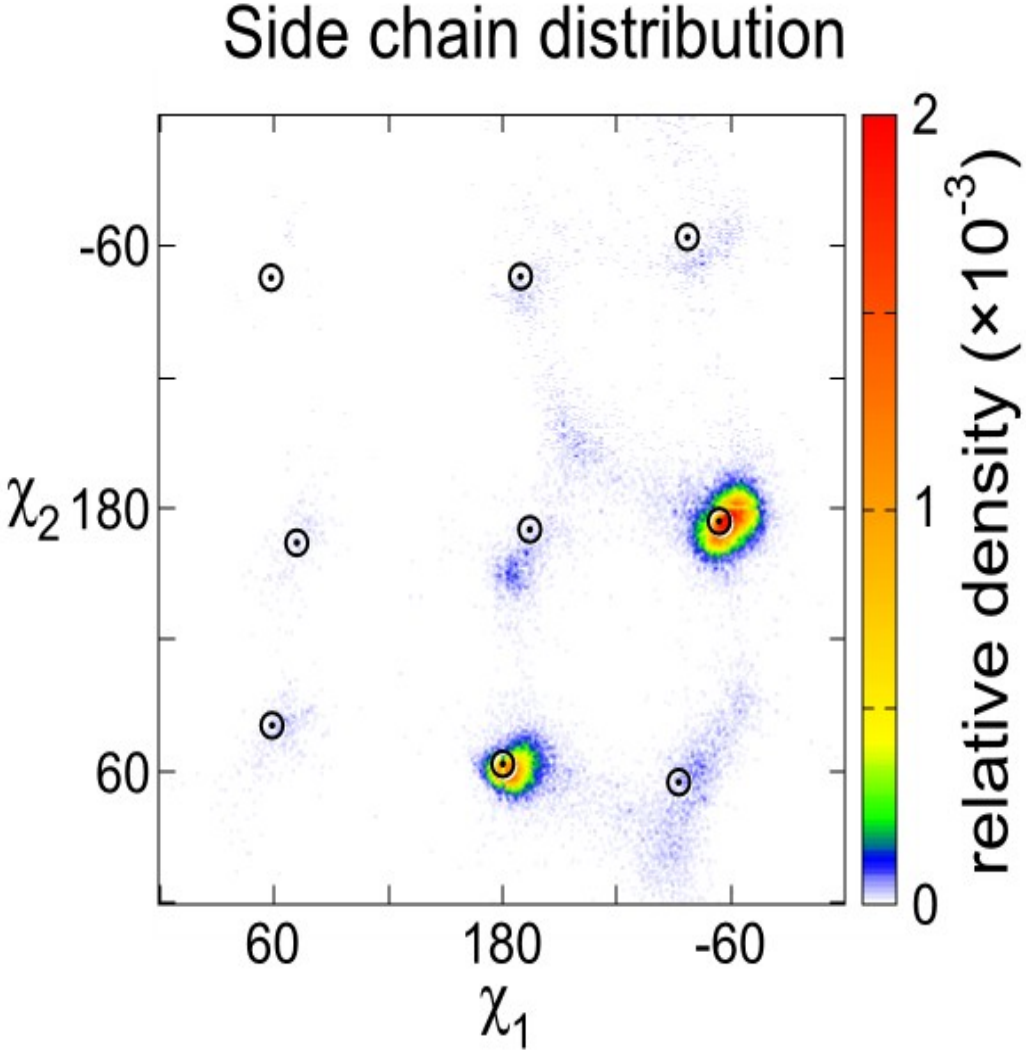


*Figures from Wikipedia

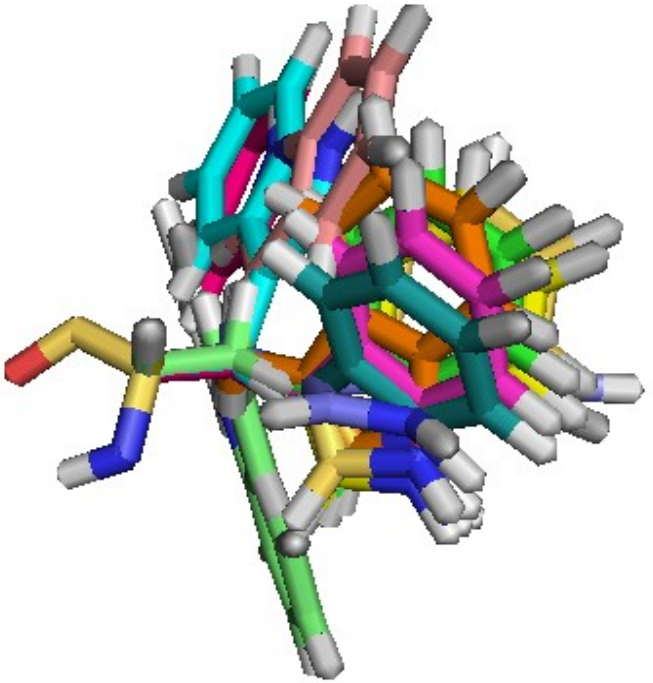
STATISTICS OF DIHEDRAL ANGLES







DISCRETIZED CONFORMATION LIBRARIES

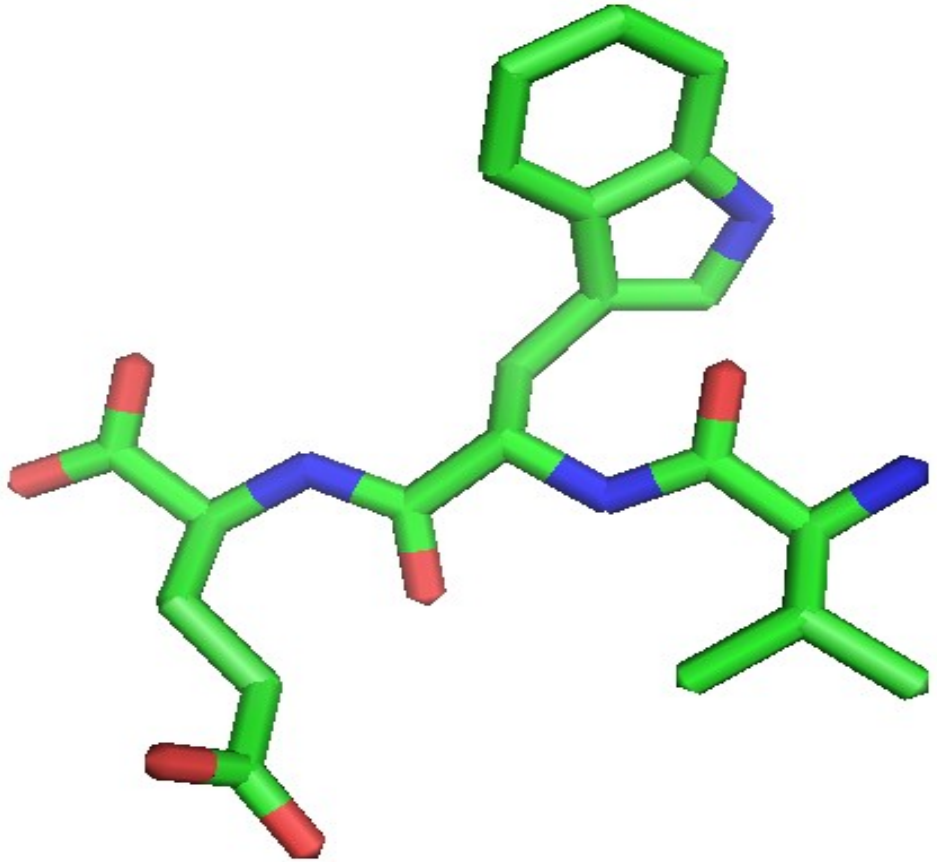


SIDECHAIN OPTIMIZATION

GLU

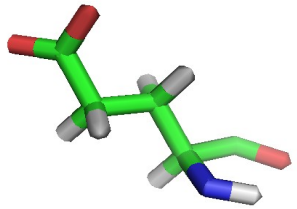
TRP

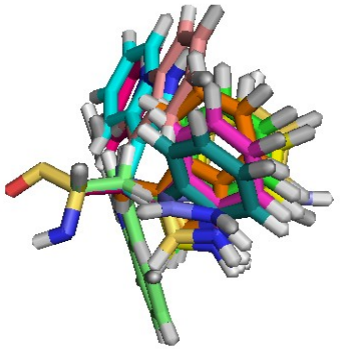
VAL



COMBINATORIAL SEARCH SPACE (3-D JIGSAW PUZZLE)

GLU

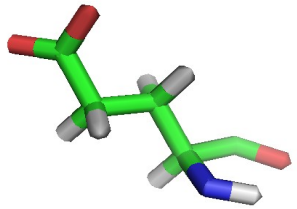


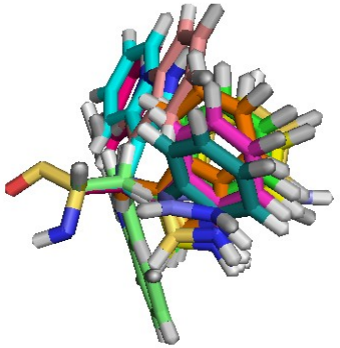


No of conformations to search 36

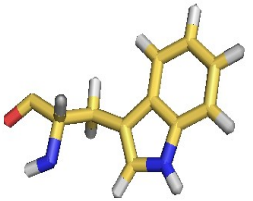
COMBINATORIAL SEARCH SPACE (3-D JIGSAW PUZZLE)

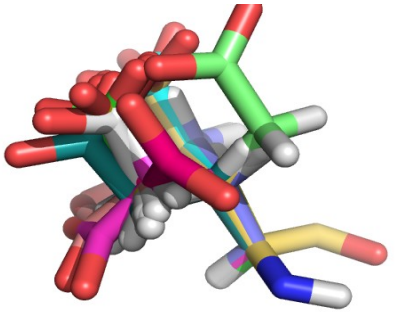
GLU





TRP

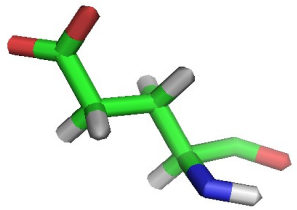


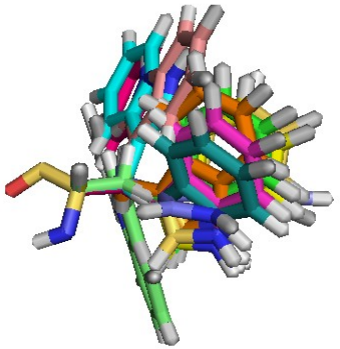


No of conformations to search 36* 54

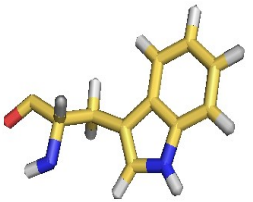
COMBINATORIAL SEARCH SPACE (3-D JIGSAW PUZZLE)

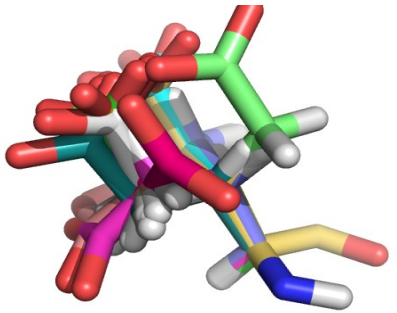
GLU



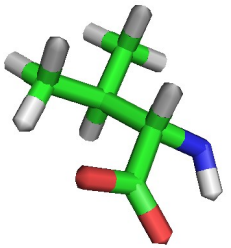


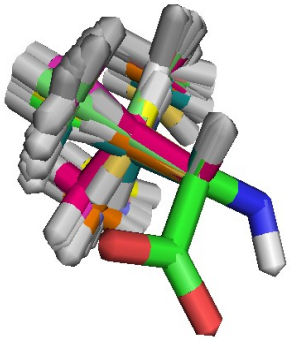
TRP





VAL

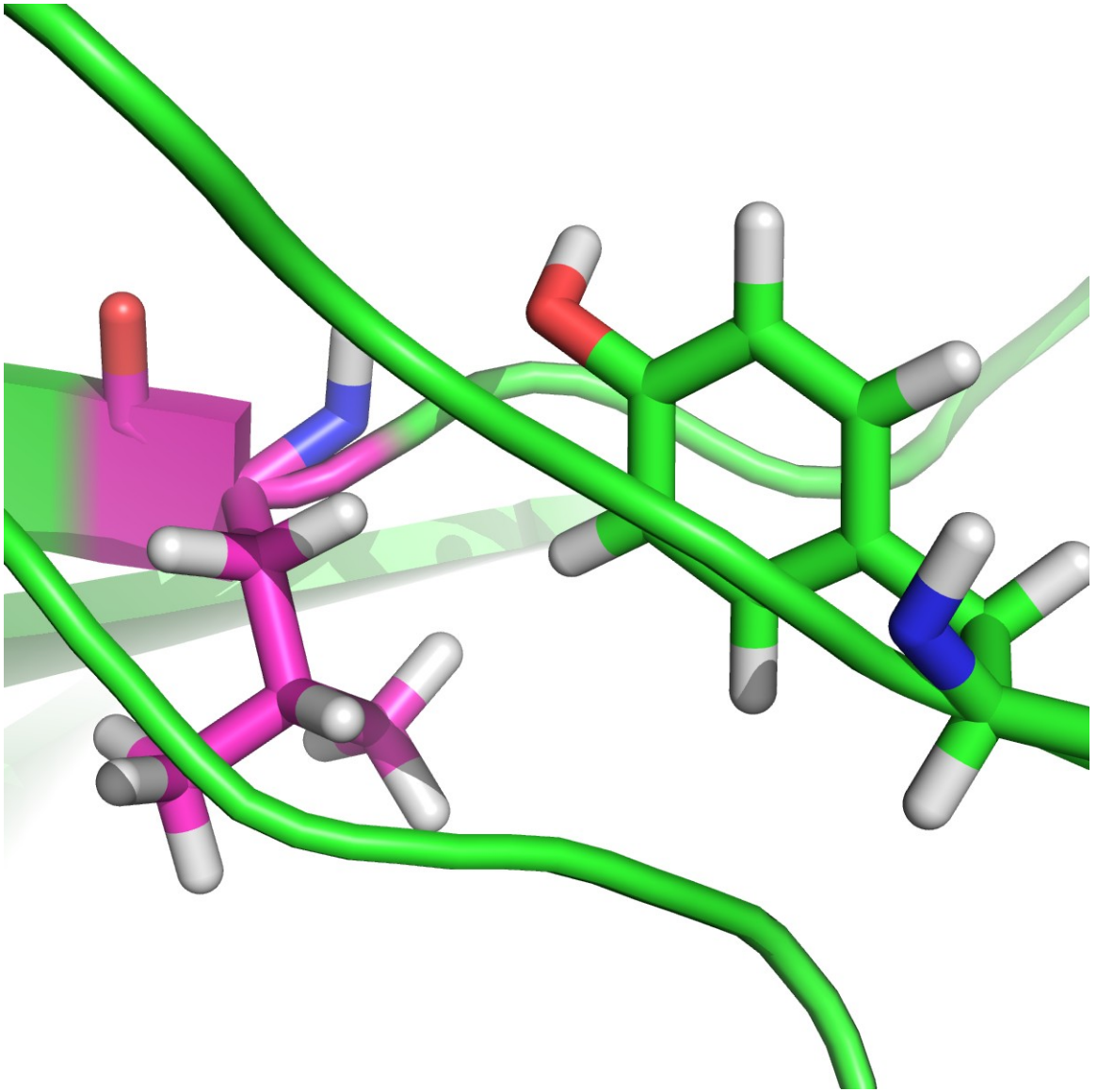


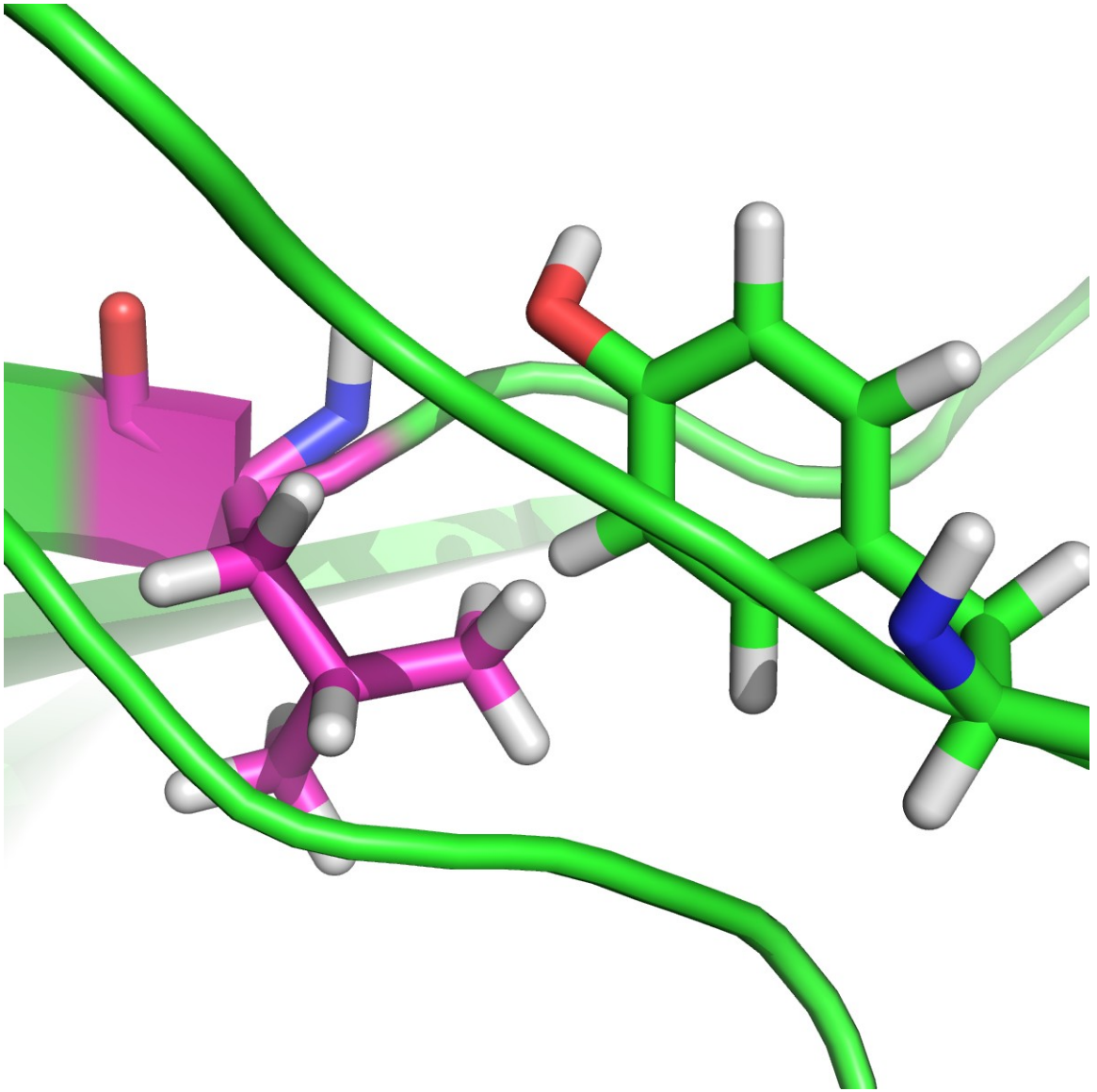


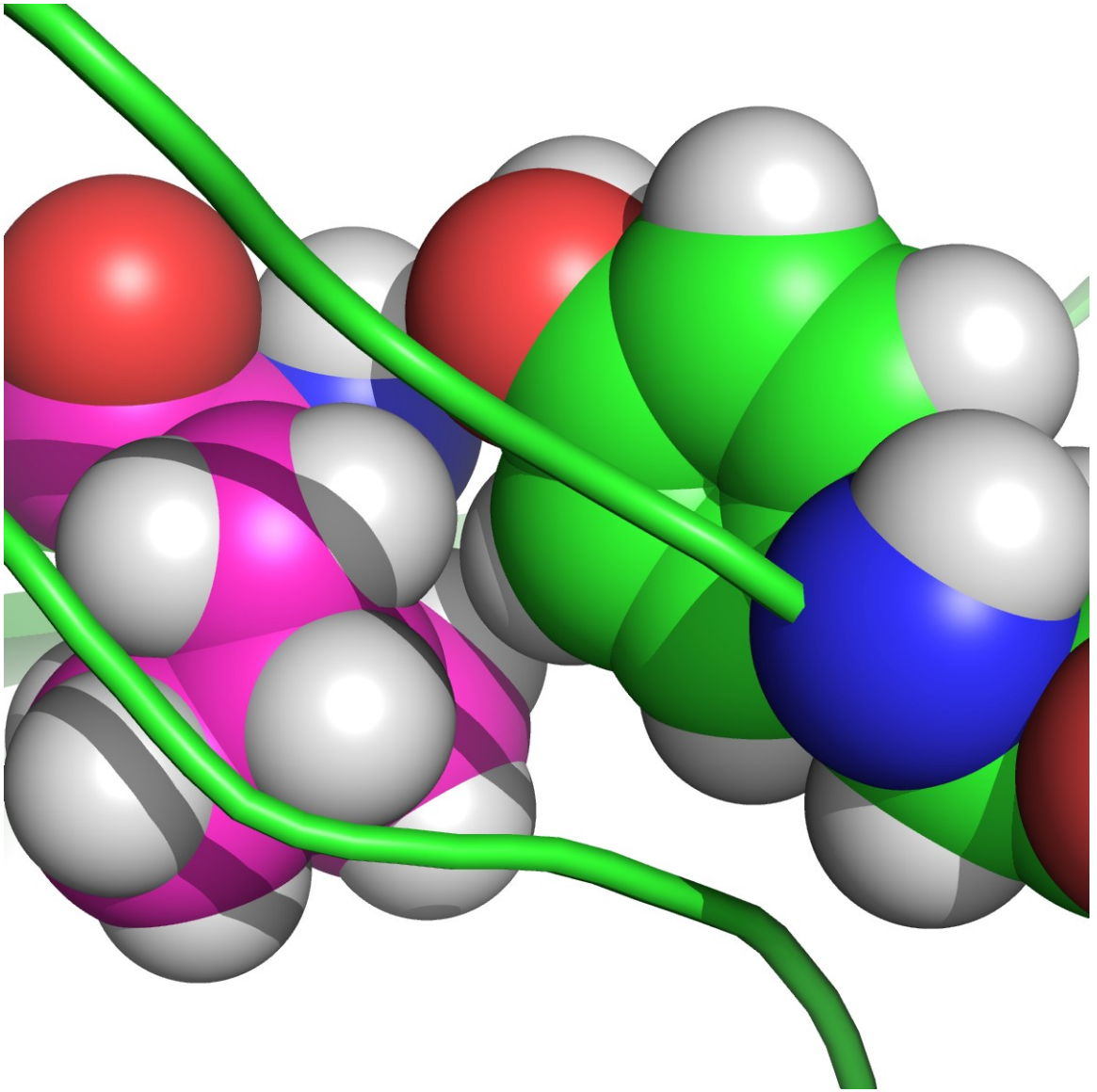
No of conformations to search 36* 54 * 3 = 5832
Typically > 10 ^ 60

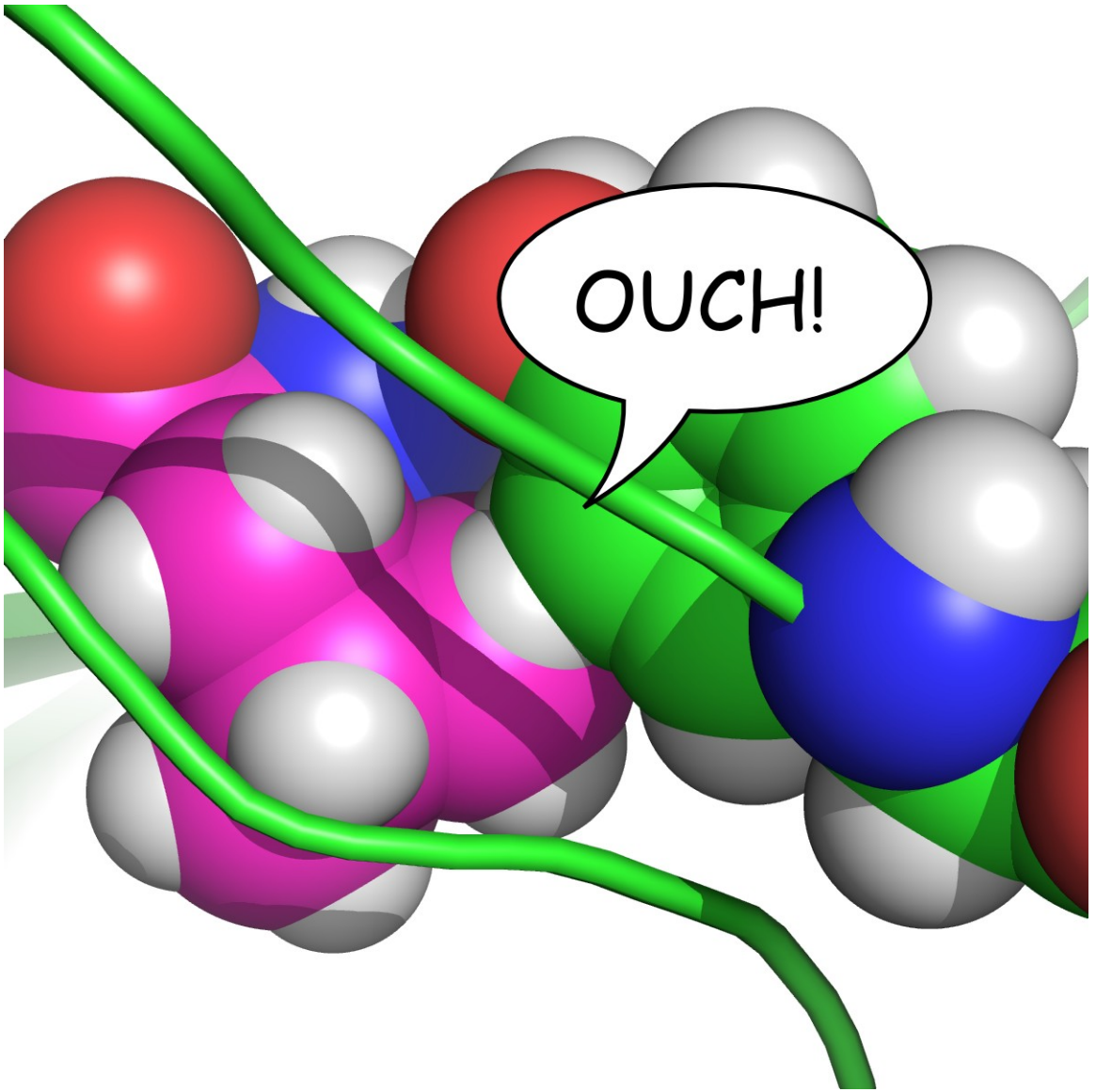
SEARCH ALGORITHMS

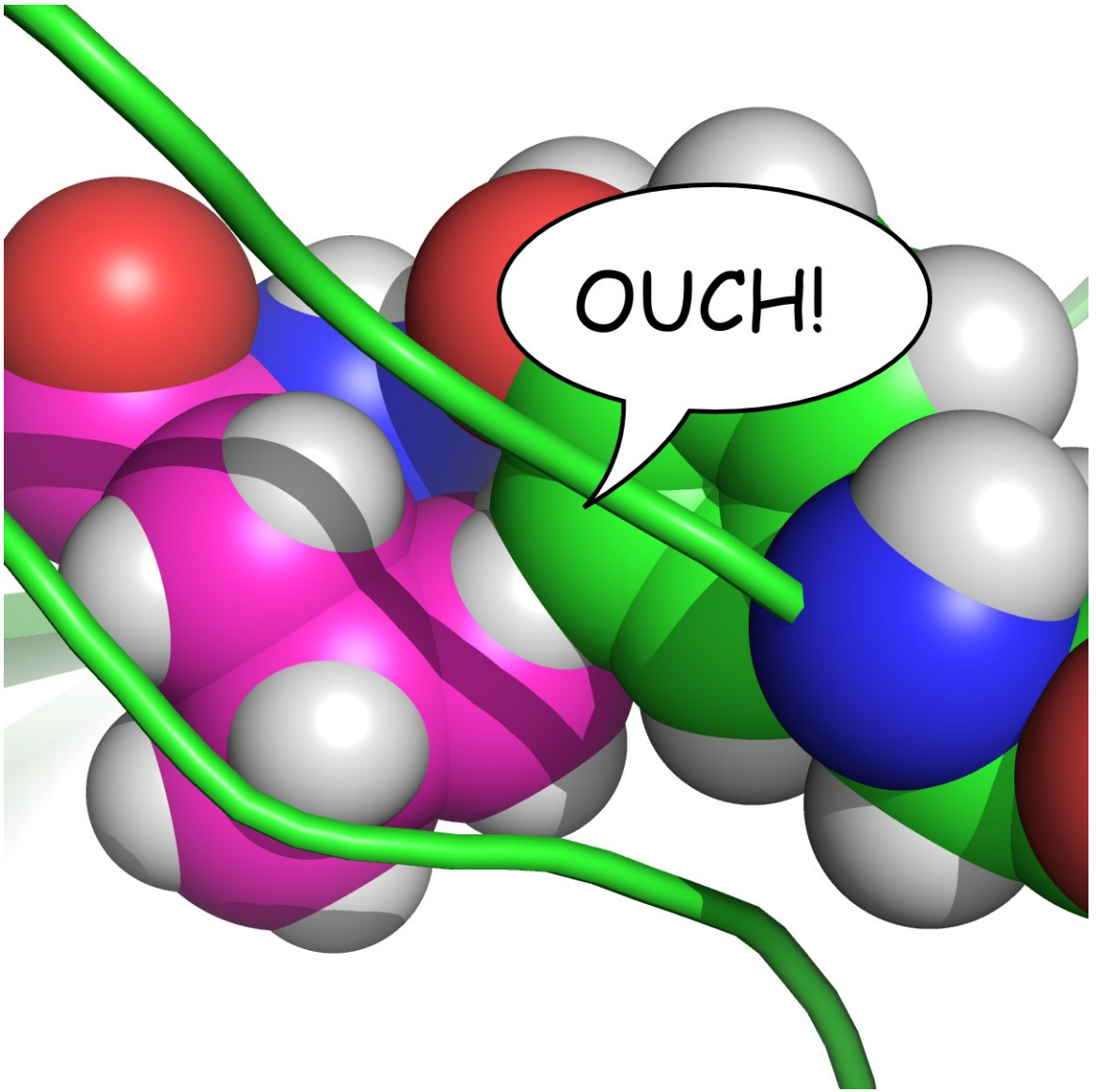
- Dead End Elimination
- Self Consistent Mean Field
- A* search
- Monte Carlo Simulated Annealing
- Graph decomposition





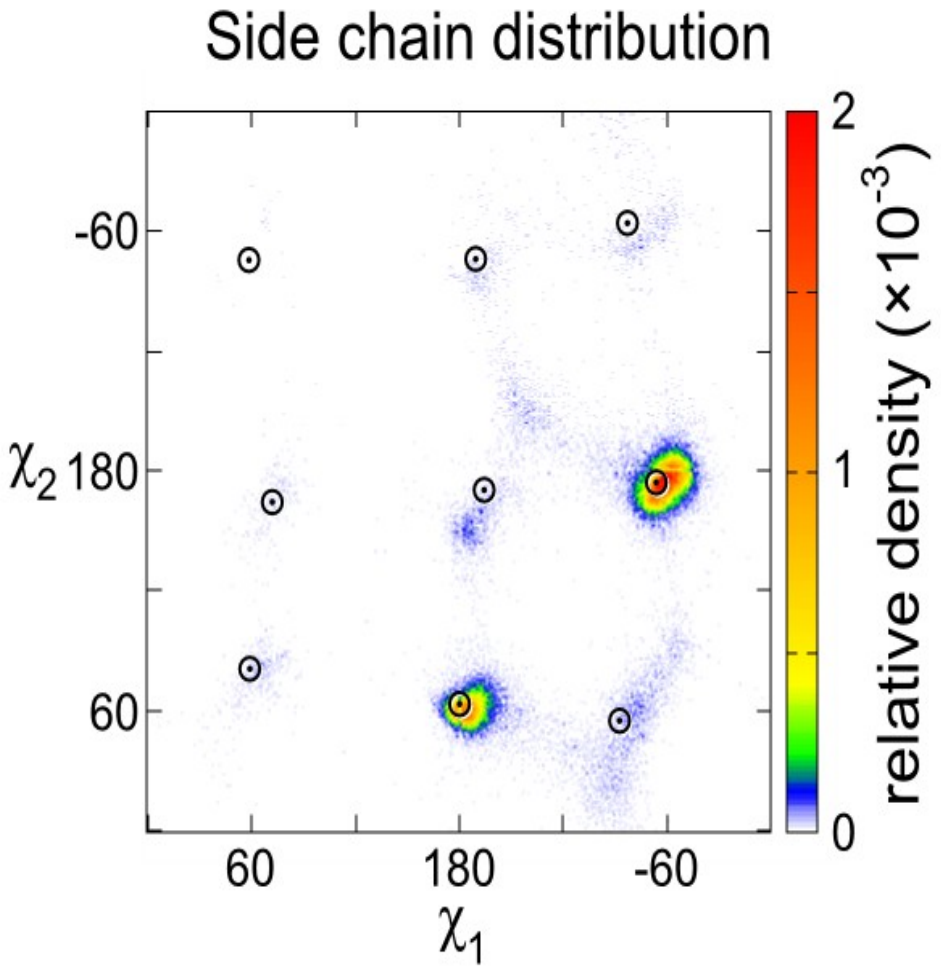



OUCH!

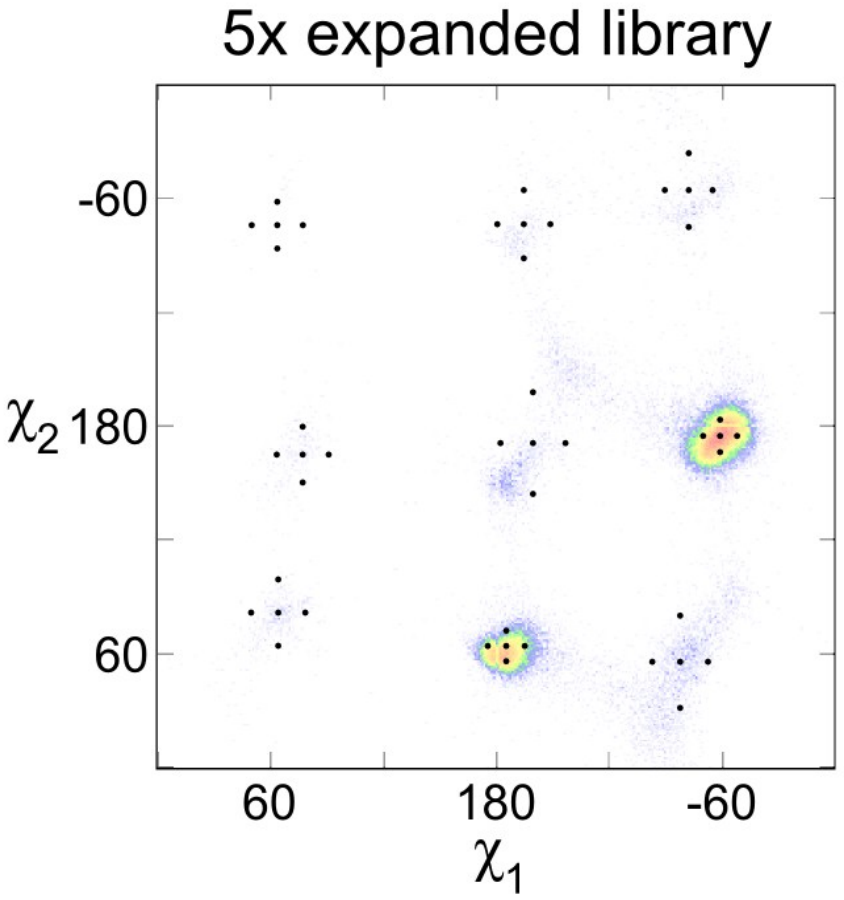


We need more sampling

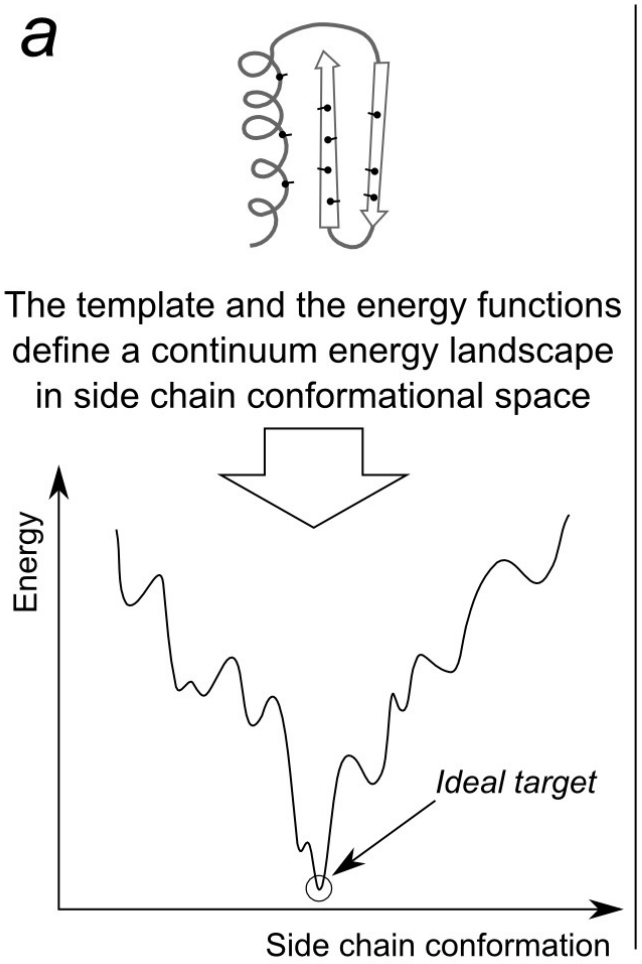
MORE SAMPLING



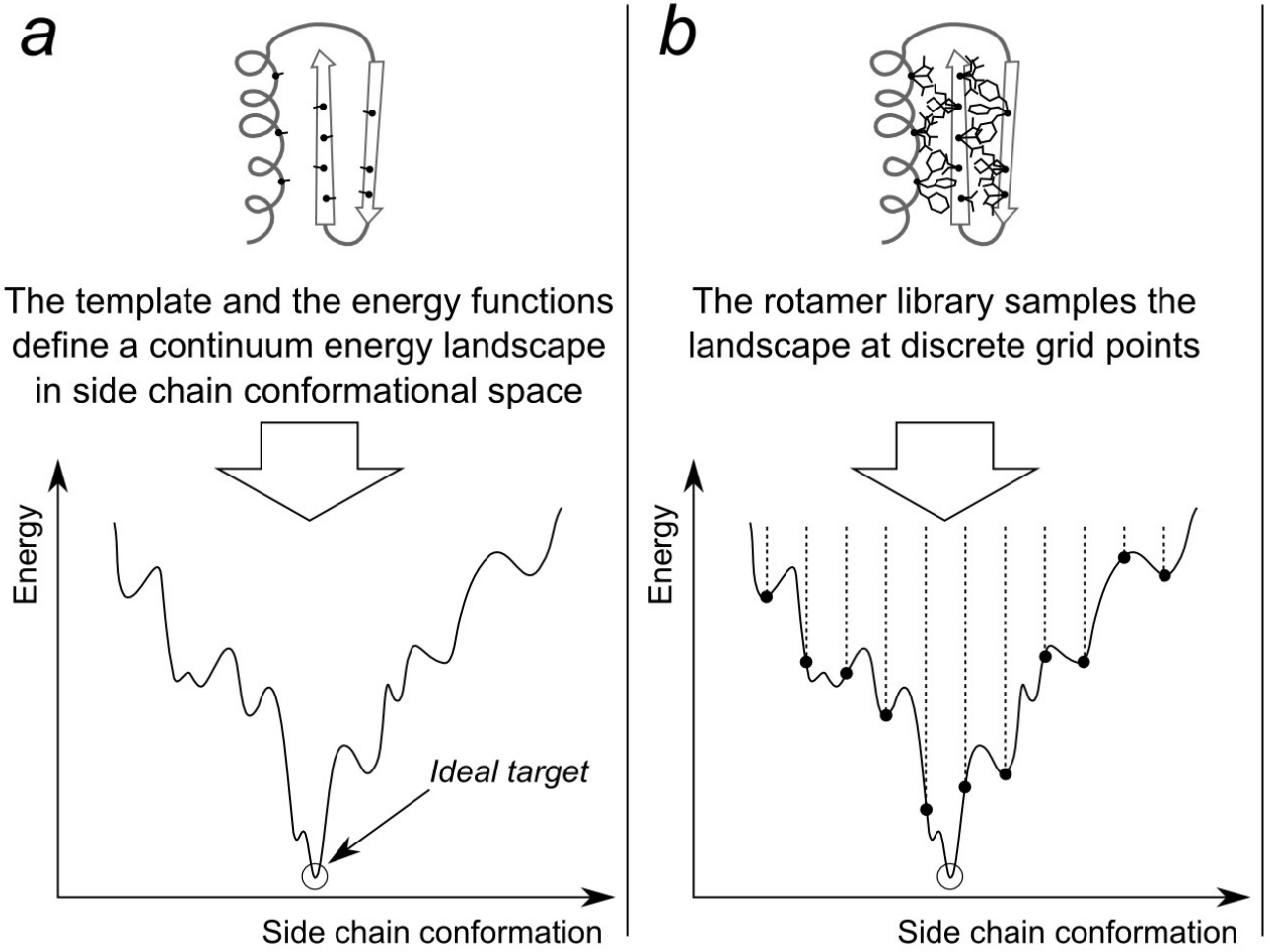
MORE SAMPLING



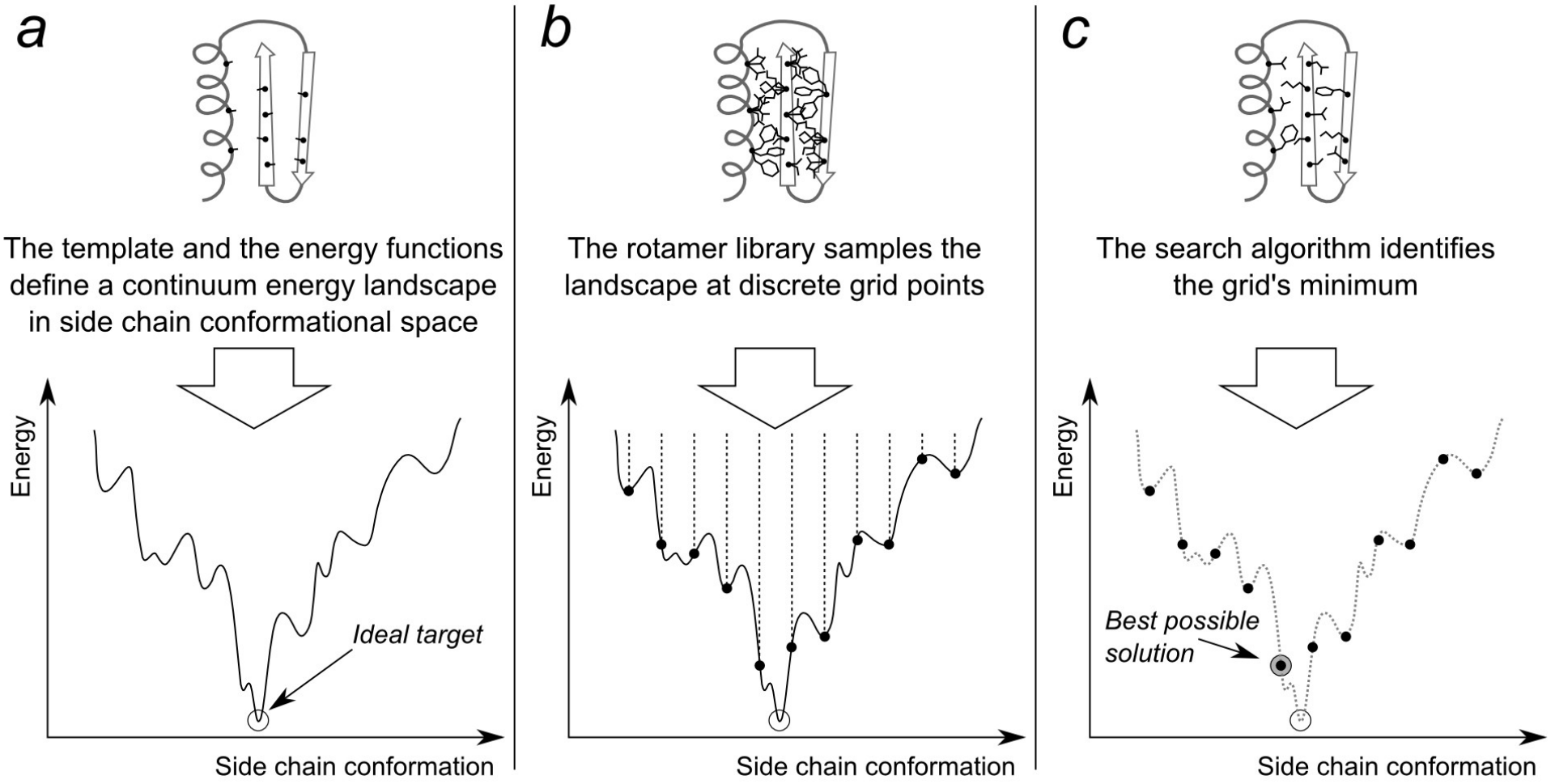
ROTAMER LIBRARY DETERMINES QUALITY OF SOLUTION



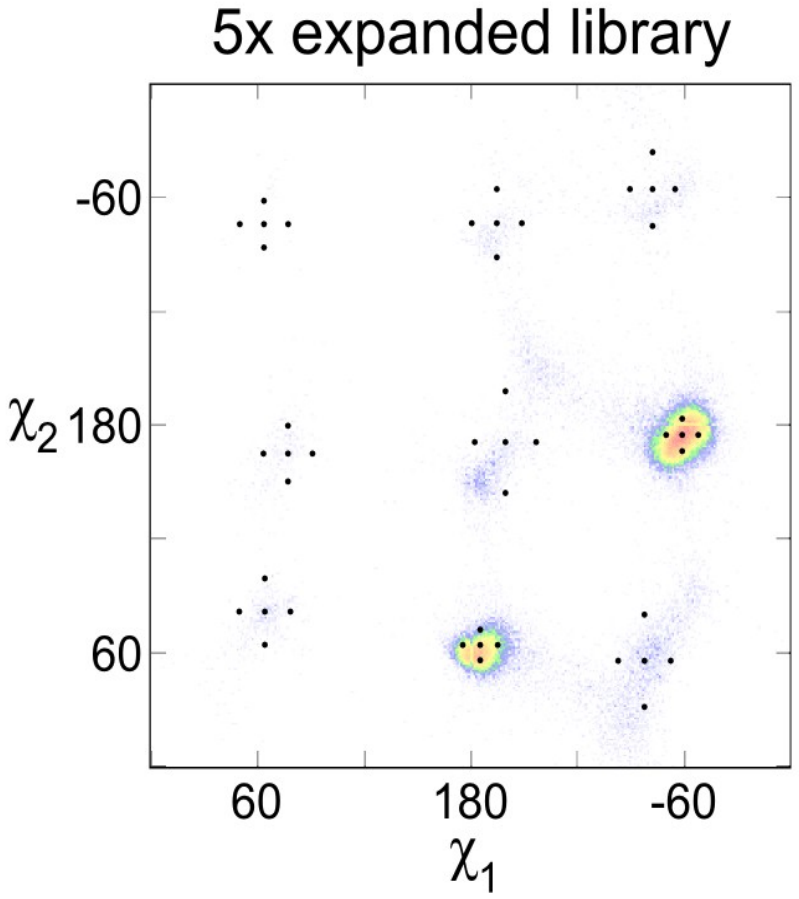
ROTAMER LIBRARY DETERMINES QUALITY OF SOLUTION



ROTAMER LIBRARY DETERMINES QUALITY OF SOLUTION



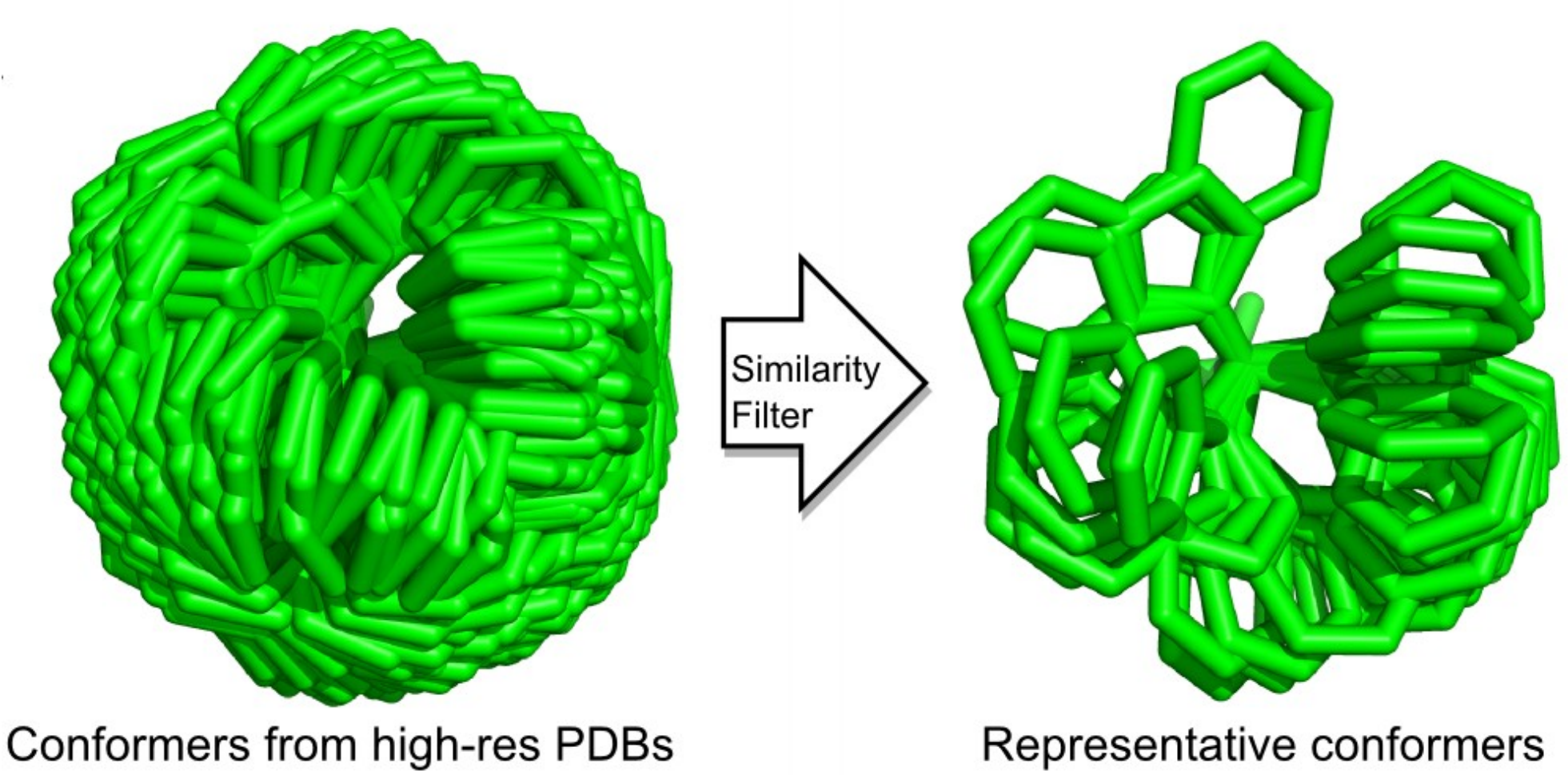
FIXED NUMBER OF CONFORMERS



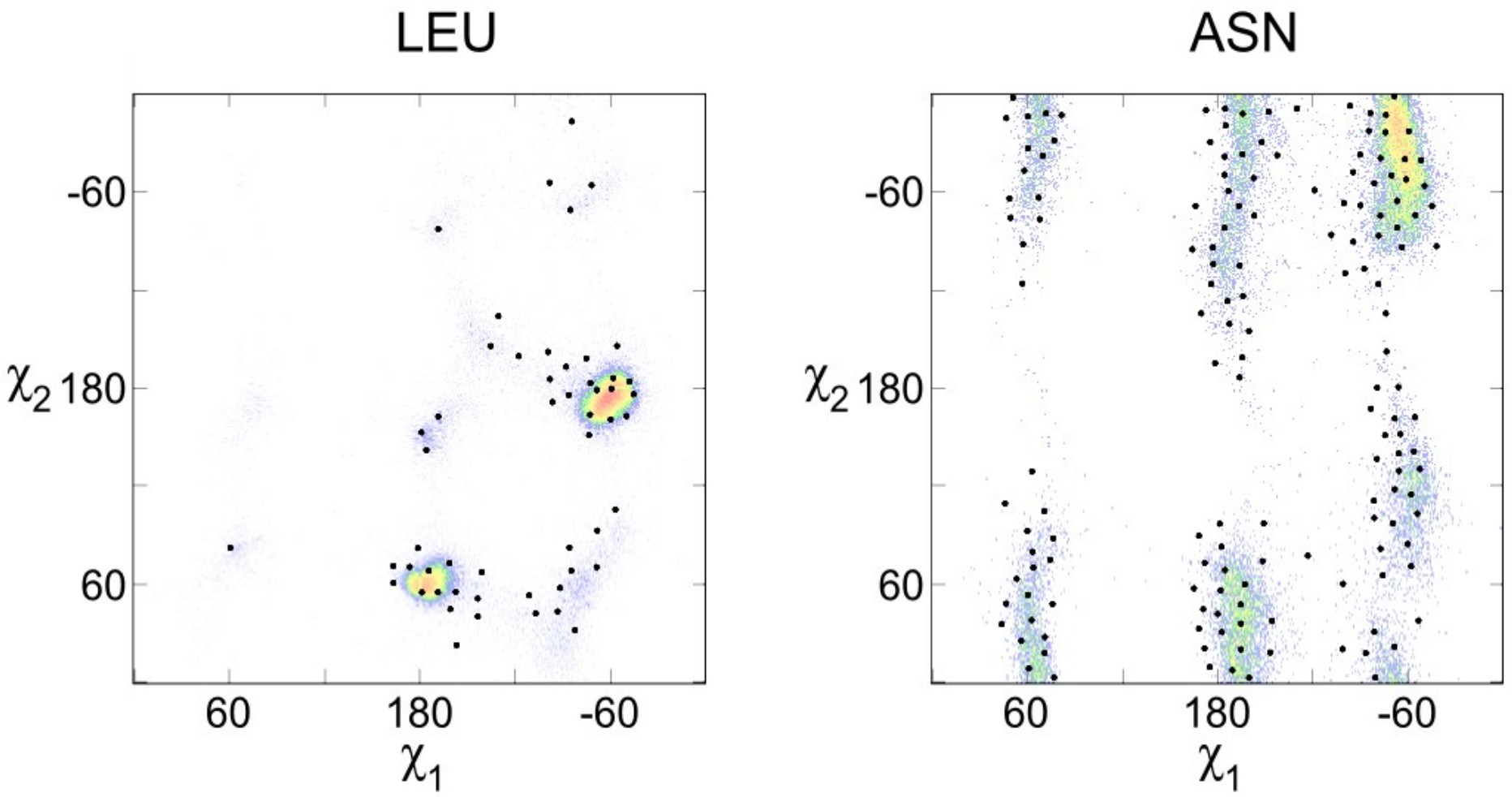
9 → 45 rotamers

No number in between

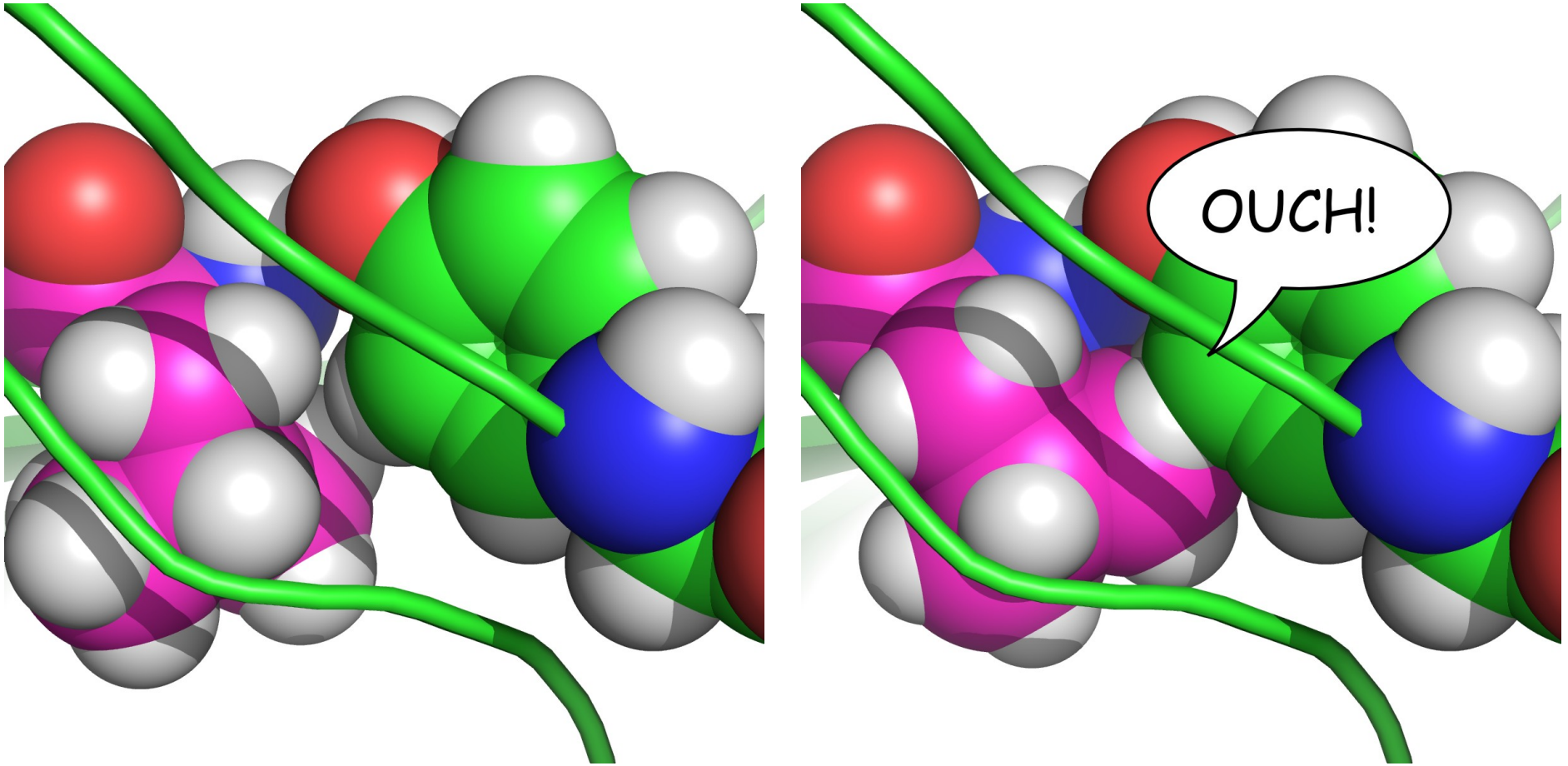
GEOMETRIC FILTERS LEAD TO CONFORMER LIBRARIES



IGNORES THE NATURAL DISTRIBUTION



- People have been looking for solutions using the statistical distribution in structures
- However, the problem is that sampling is related to the energetics in a way that is difficult to predict



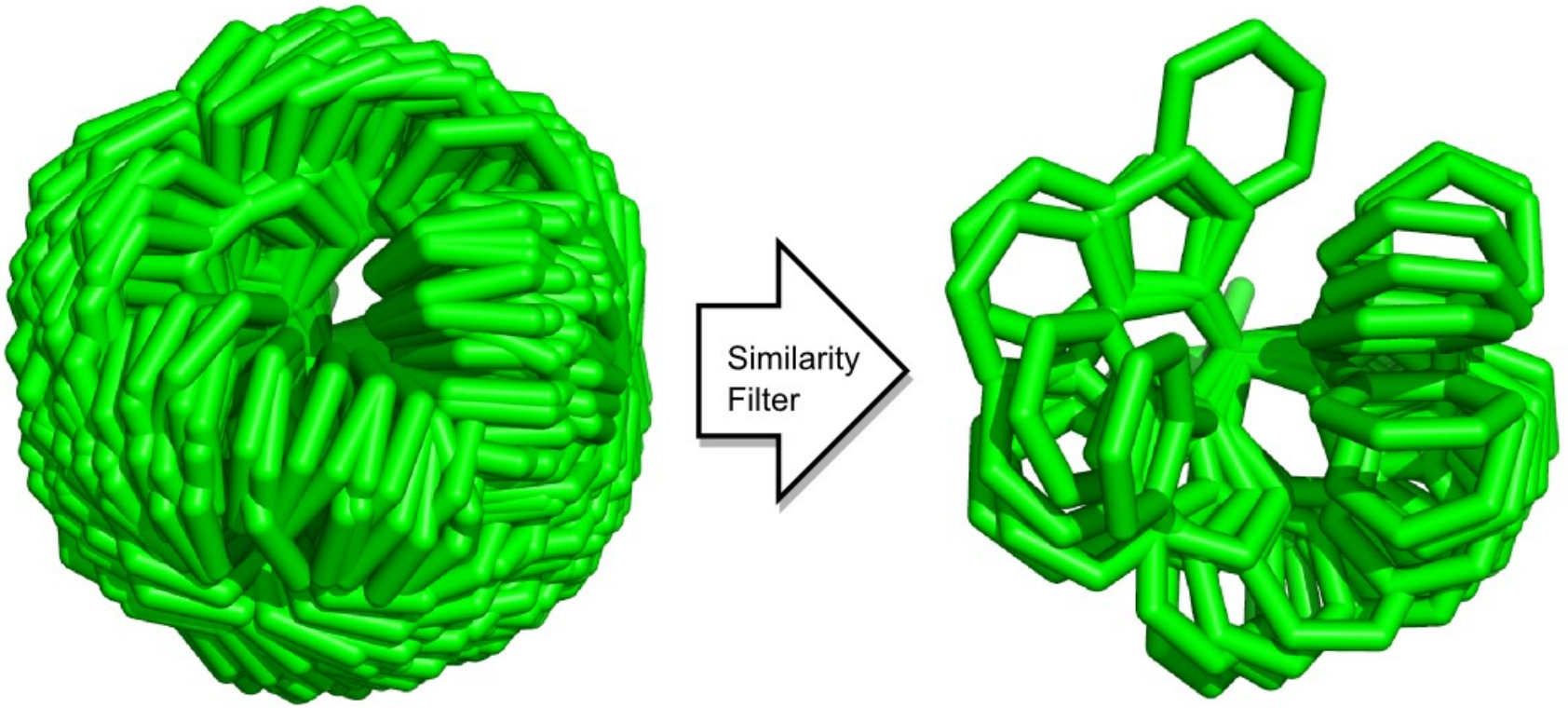
- People have been looking for solutions using the statistical distribution in structures
- However, the problem is that sampling is related to the energies in a way that is difficult to predict
- Solution: use energetics to identify the best sampling strategy for side chain optimization

GOALS

Can we create a library that can outperform existing libraries in terms of speed and/or Energies ?

Can we create a flexible library where the conformers are in some useful order?

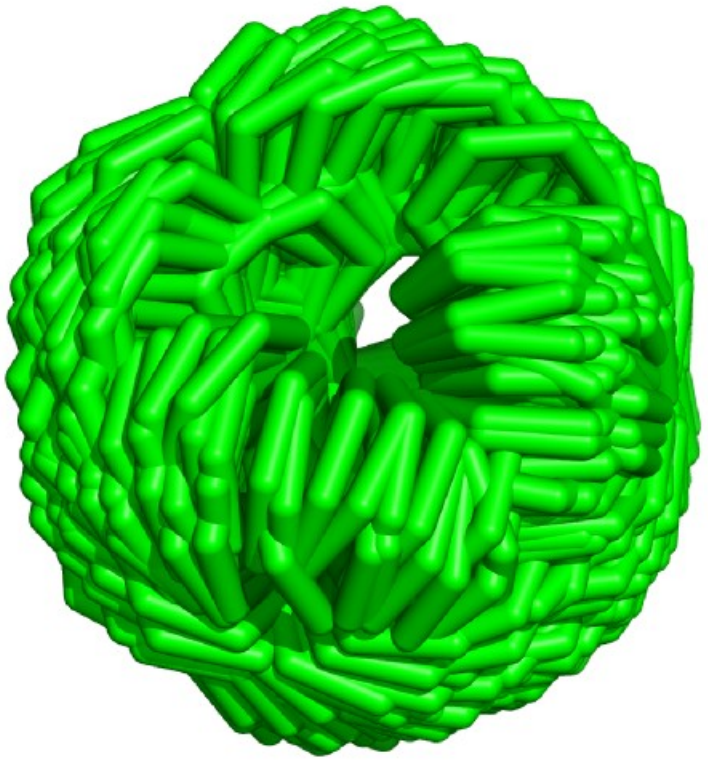
Can we sort the conformers instead of extracting a fixed-size subset?



5,000 conformers from high-res PDBs

50 representative conformers

Can we sort the conformers instead of extracting a fixed-size subset?



5,000 conformers from high-res PDBs

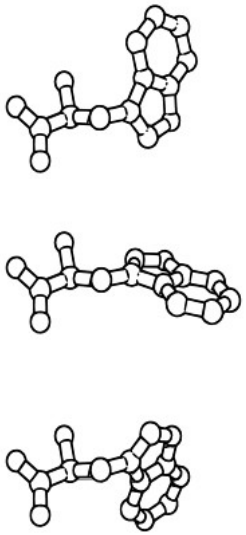
Energy based sorting

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
.....
.....
4999.
5000.

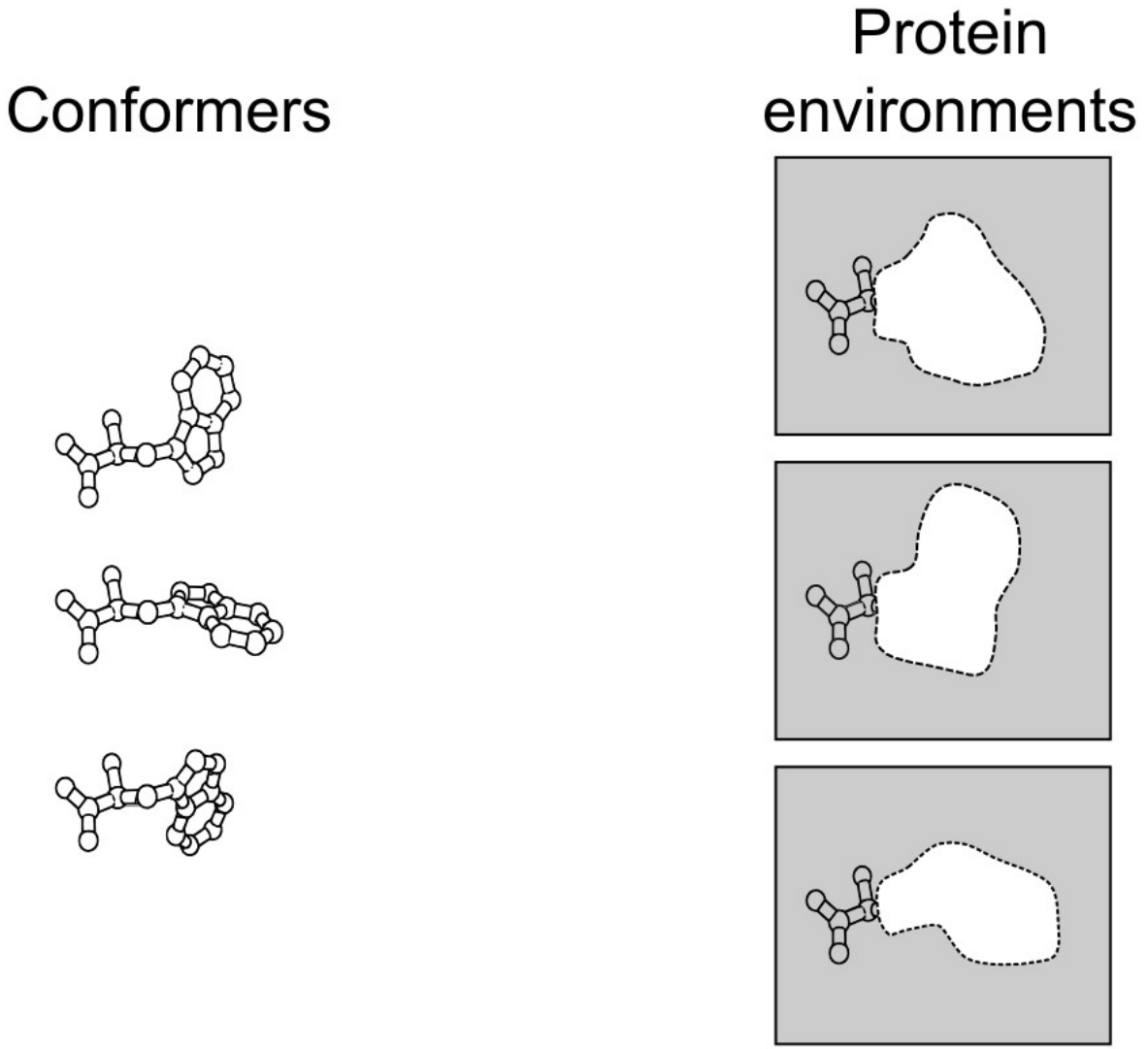
Sorted list of 5,000 conformers

- Use energetics to identify the best sampling strategy for side chain optimization

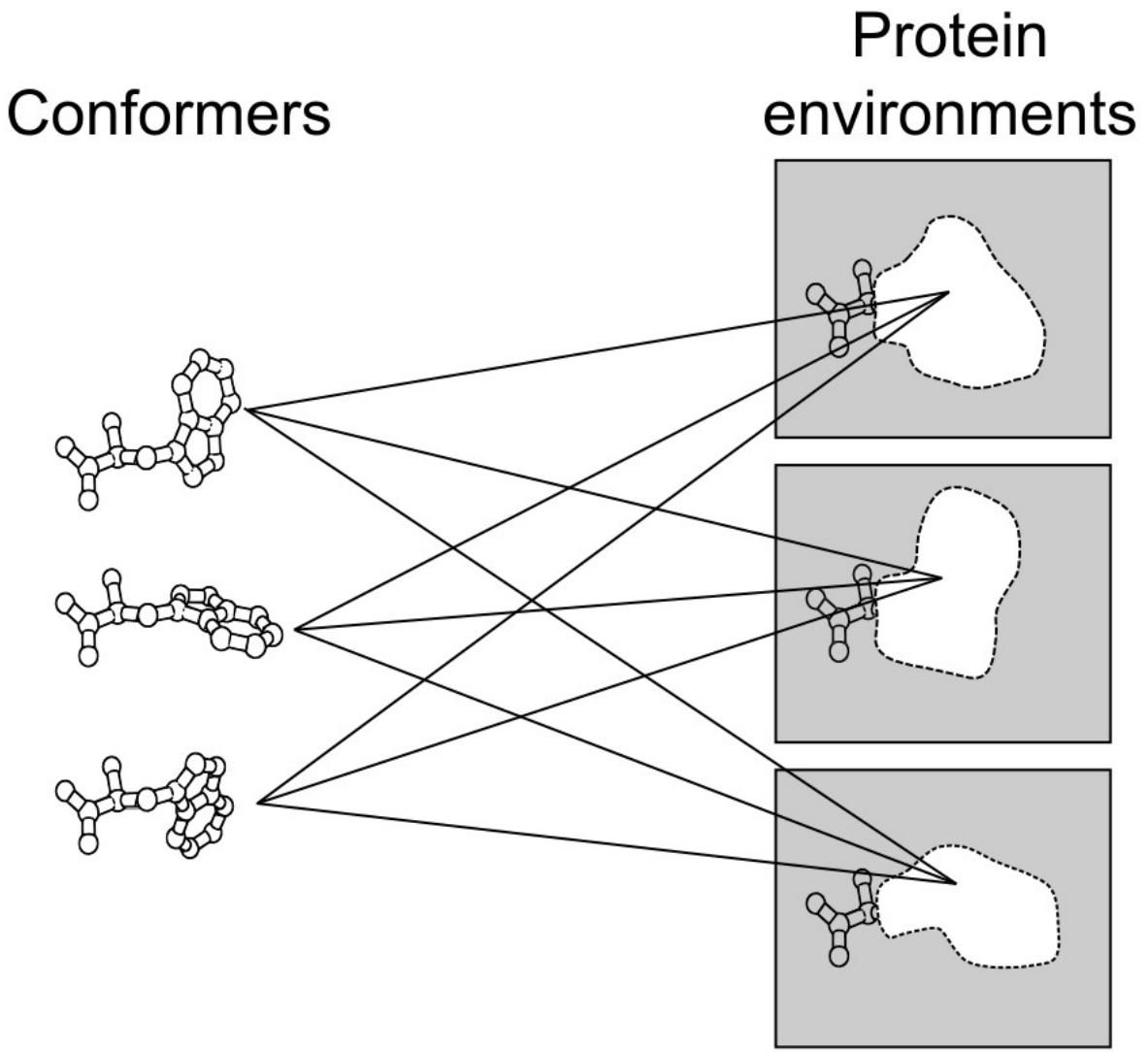
Conformers



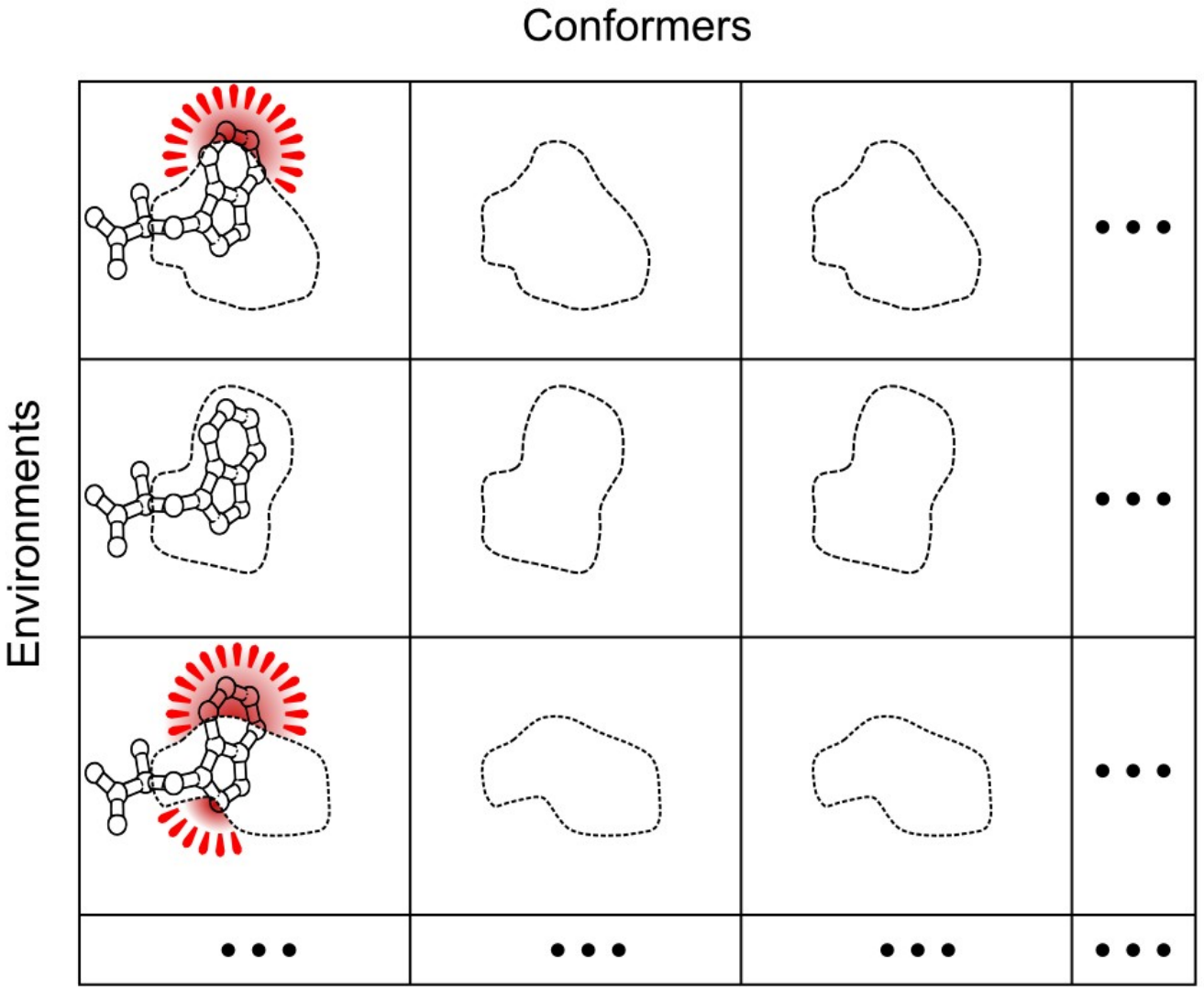
- Use energetics to identify the best sampling strategy for side chain optimization



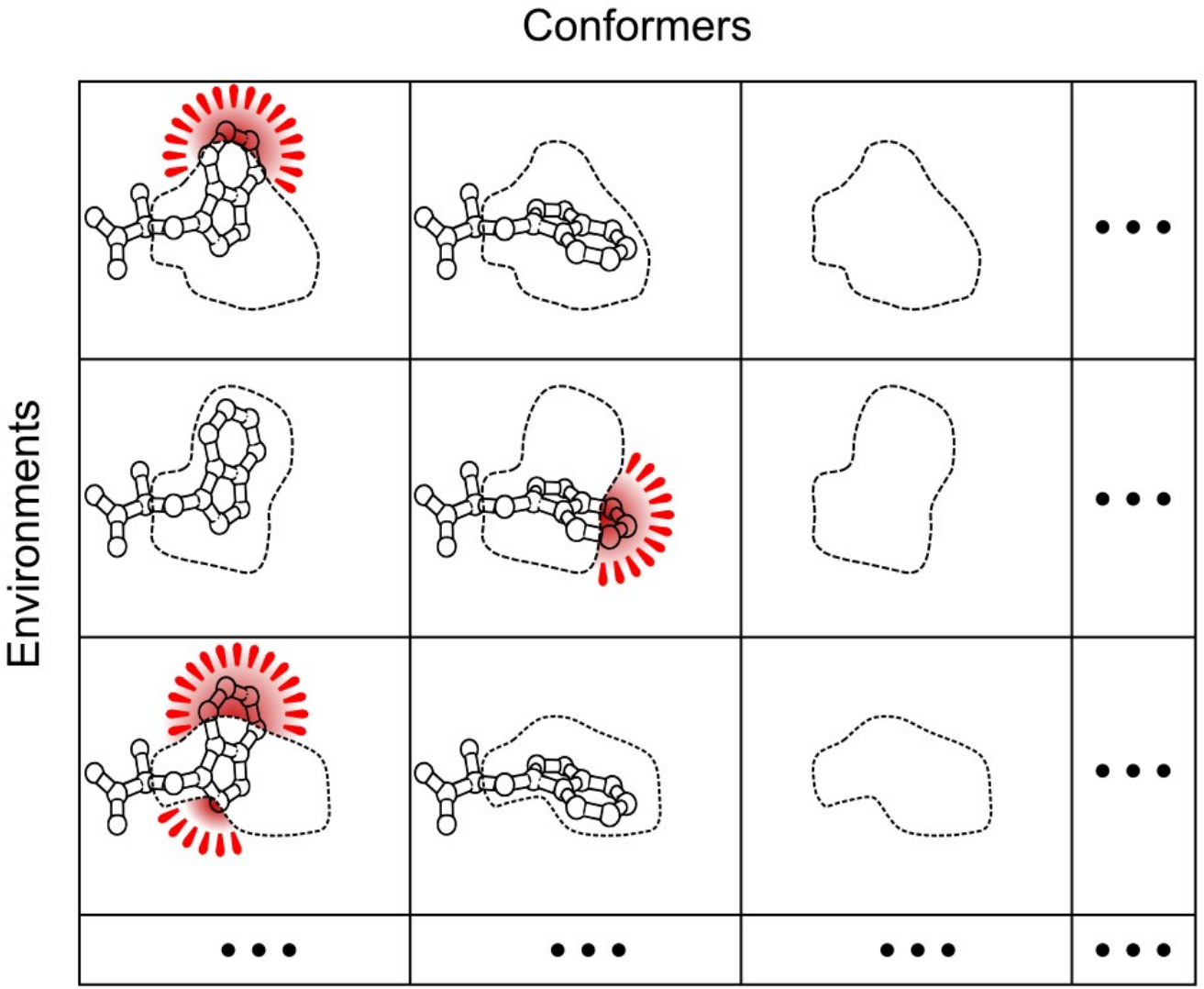
- Use energetics to identify the best sampling strategy for side chain optimization



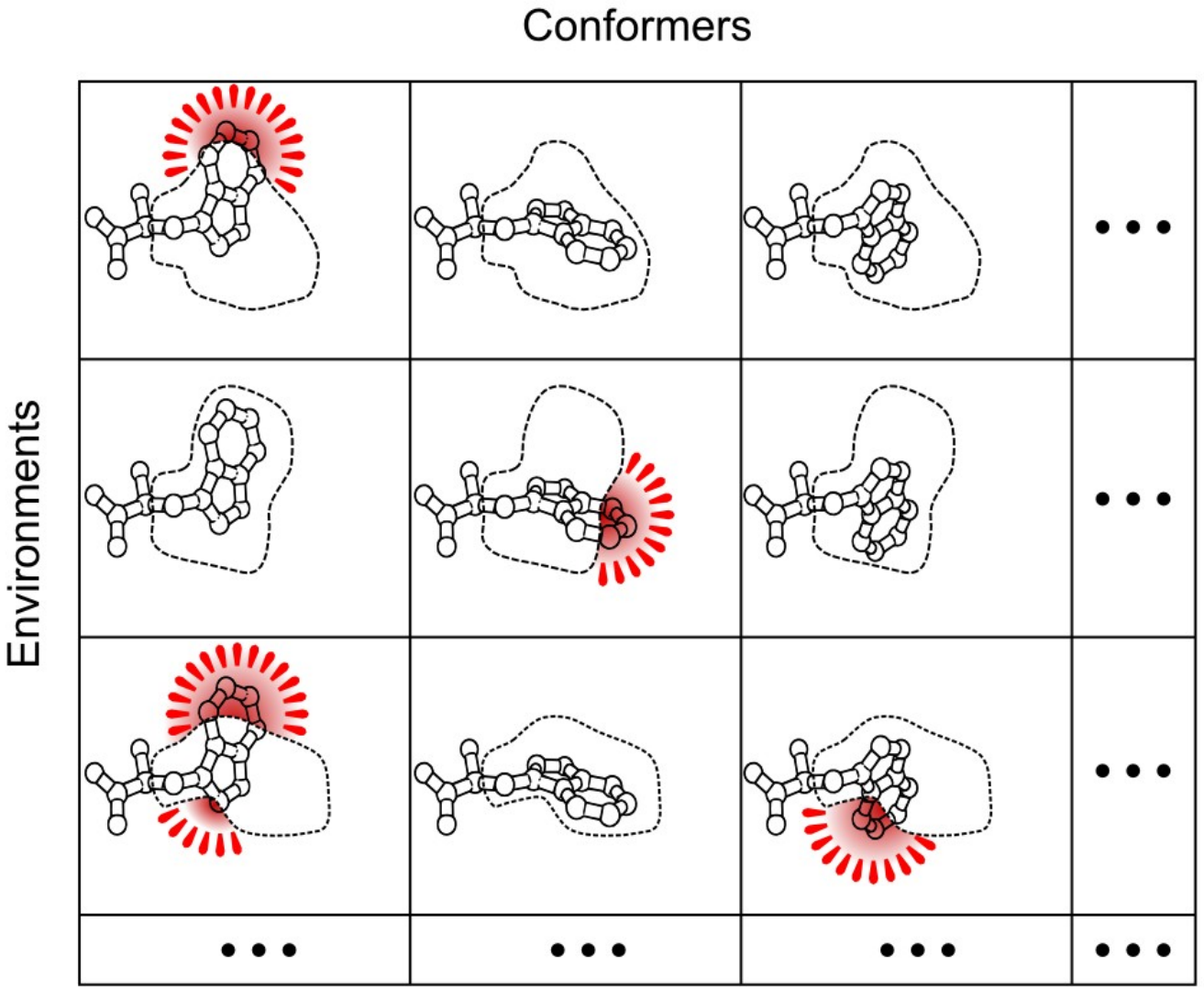
- Use energetics to identify the best sampling strategy for side chain optimization



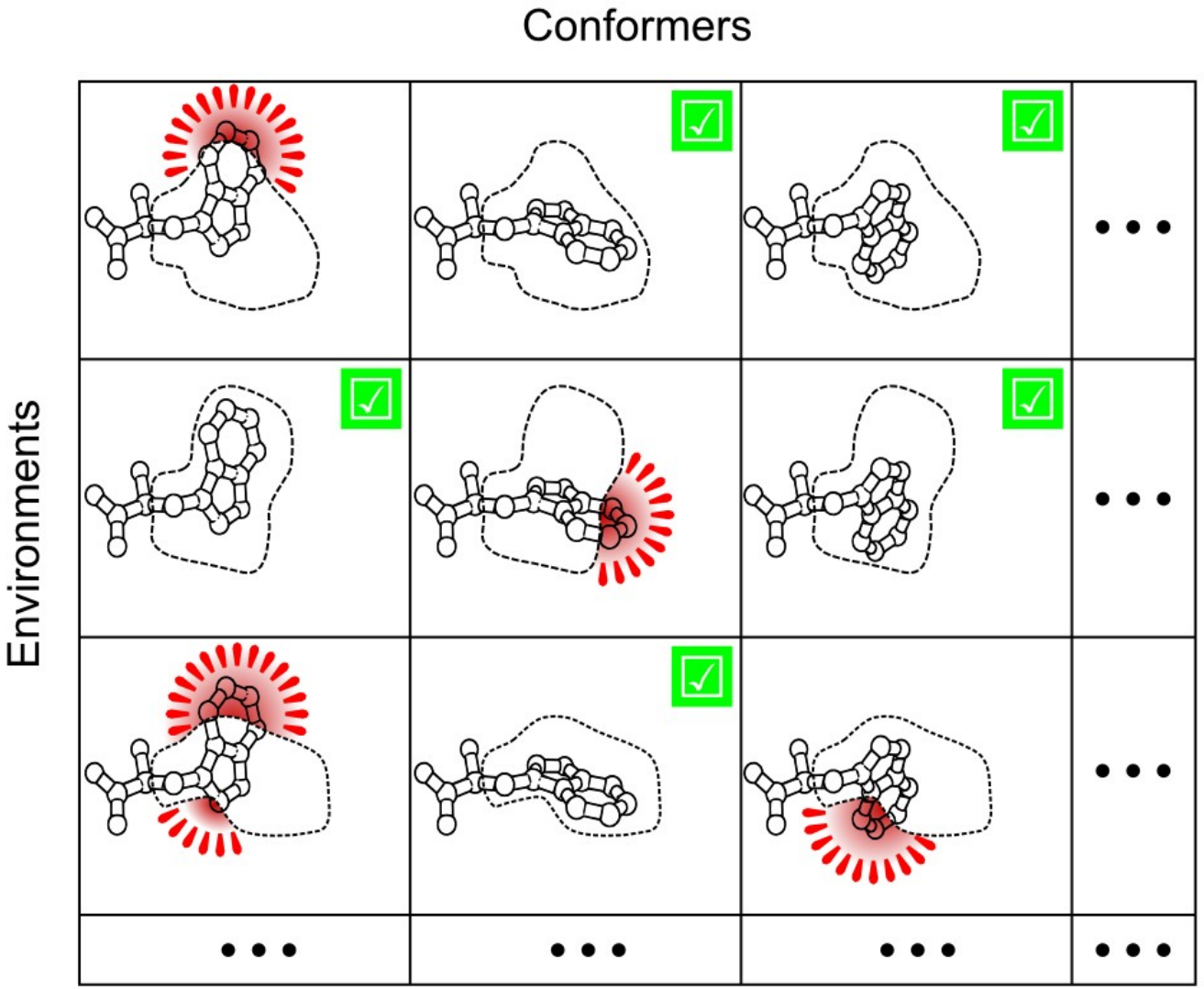
- Use energetics to identify the best sampling strategy for side chain optimization

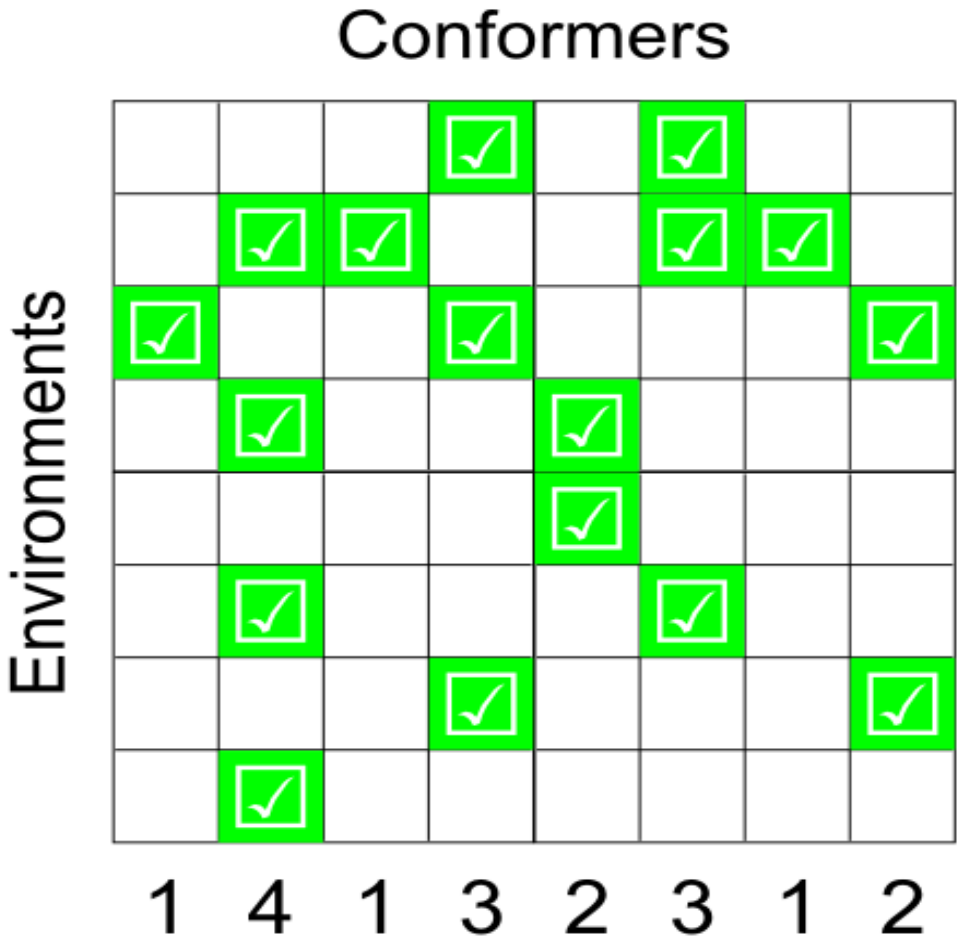


- Use energetics to identify the best sampling strategy for side chain optimization

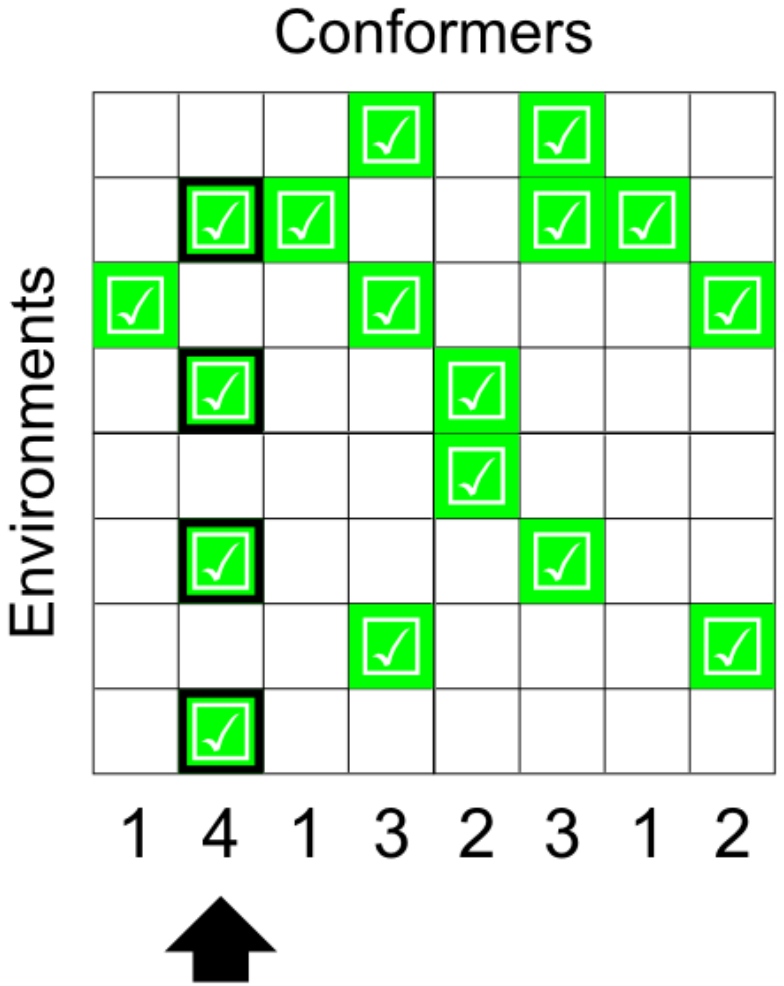


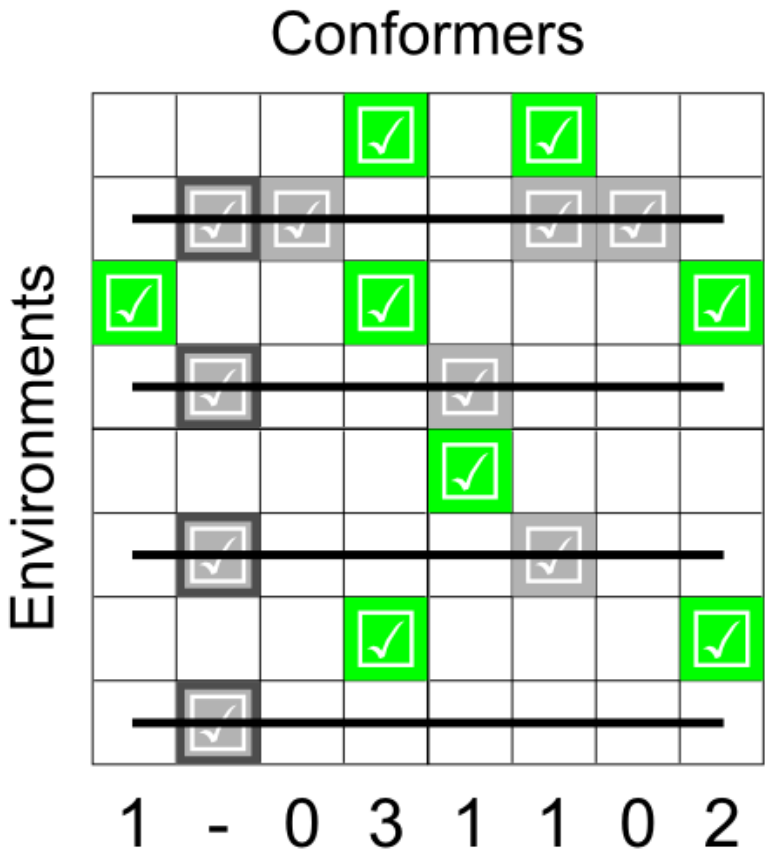
- Use energetics to identify the best sampling strategy for side chain optimization




Conformers
Environments
1 4 1 3 2 3 1 2

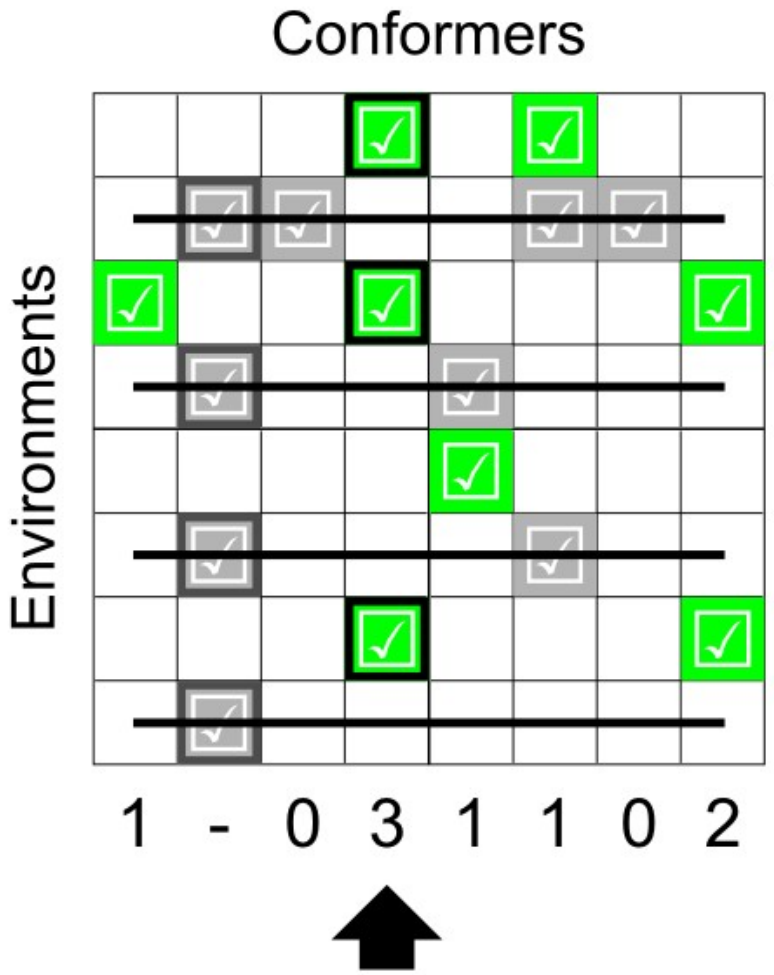
THE first conformer

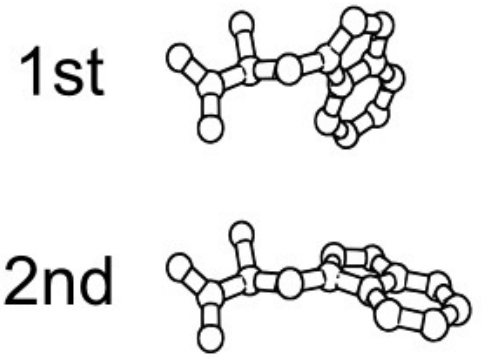




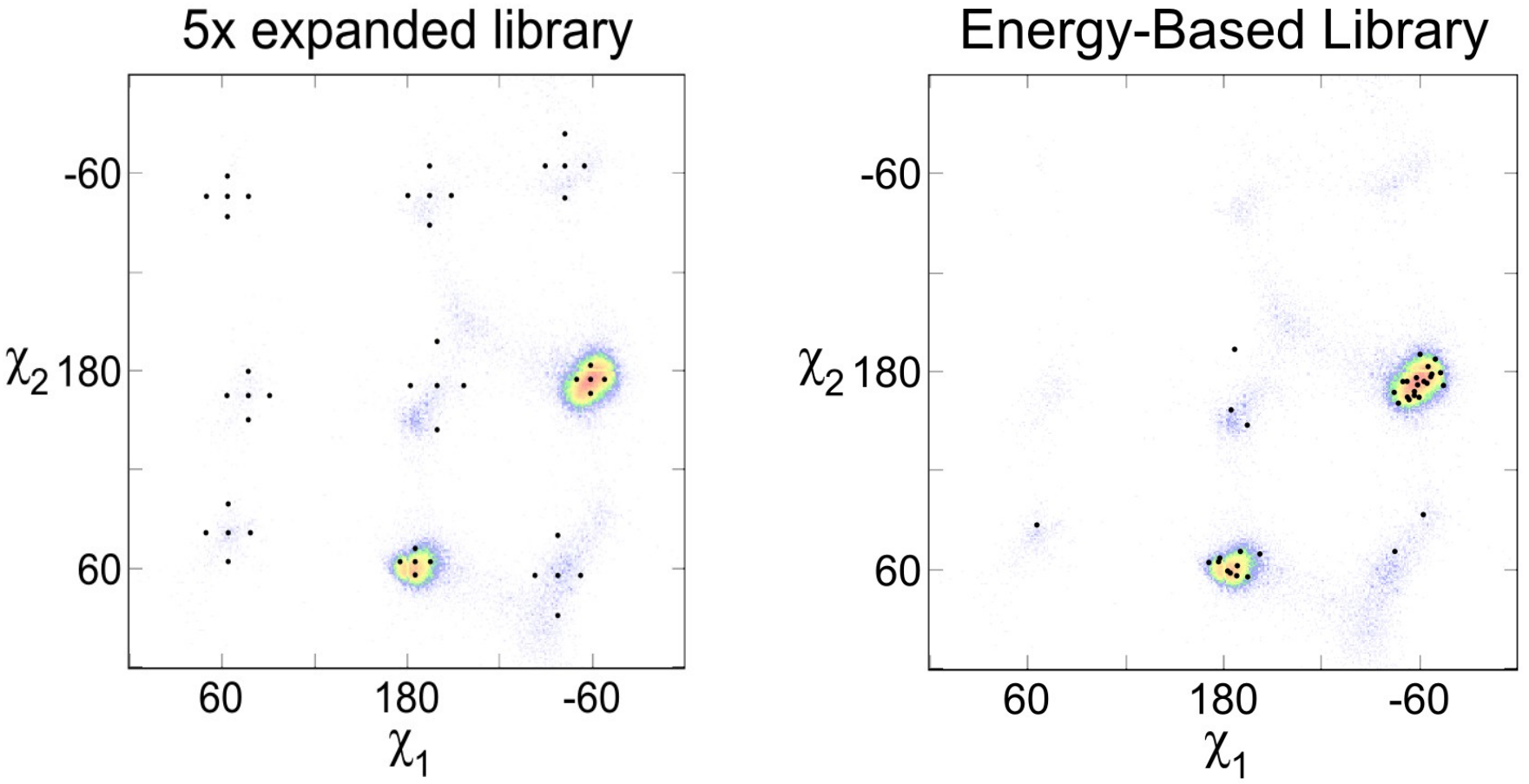
1st

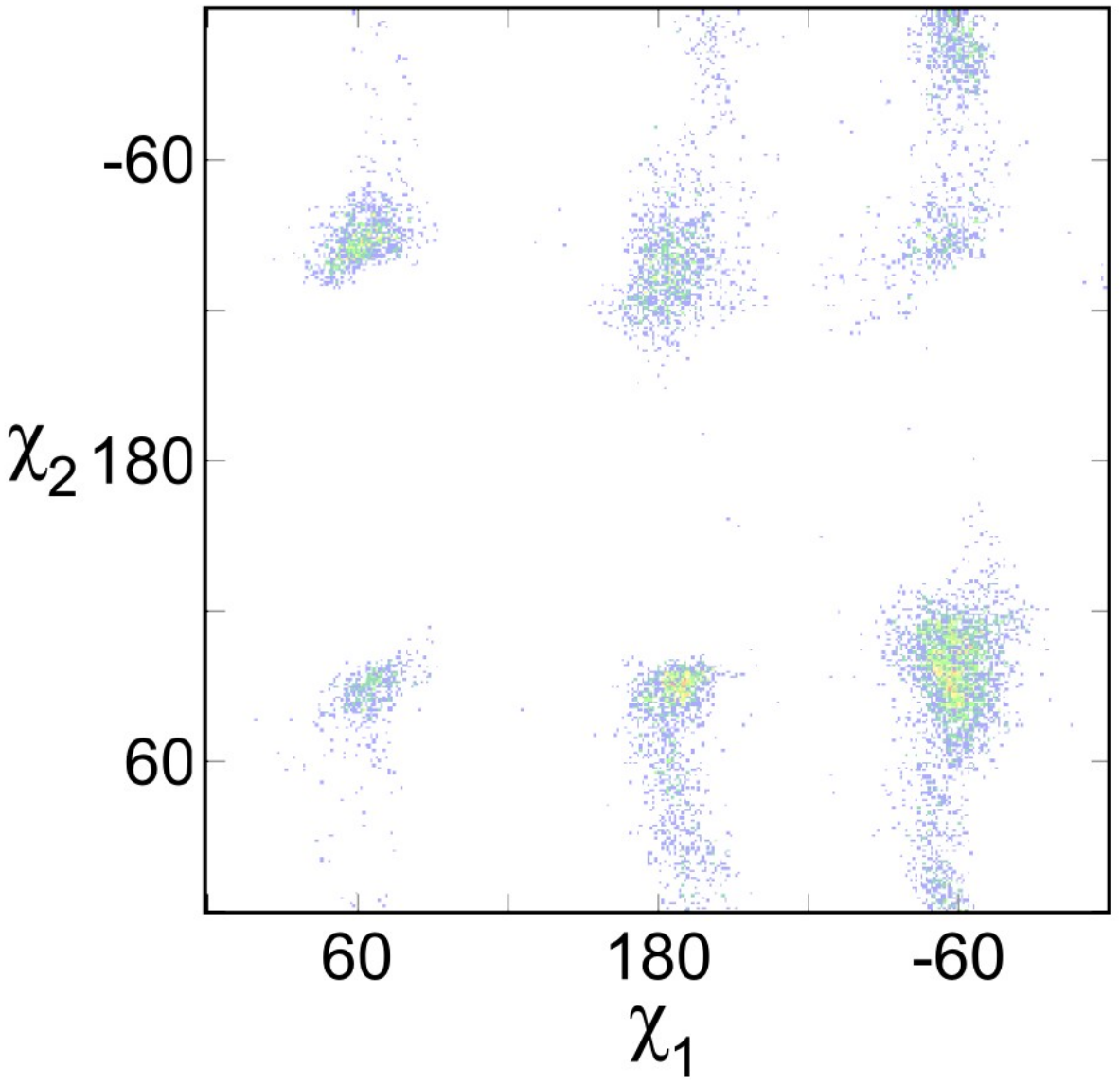
THE second conformer



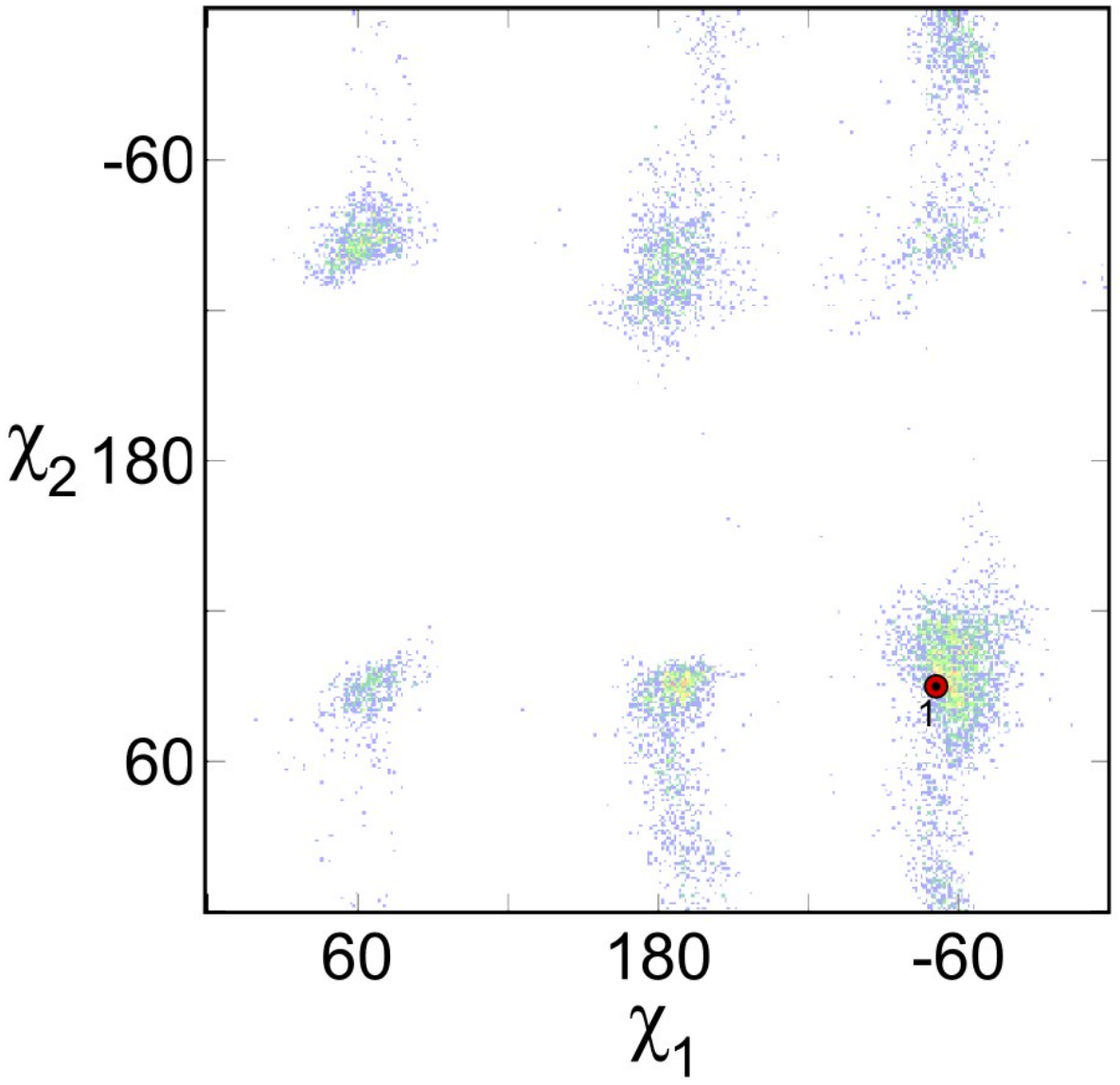


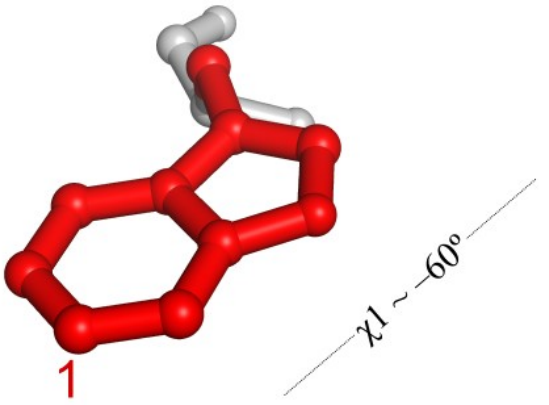
CONFORMERS IN PROPORTION TO DISTRIBUTION

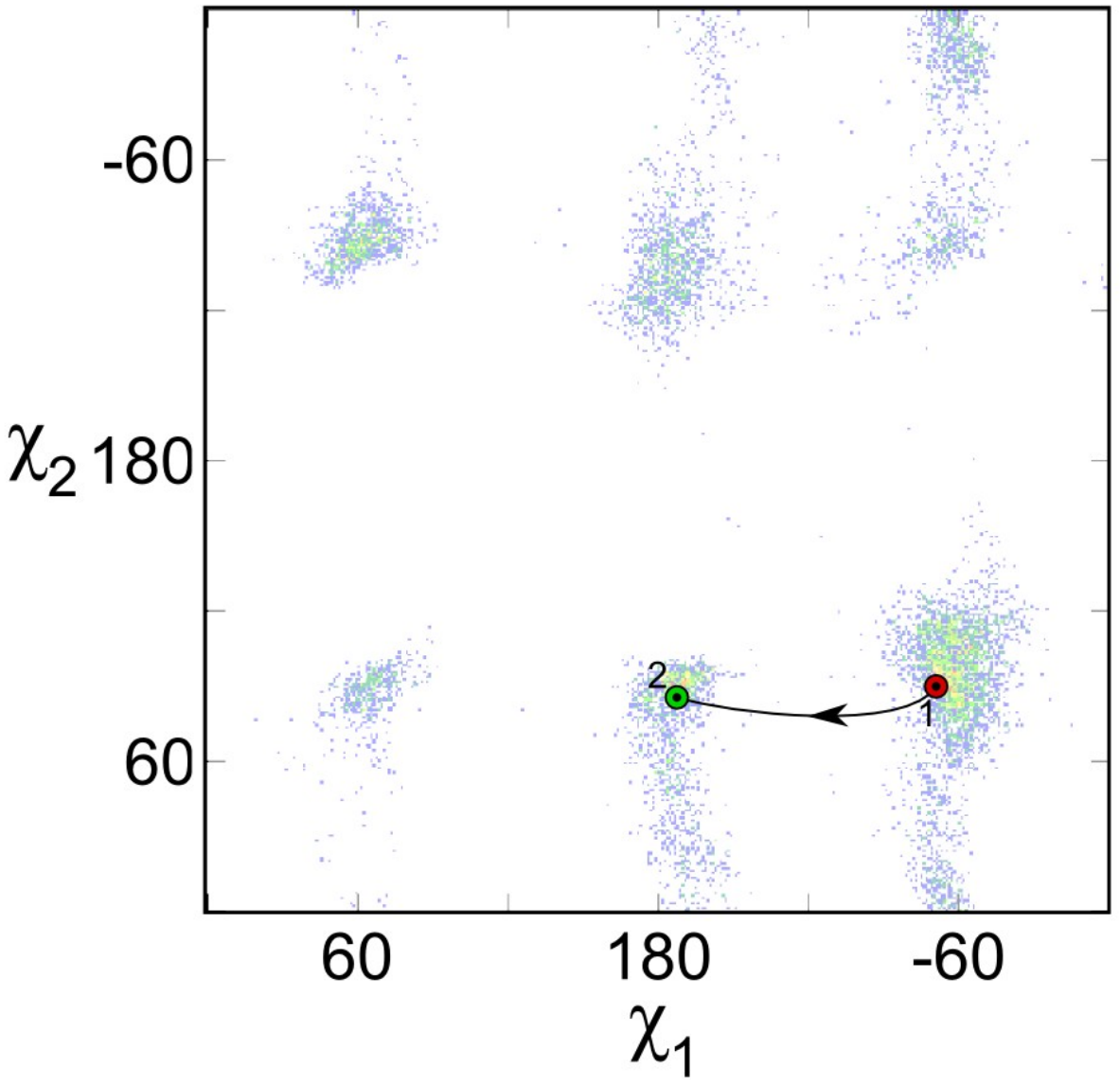


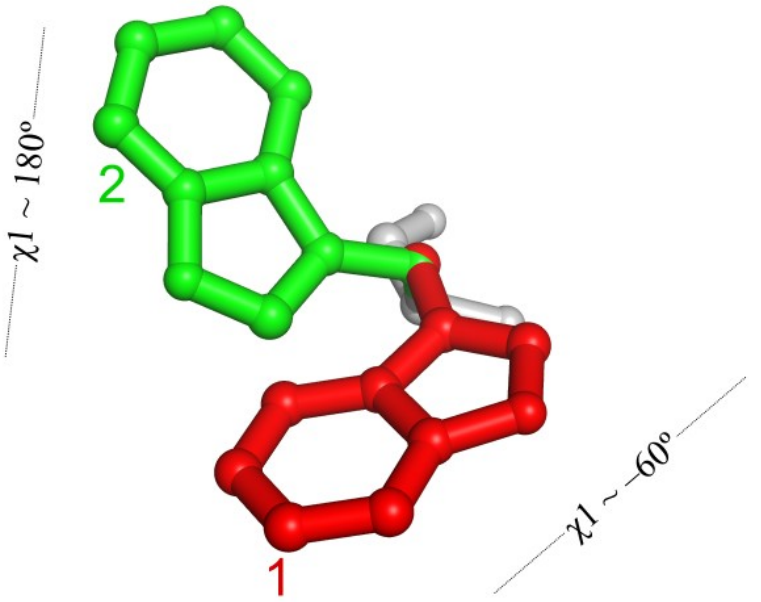


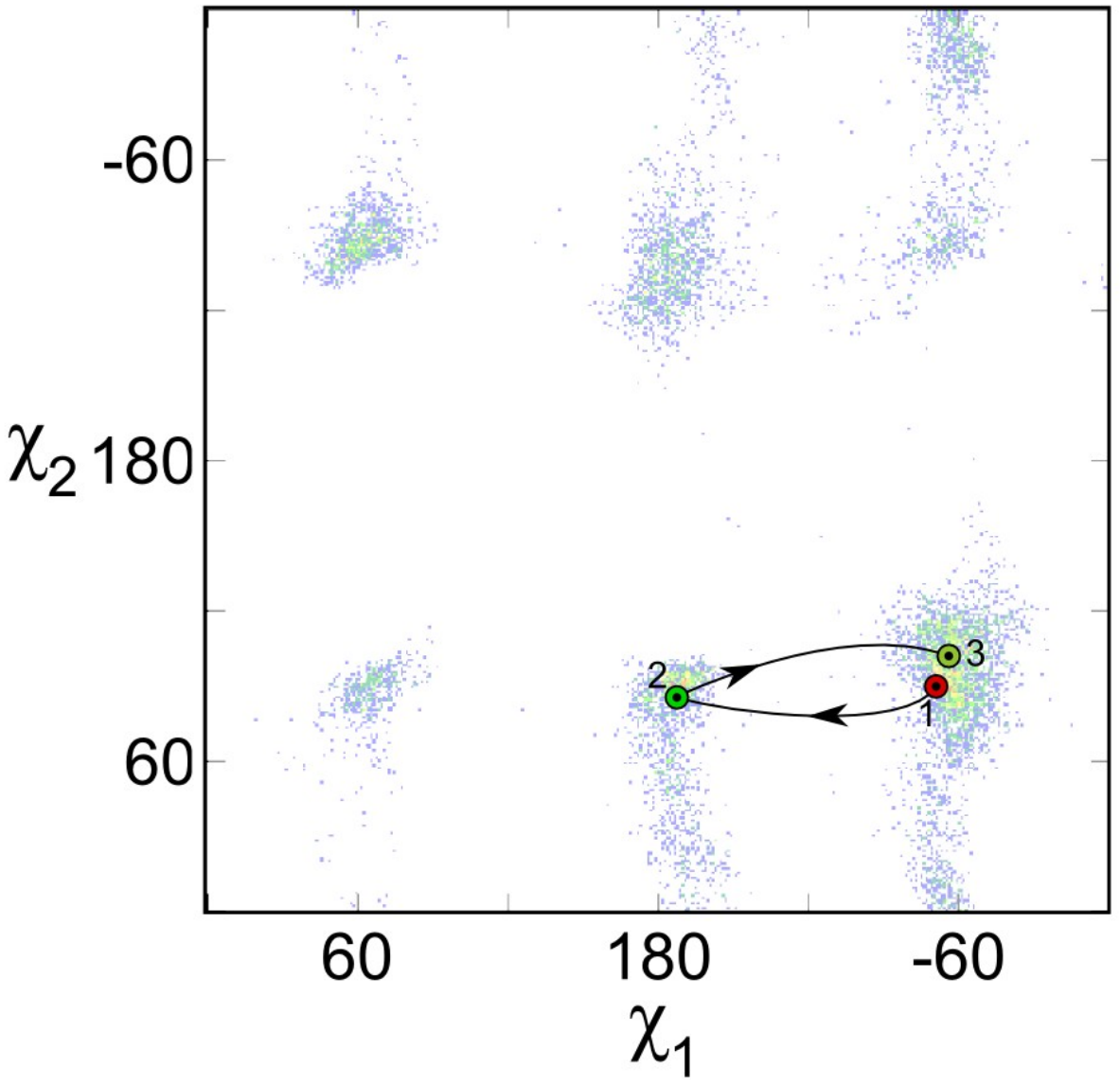
A walk in Trp space

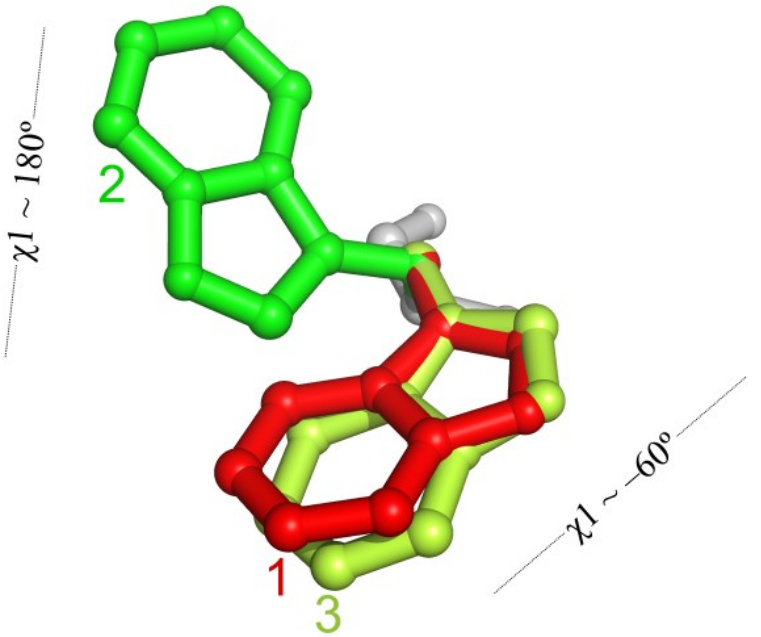

-60
χ_2 180
60
60
180
-60
χ_1
1

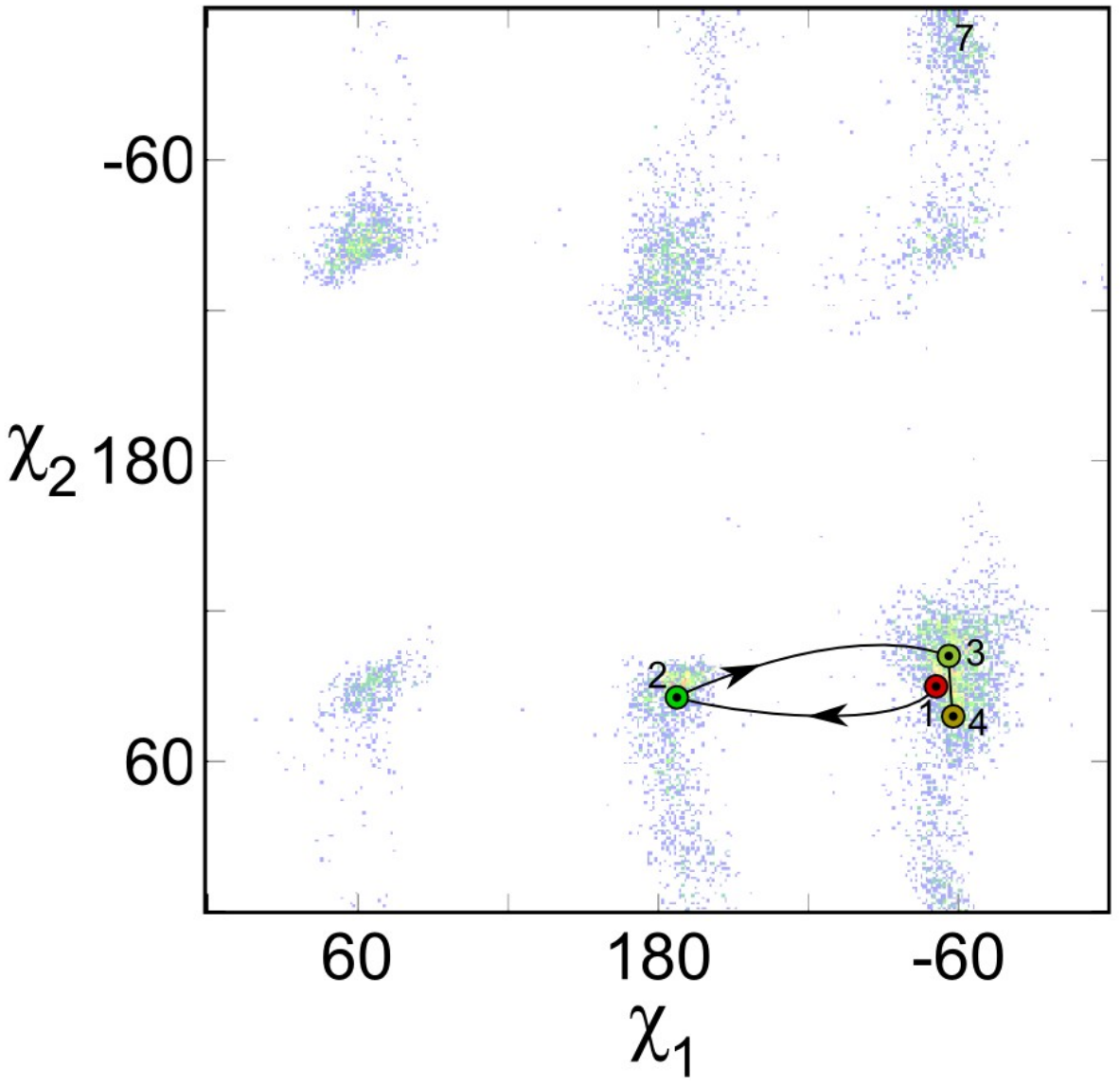

1
χ1 ~ -60°

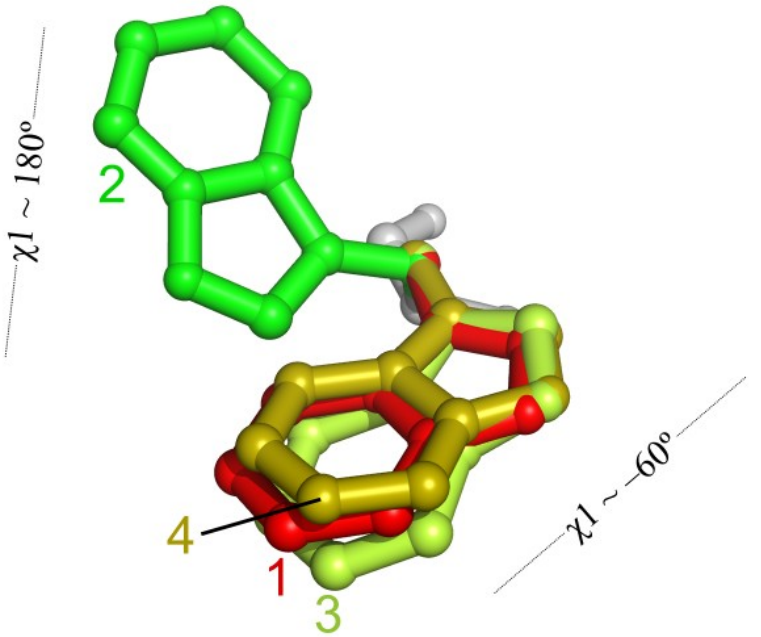

-60
χ_2 180
60
60
180
-60
χ_1
2
1

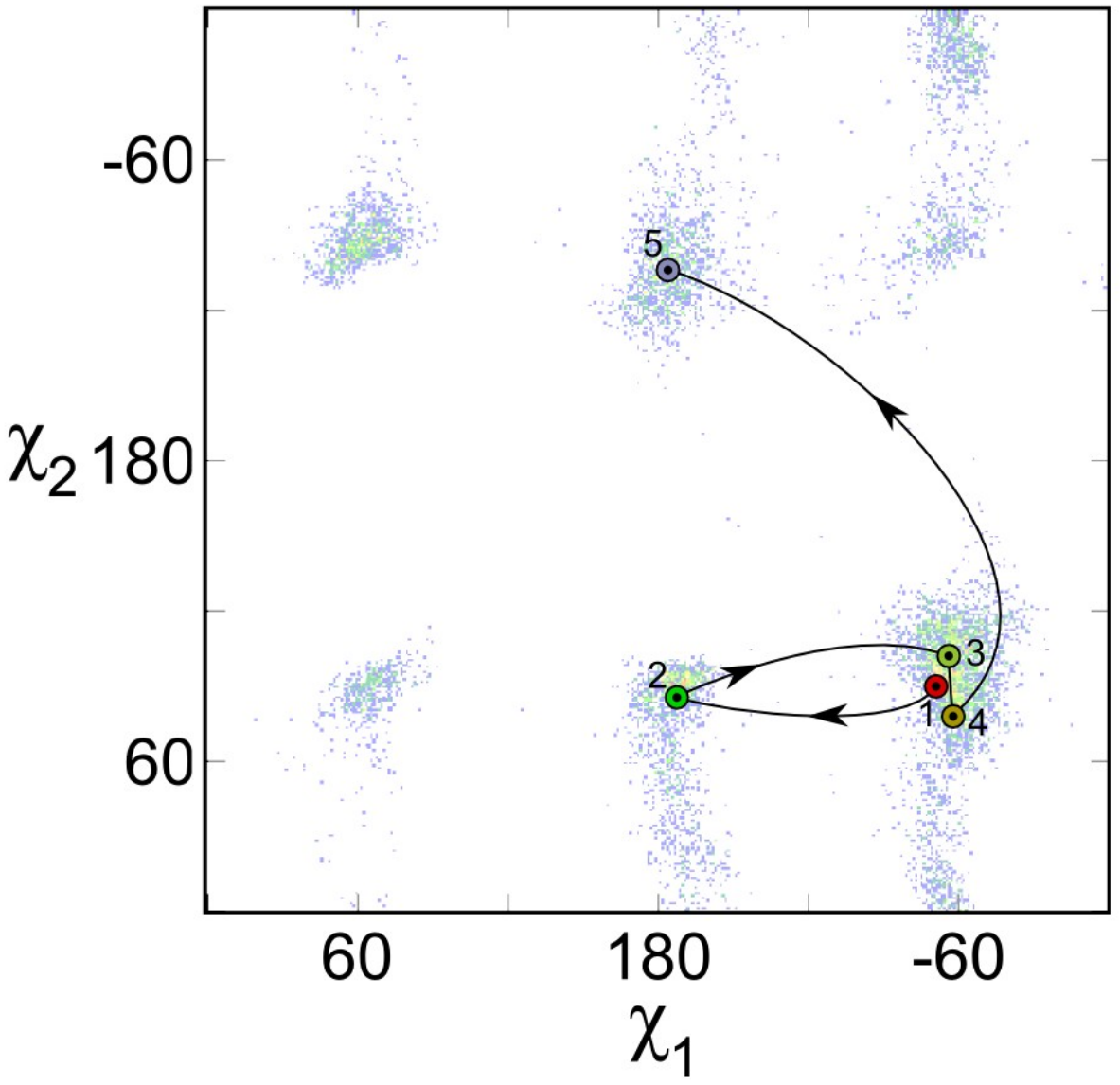

χ1 ~ 180°
2
1
χ1 ~ −60°

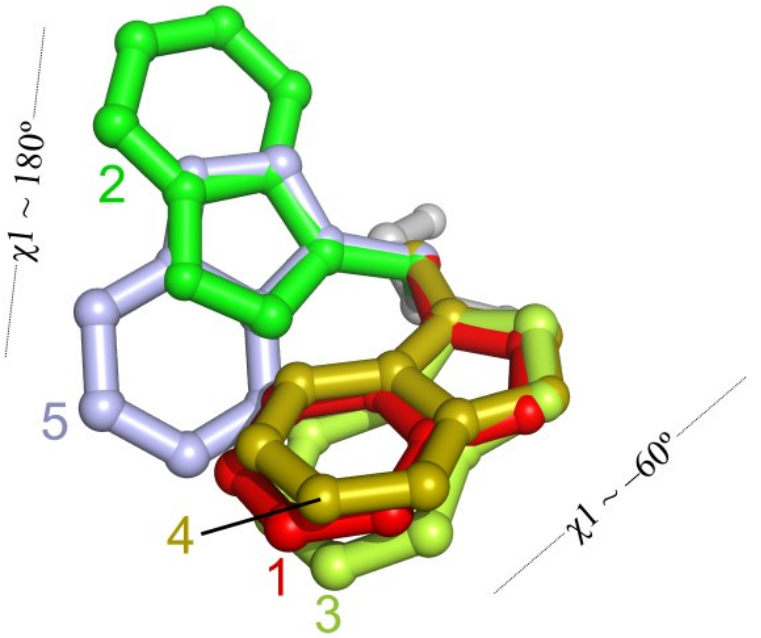

-60
χ_2 180
60
60
180
-60
χ_1
2
3
1

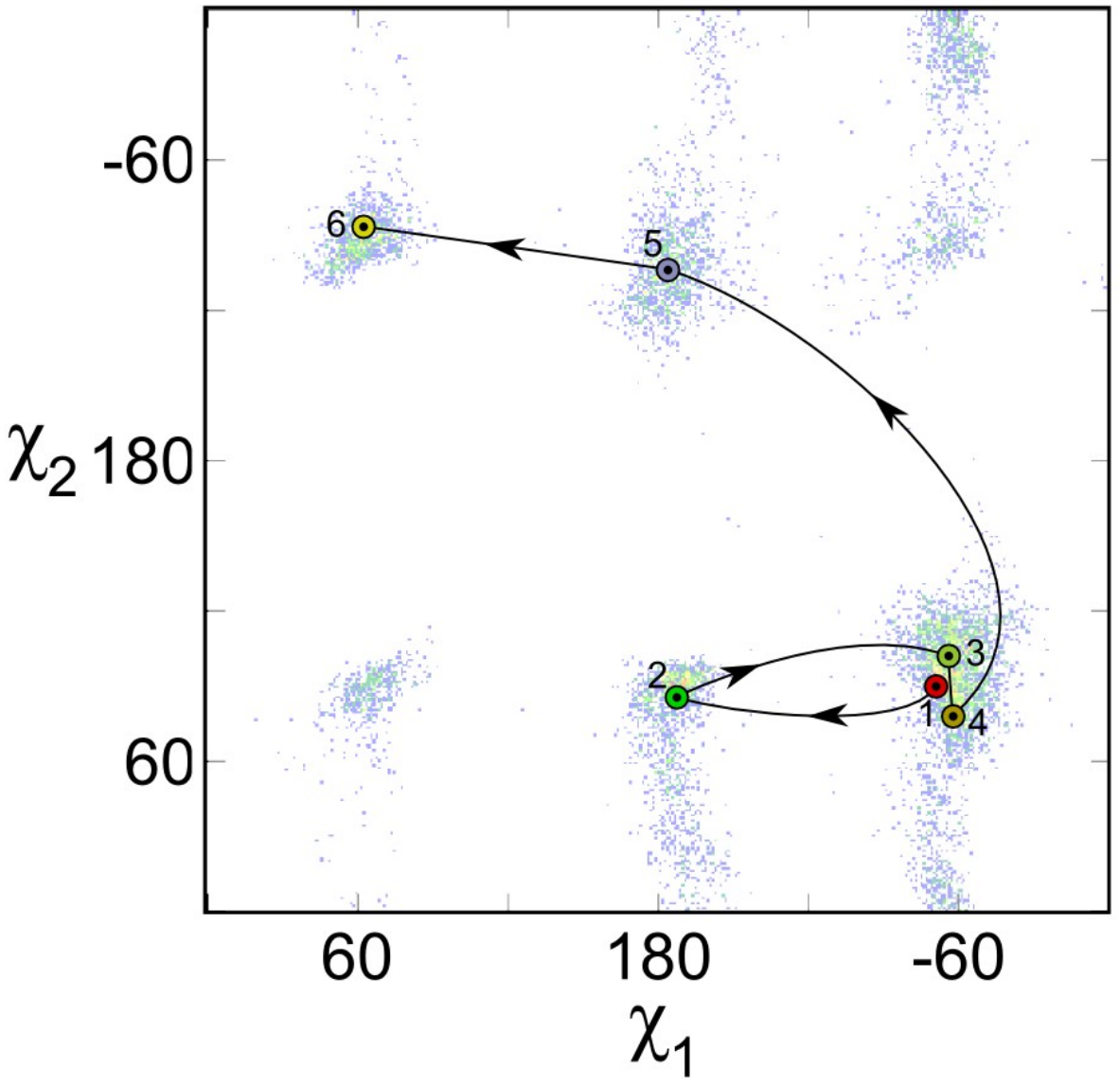

χ1 ~ 180°
2
1
3
χ1 ~ −60°

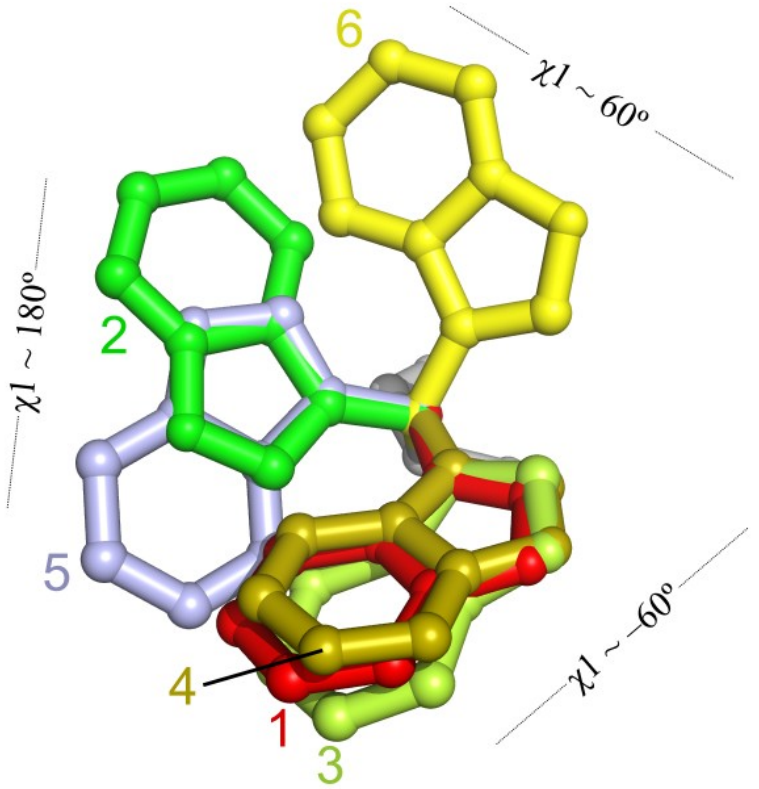

7
-60
χ_2 180
60
2
3
1
4
60
180
-60
χ_1

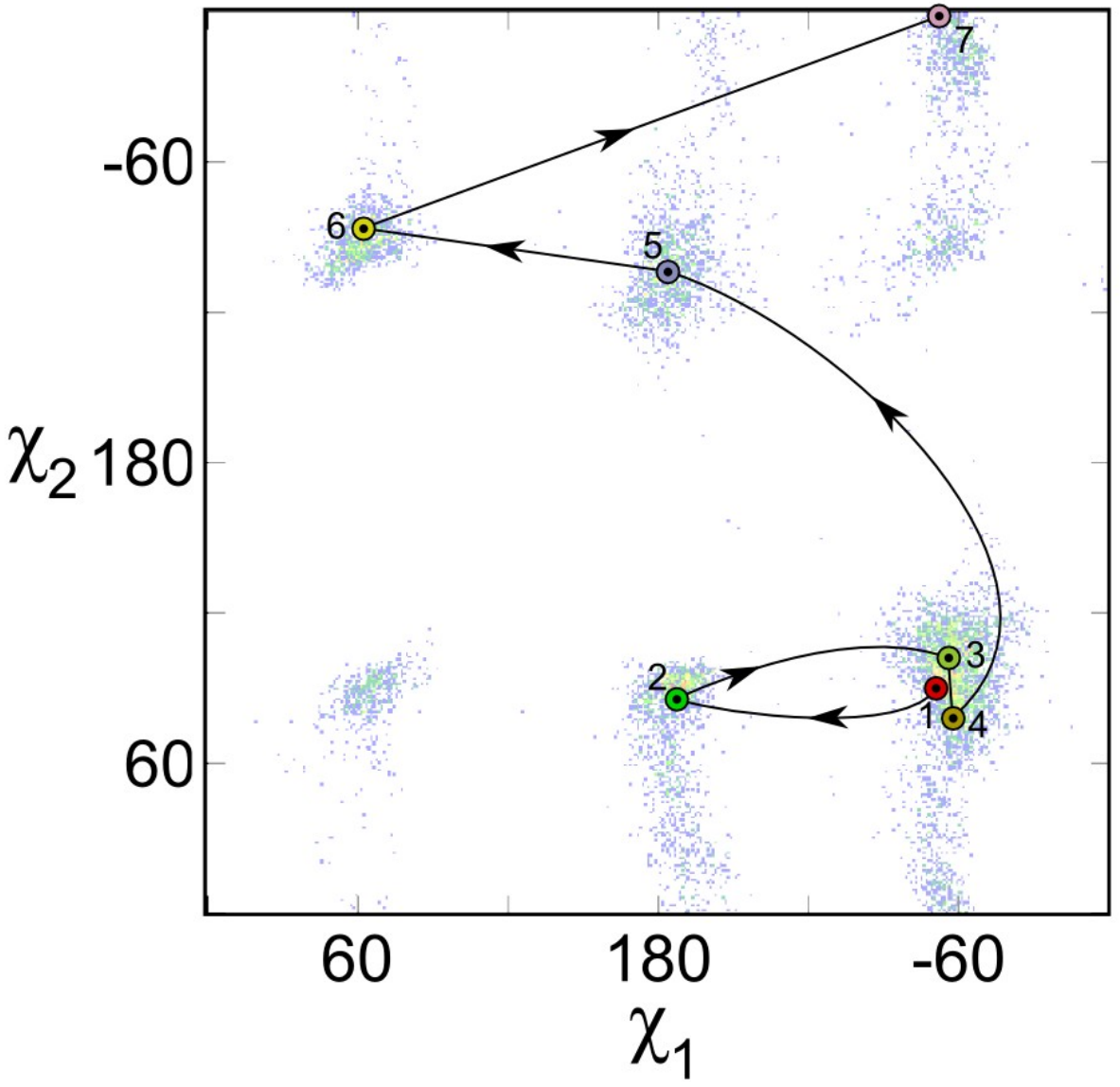

χ1 ~ 180°
2
4
1
3
χ1 ~ −60°

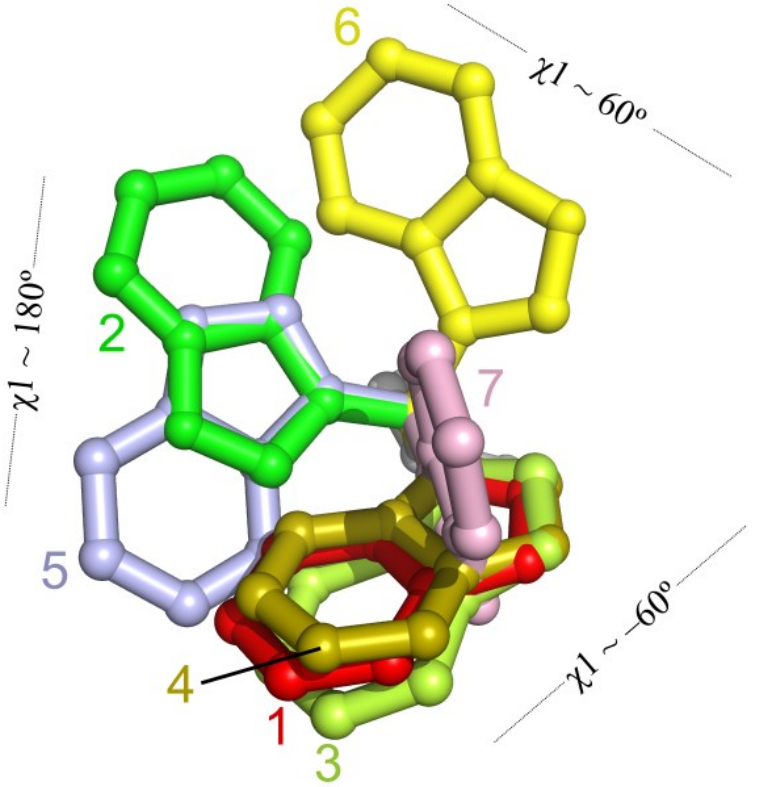

-60
χ_2 180
60
60
180
-60
χ_1
5
2
3
1
4

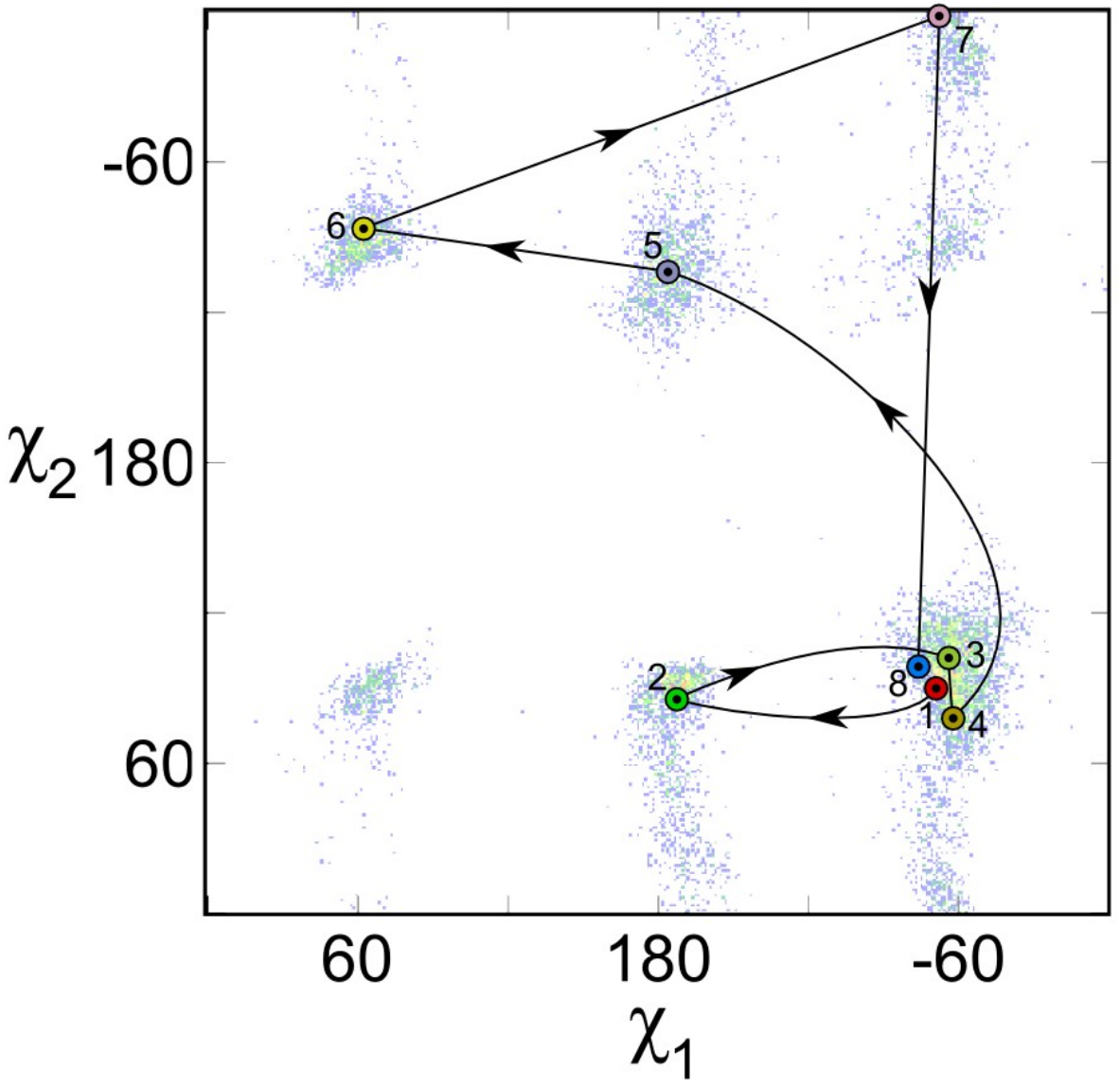

χ1 ~ 180°
2
5
4
1
3
χ1 ~ −60°

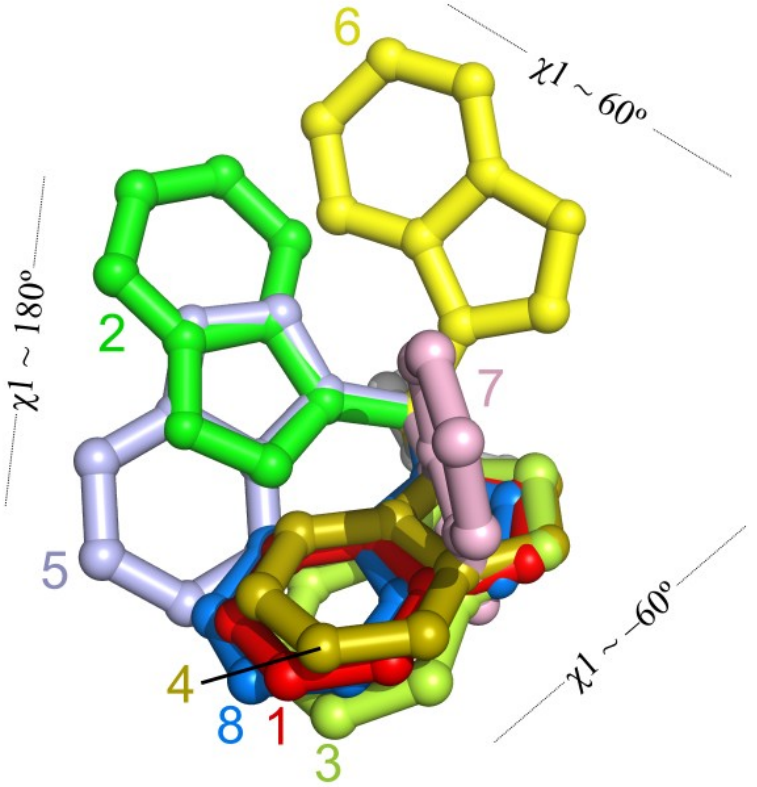

-60
60
χ2 180
60
180
-60
χ1
6
5
2
3
1
4

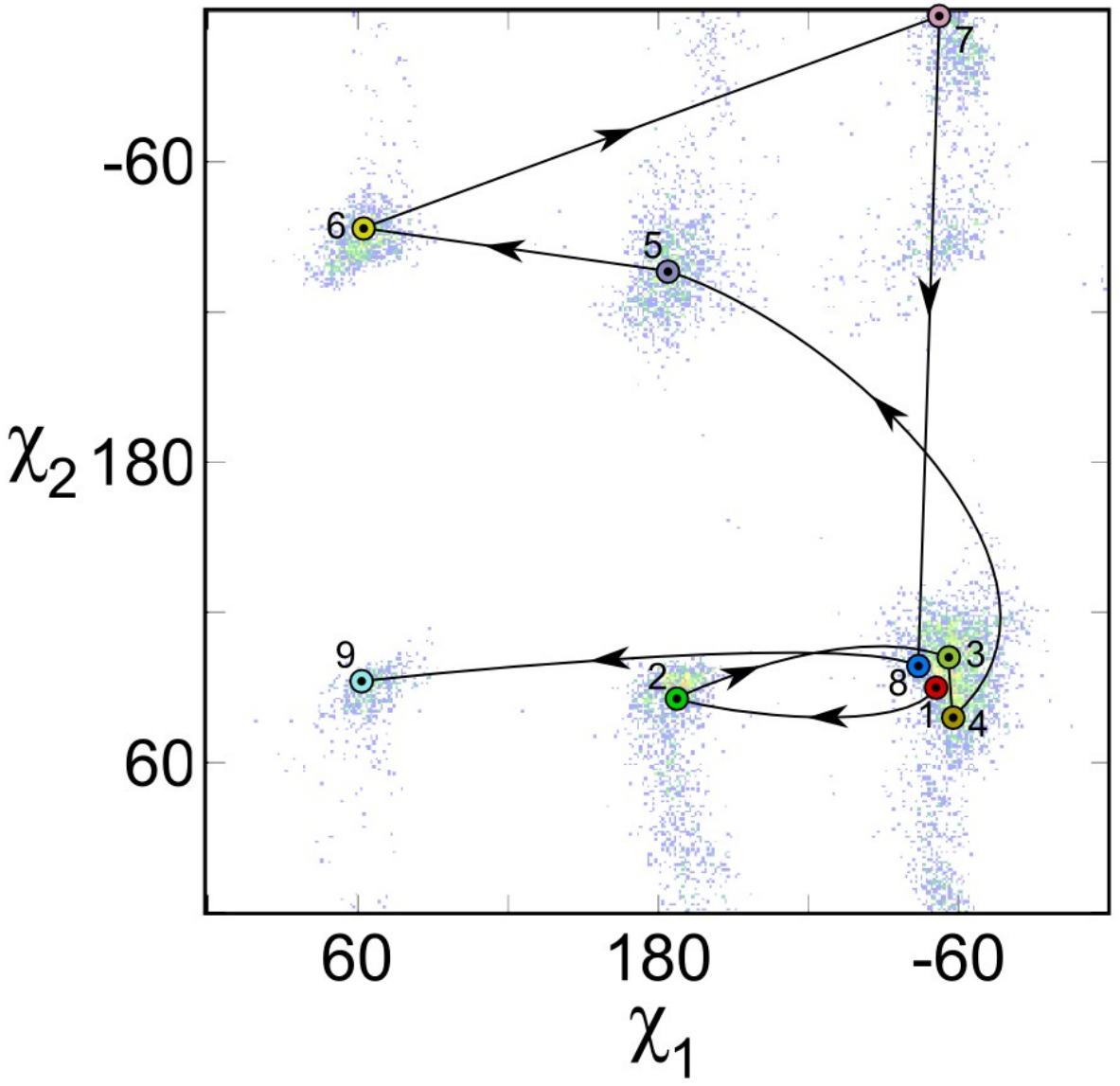

χ1 ~ 60°
χ1 ~ 180°
χ1 ~ -60°
6
2
5
4
1
3

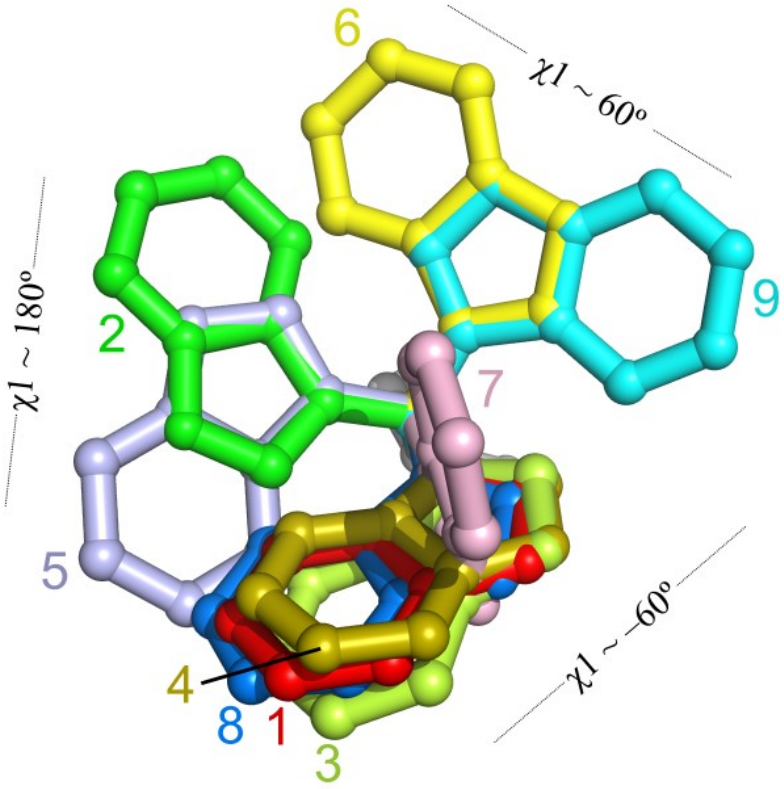

-60
χ_2 180
60
60
180
-60
χ_1
6
5
7
2
3
1
4

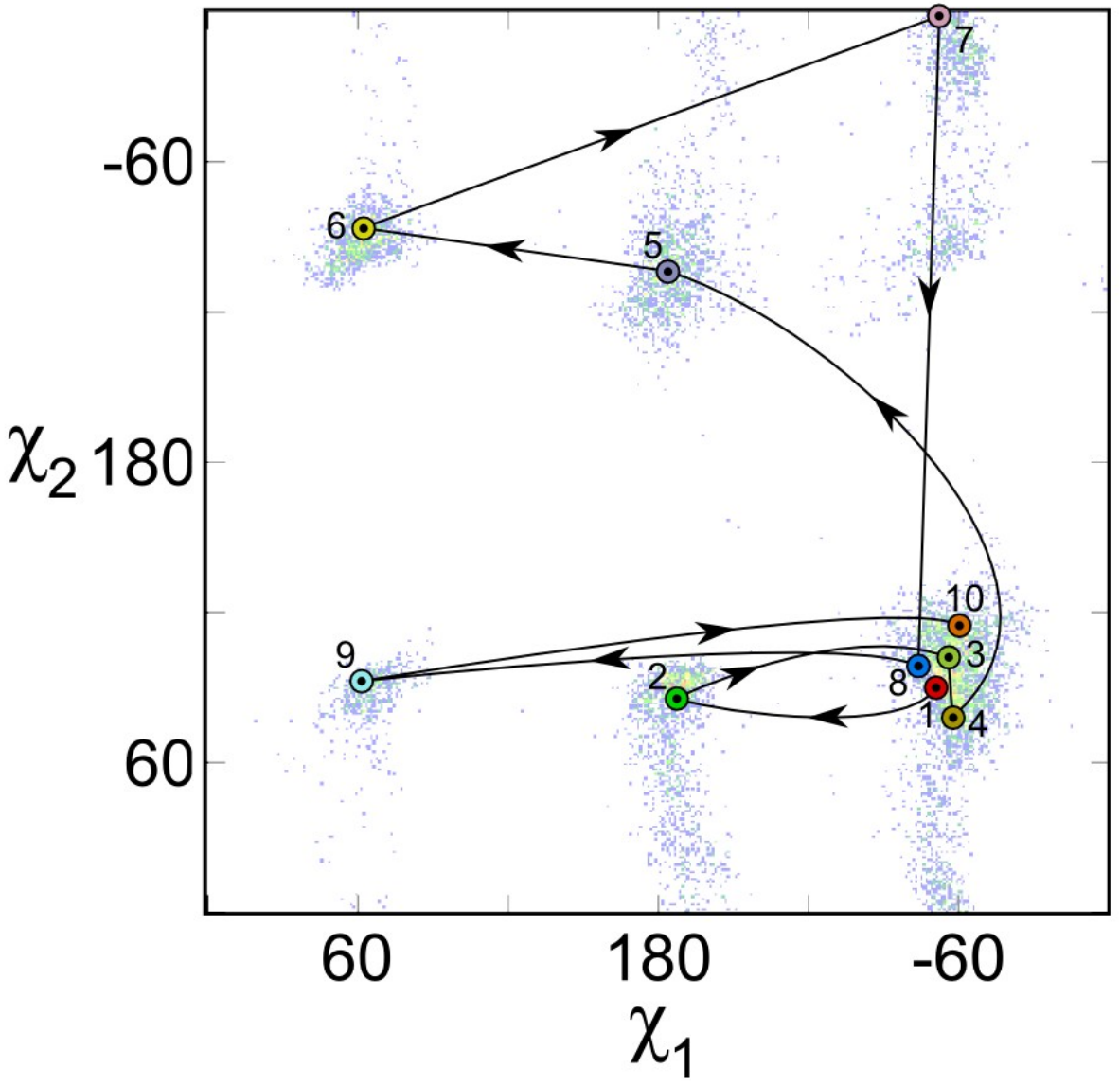

6
χ1 ~ 60°
χ1 ~ 180°
2
7
5
4
1
3
χ1 ~ -60°

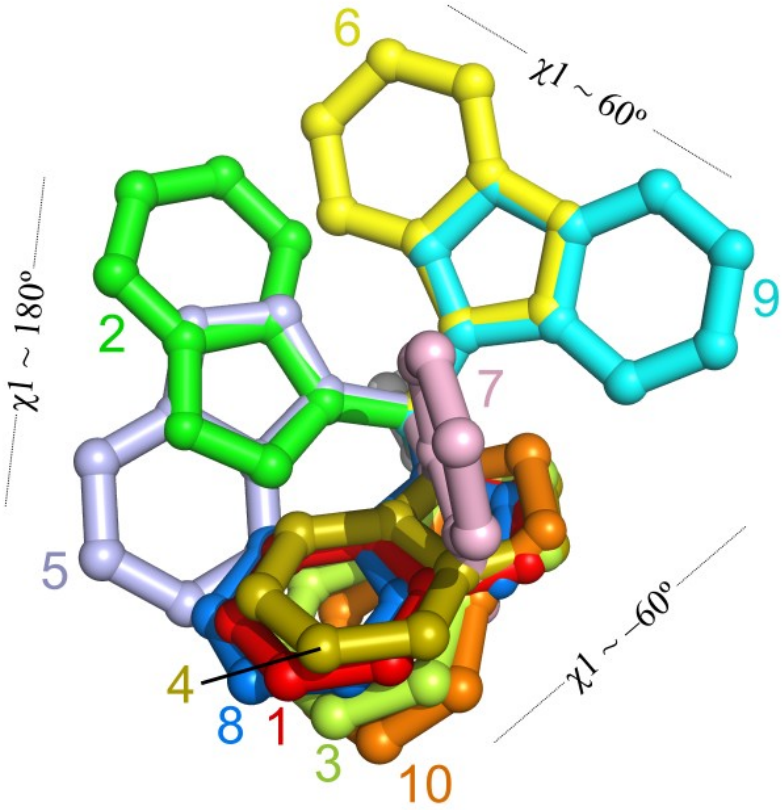

-60
χ_2 180
60
60
180
-60
χ_1
6
5
7
2
8
3
1
4

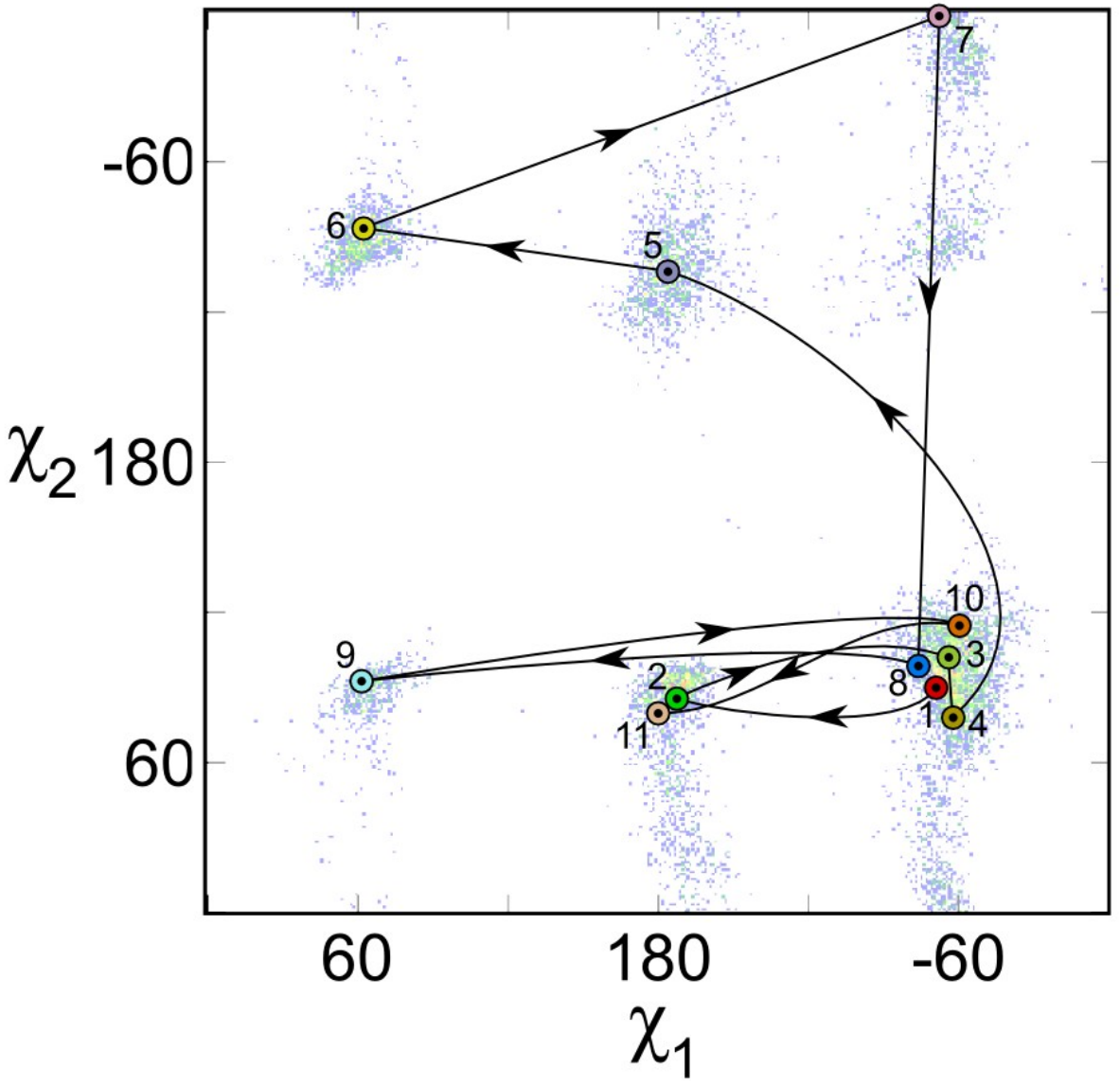

6
χ1 ~ 60°
χ1 ~ 180°
2
7
5
4
8
1
3
χ1 ~ -60°

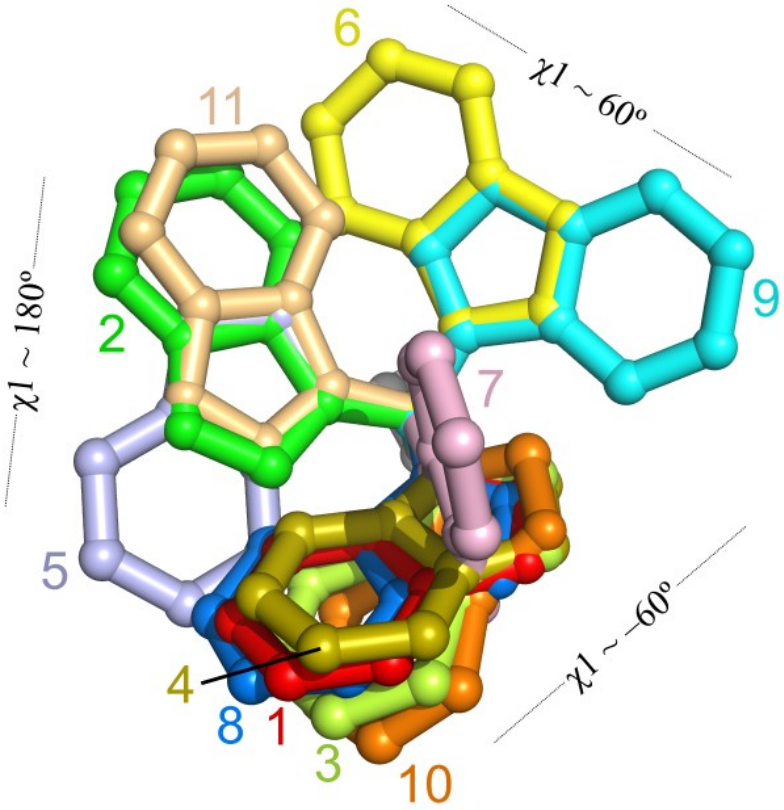

-60
χ_2 180
60
60
180
-60
χ_1
6
5
7
9
2
8
3
1
4

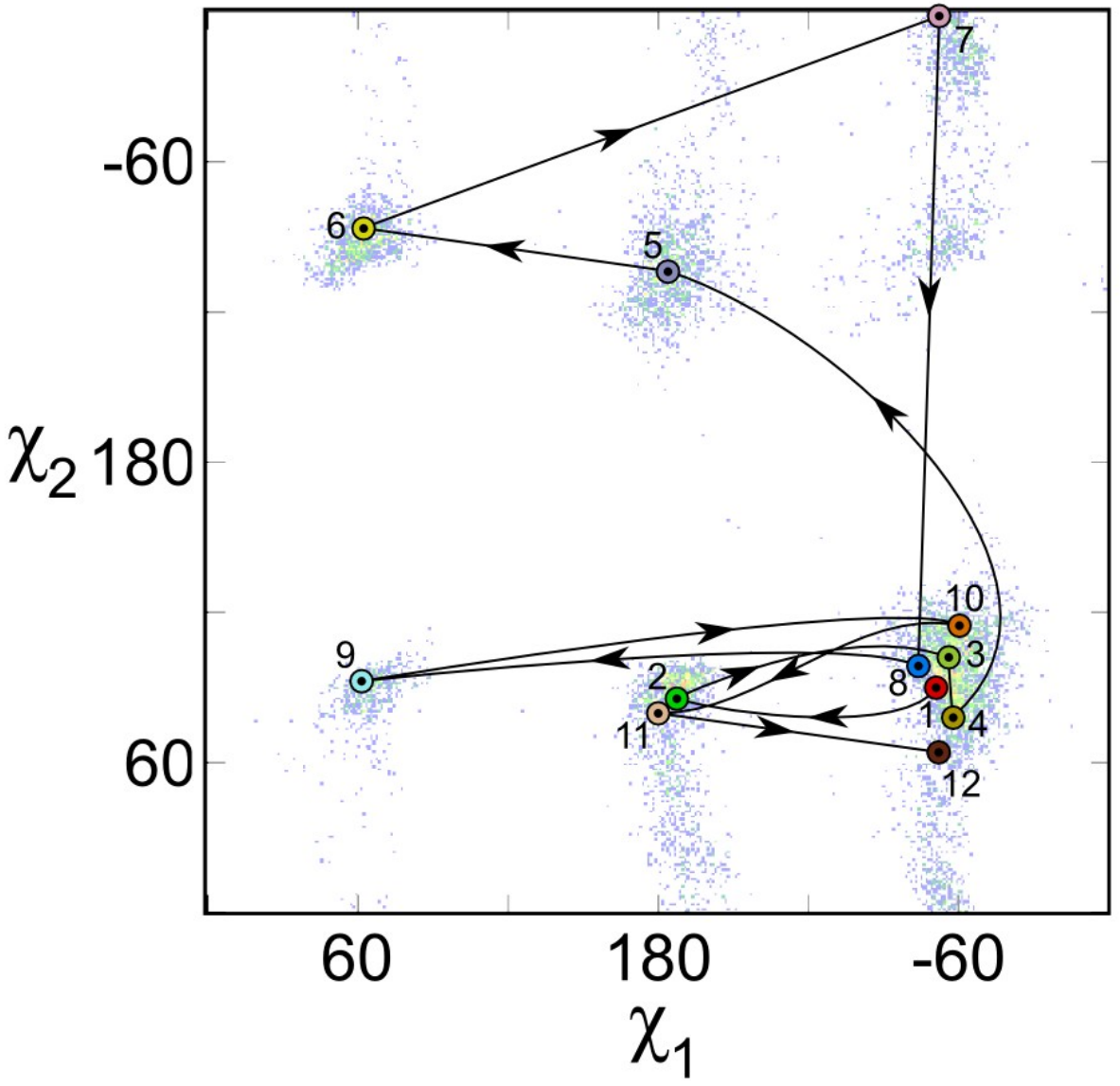

6
χ1 ~ 60°
9
χ1 ~ 180°
2
7
5
χ1 ~ -60°
4
8
1
3

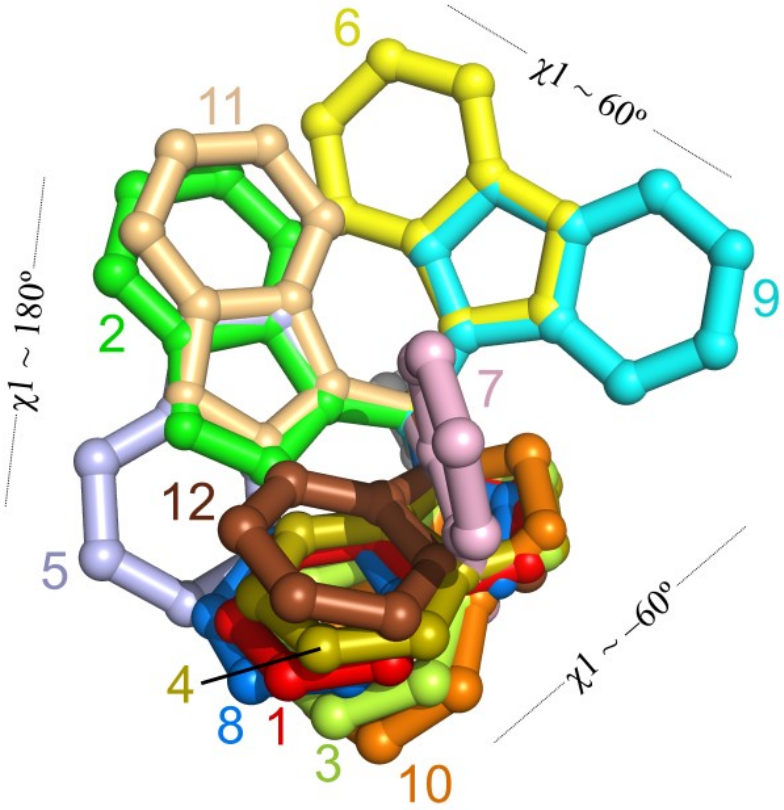

-60
χ_2 180
60
60
180
-60
χ_1
6
5
7
9
2
10
3
8
1
4

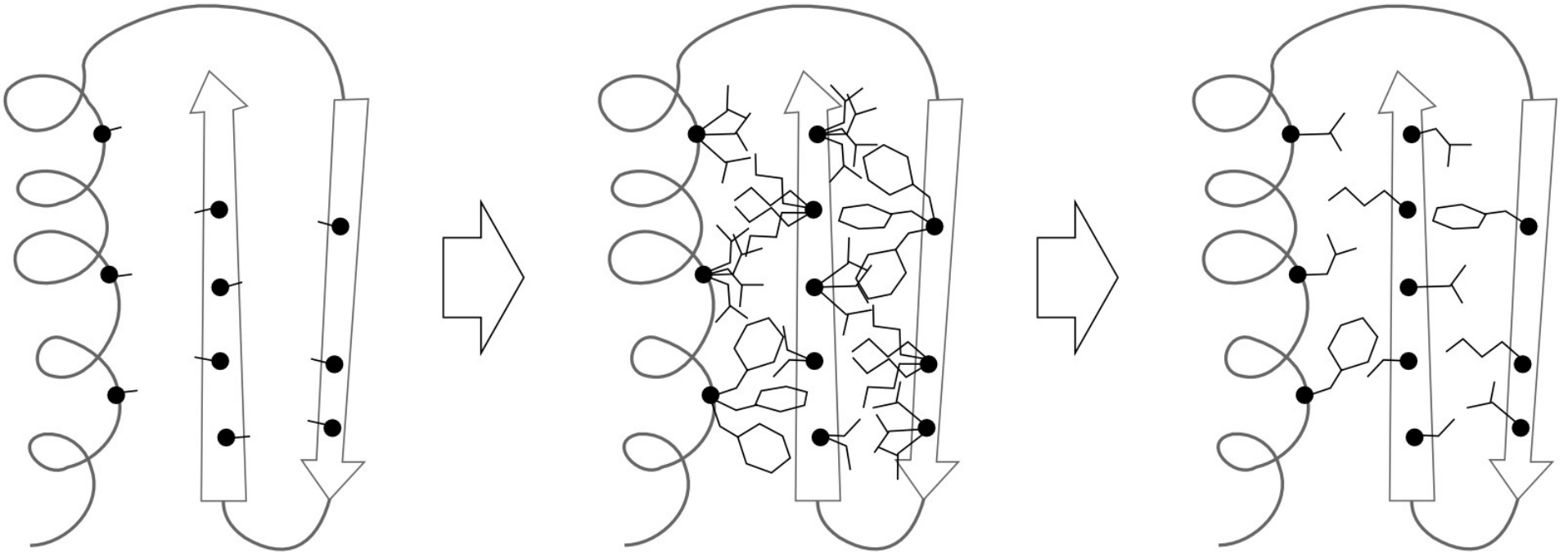

6
χ1 ~ 60°
9
χ1 ~ 180°
2
7
5
χ1 ~ -60°
4
8
1
3
10


-60
χ_2 180
60
60
180
-60
χ_1
6
5
7
9
2
11
10
3
8
1
4

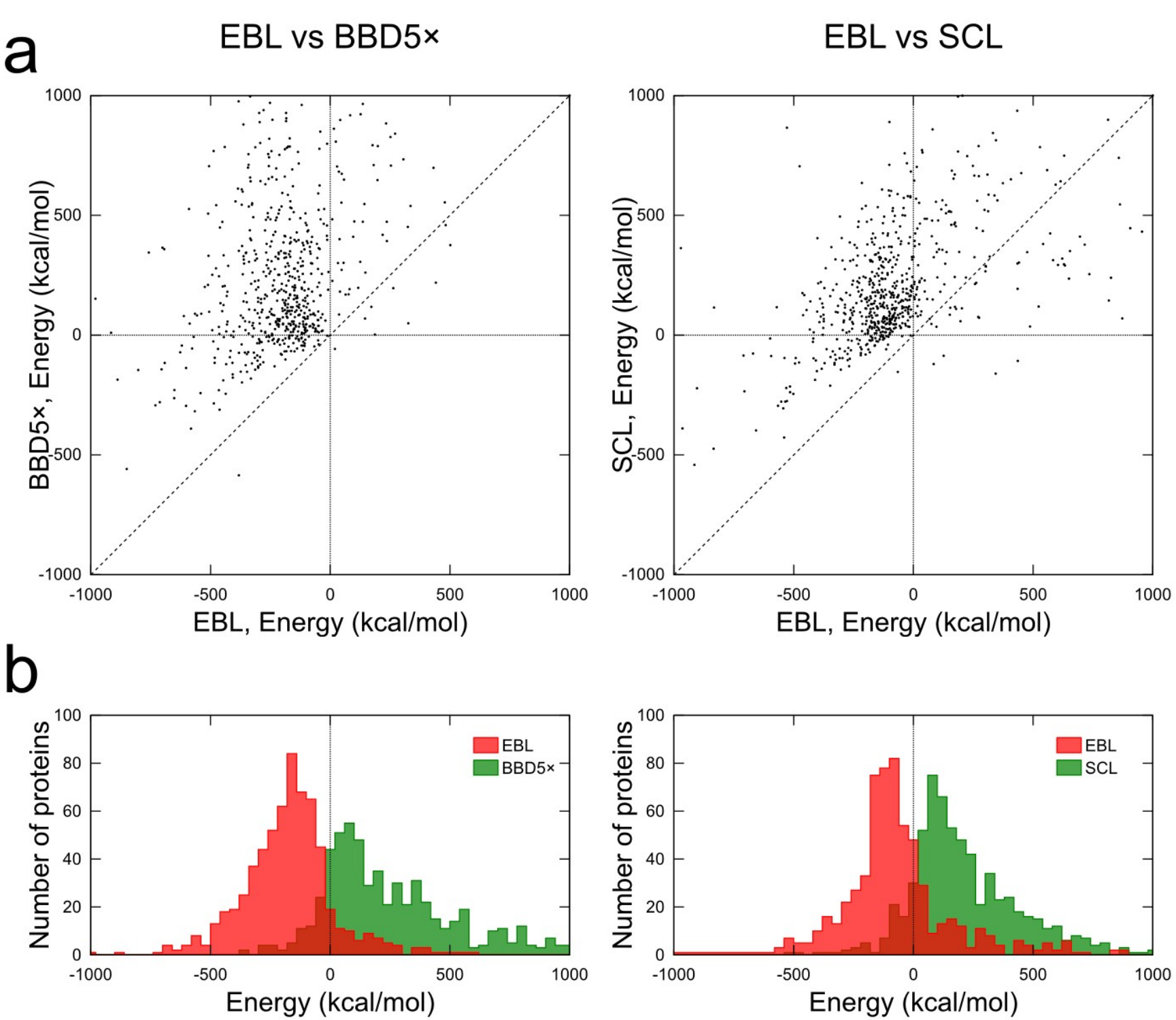

χ1 ~ 60°
χ1 ~ 180°
χ1 ~ -60°
6
11
9
2
7
5
4
8
1
3
10


-60
χ_2 180
60
60
180
-60
χ_1
6
5
7
9
2
11
10
3
8
1
4
12


6
11
χ1 ~ 60°
9
χ1 ~ 180°
2
7
12
5
4
8
1
3
10
χ1 ~ -60°



721 complete protein repacks.
The lower the energy, the better.


a
EBL vs BBD5×
EBL vs SCL
BBD5×, Energy (kcal/mol)
SCL, Energy (kcal/mol)
EBL, Energy (kcal/mol)
EBL, Energy (kcal/mol)
1000
500
0
-500
-1000
b
Number of proteins
EBL
BBD5×
SCL
Energy (kcal/mol)
Energy (kcal/mol)
100
80
60
40
20
0

Flexible High-Perfomance Conformer Library

- Conformers chosen using the same criterion as the optimization algorithm - Energy
- The new library is a sorted list of conformations
- Unprecedented flexibility – the first 'n' conformers is probably the best set of 'n' conformers

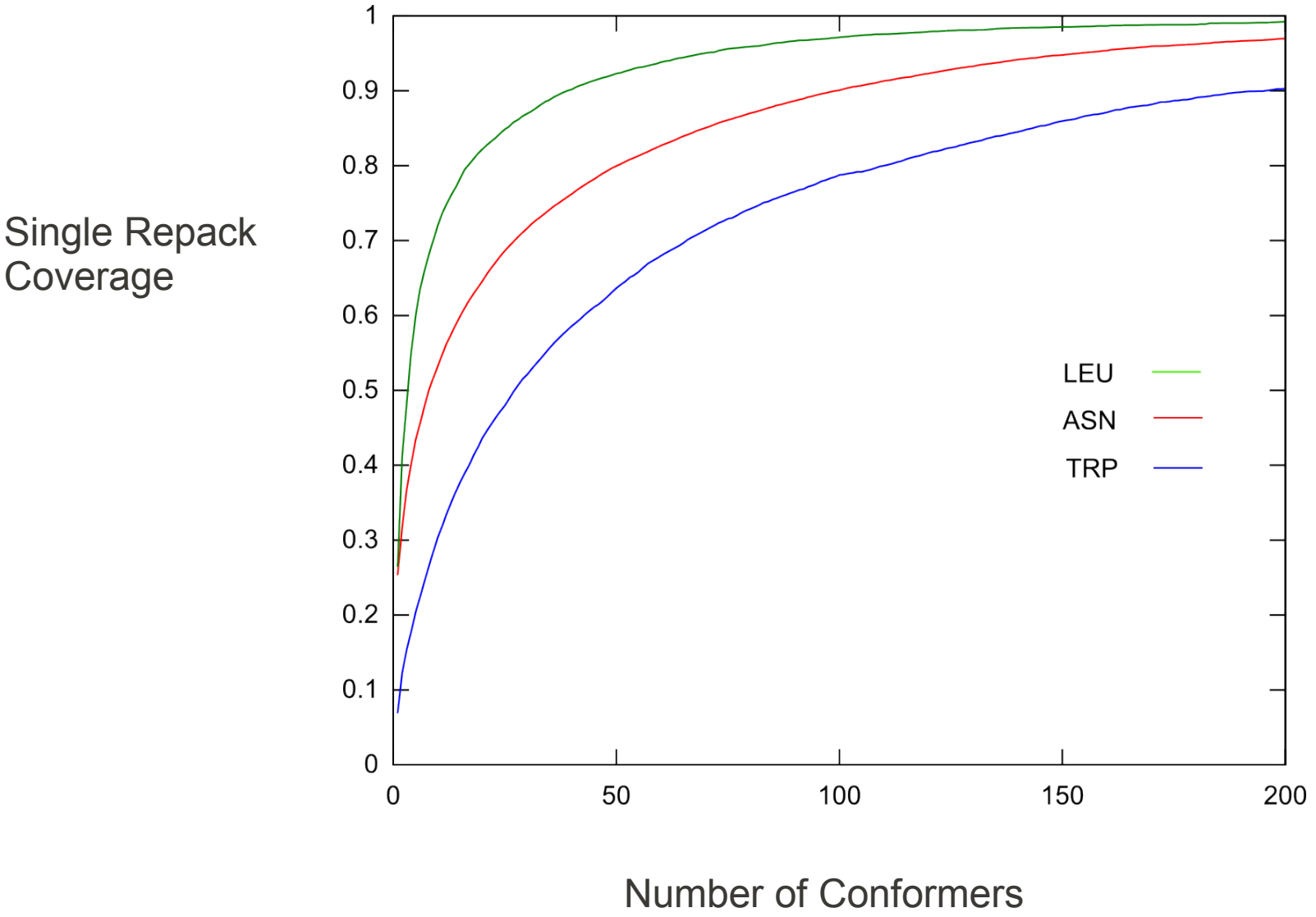
of the side chain conformation library

side chain

Optimization before ~~the~~ optimization exceeds optimistic expectation

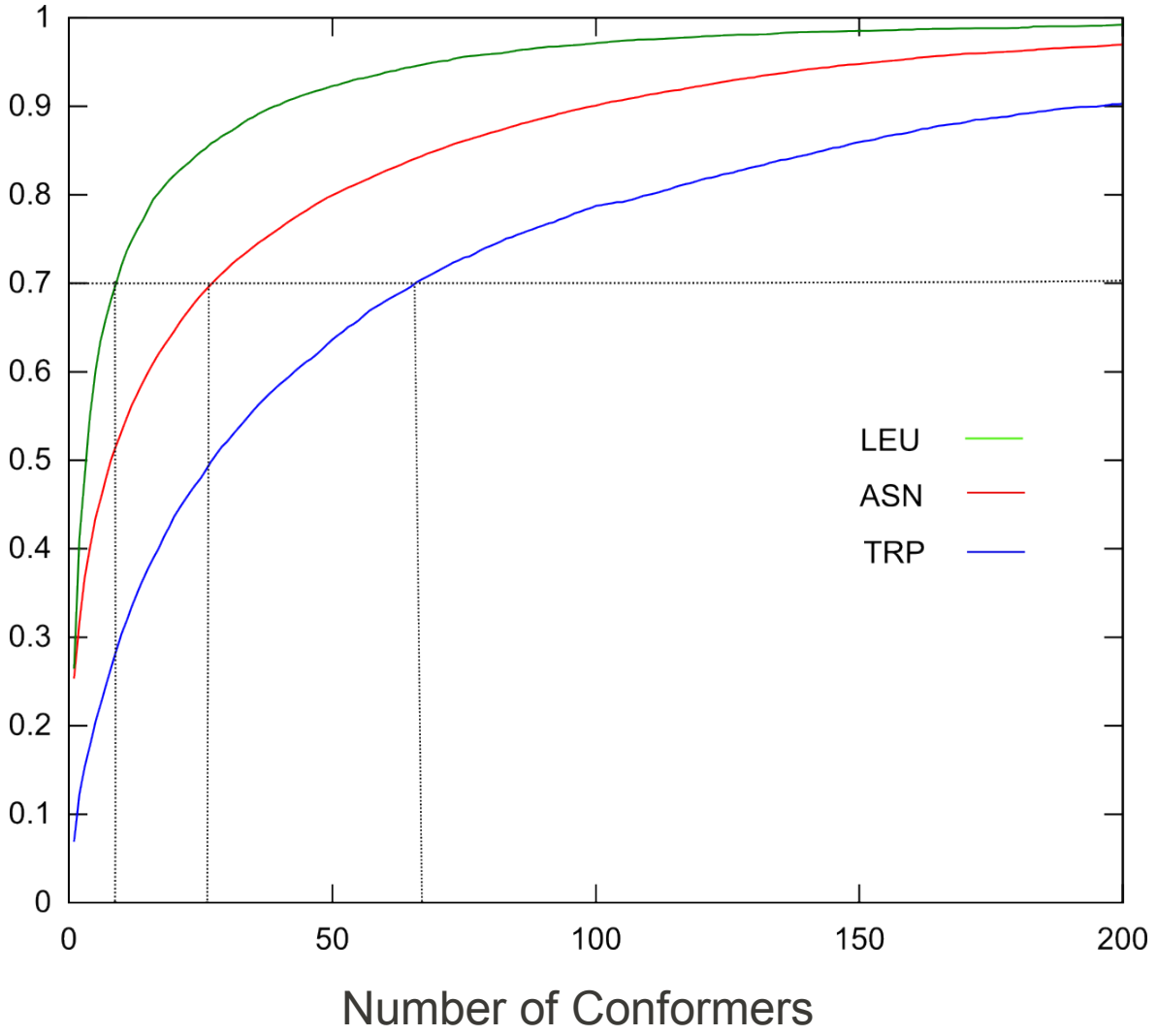
Alessandro Senes
IPiB Retreat 2011

NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE



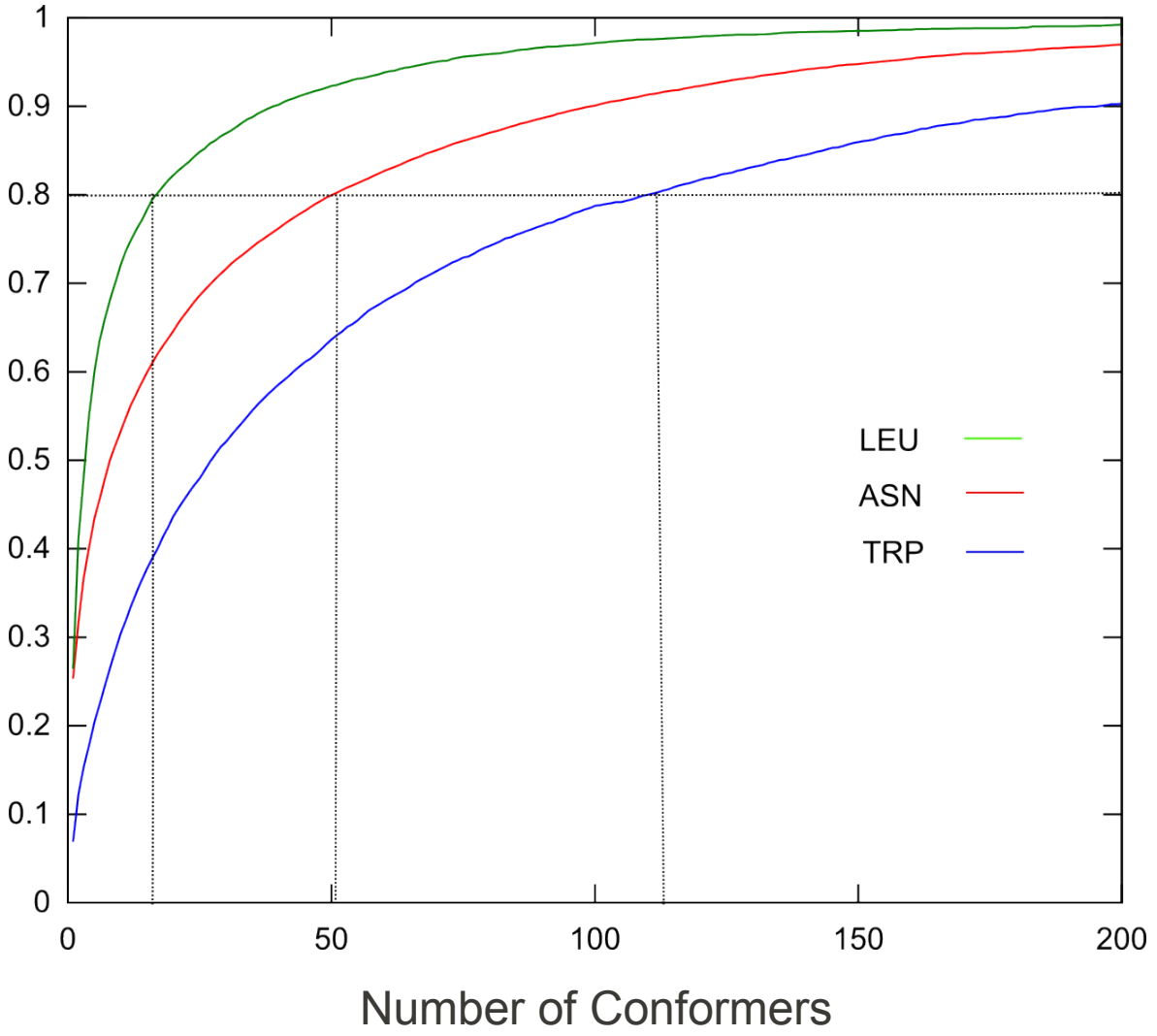
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage



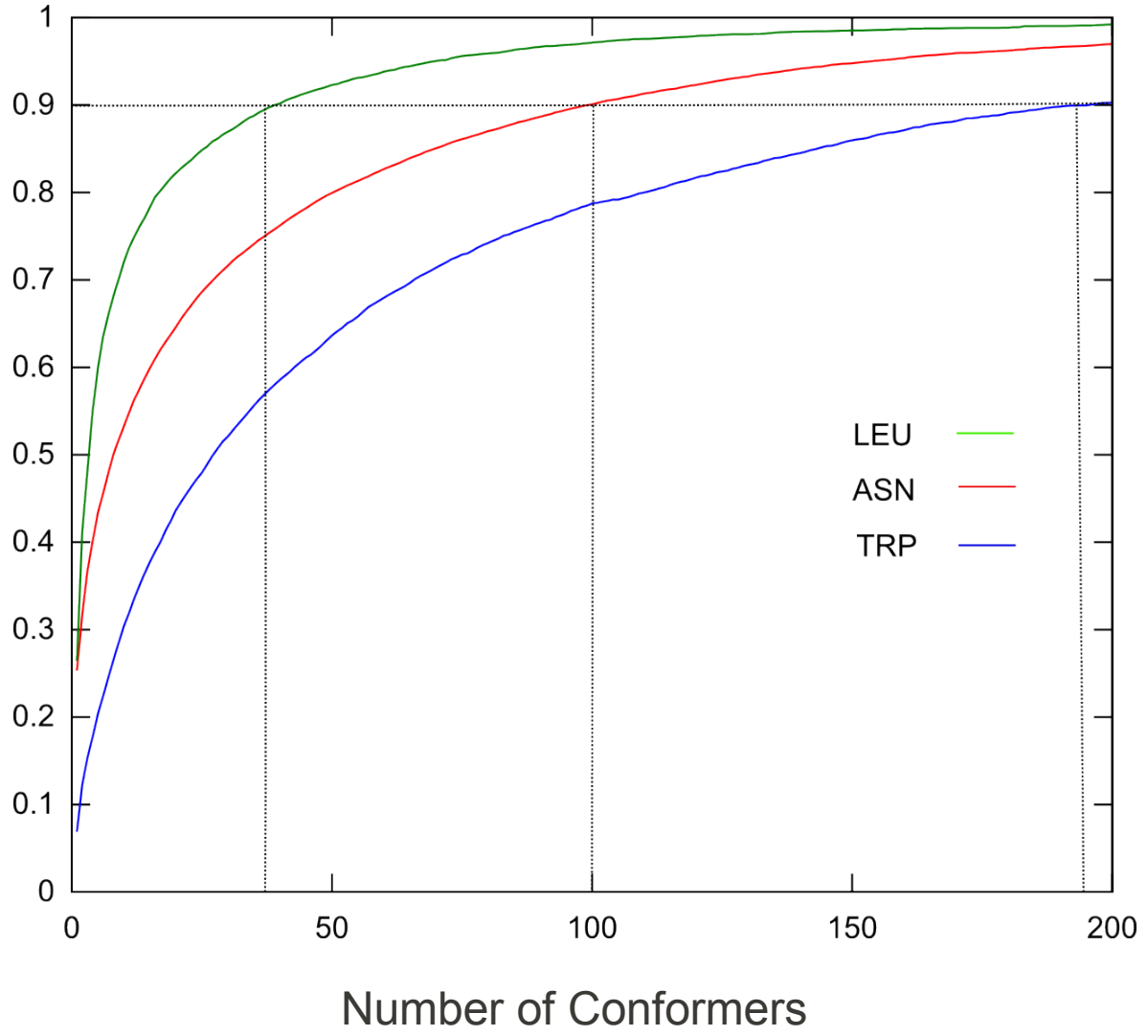
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage

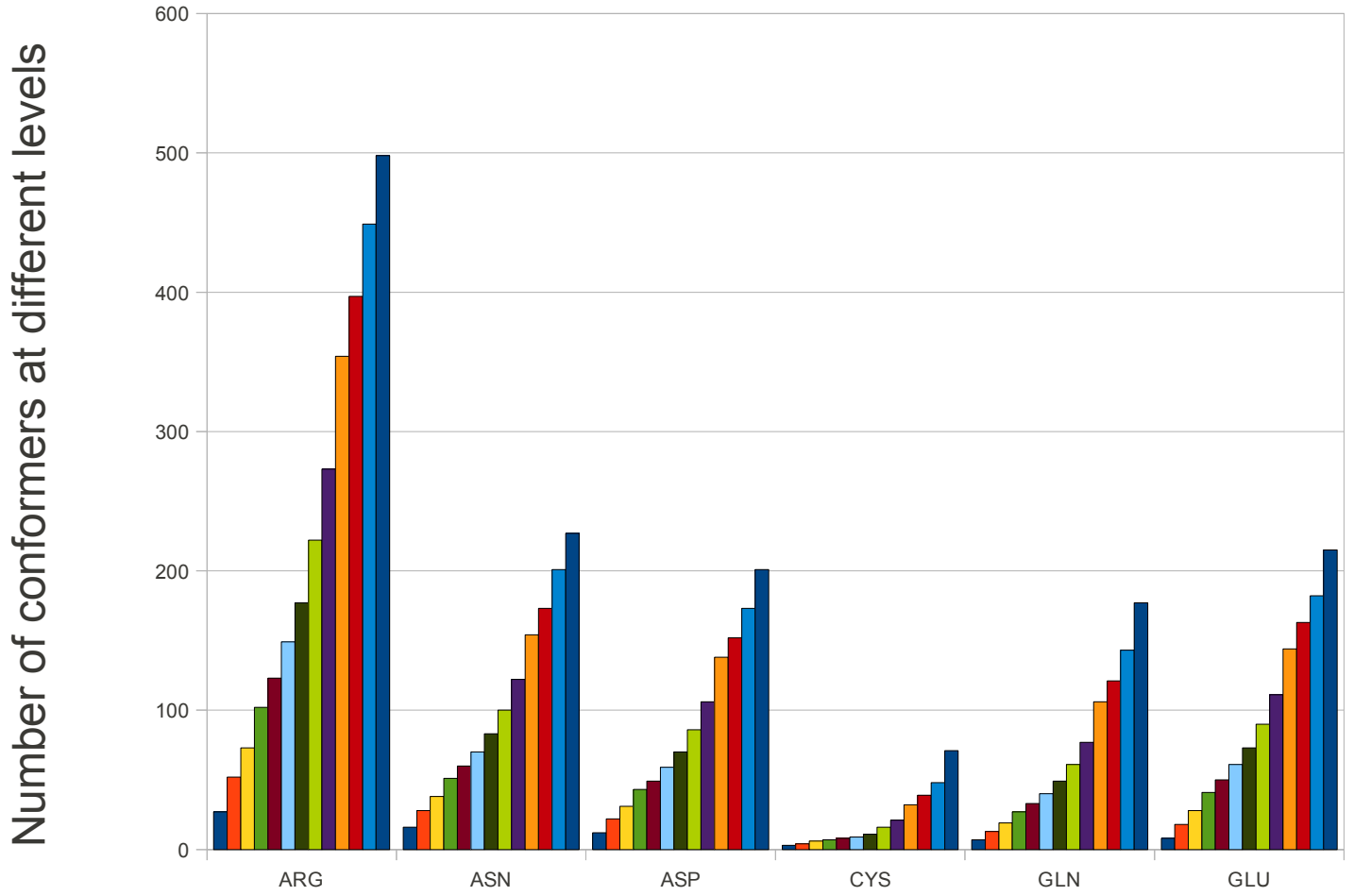


NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

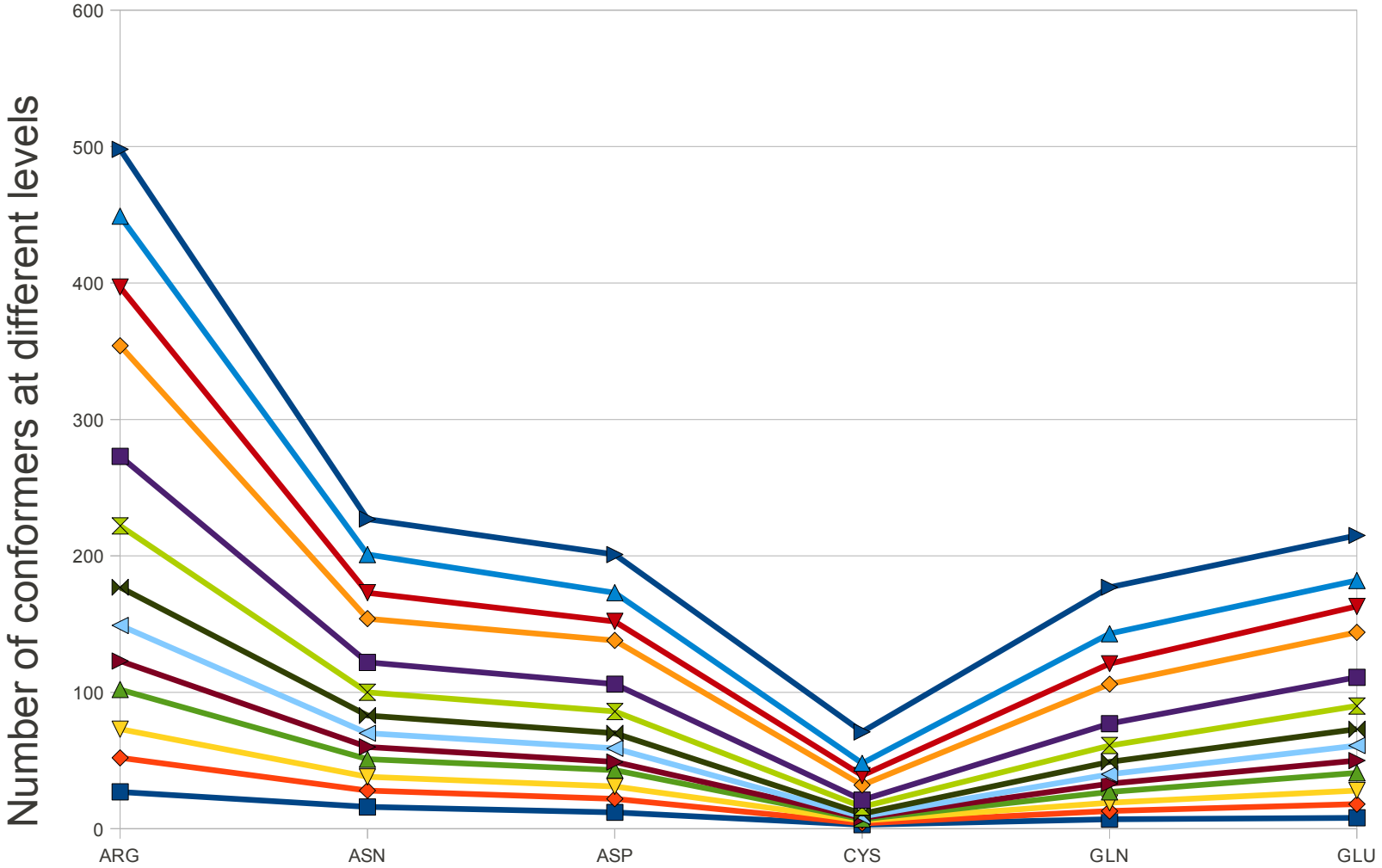
Single Repack Coverage



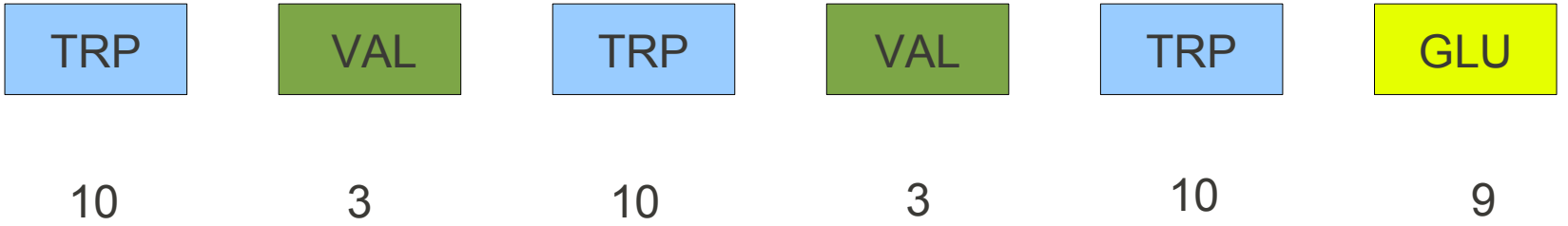
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE



NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE



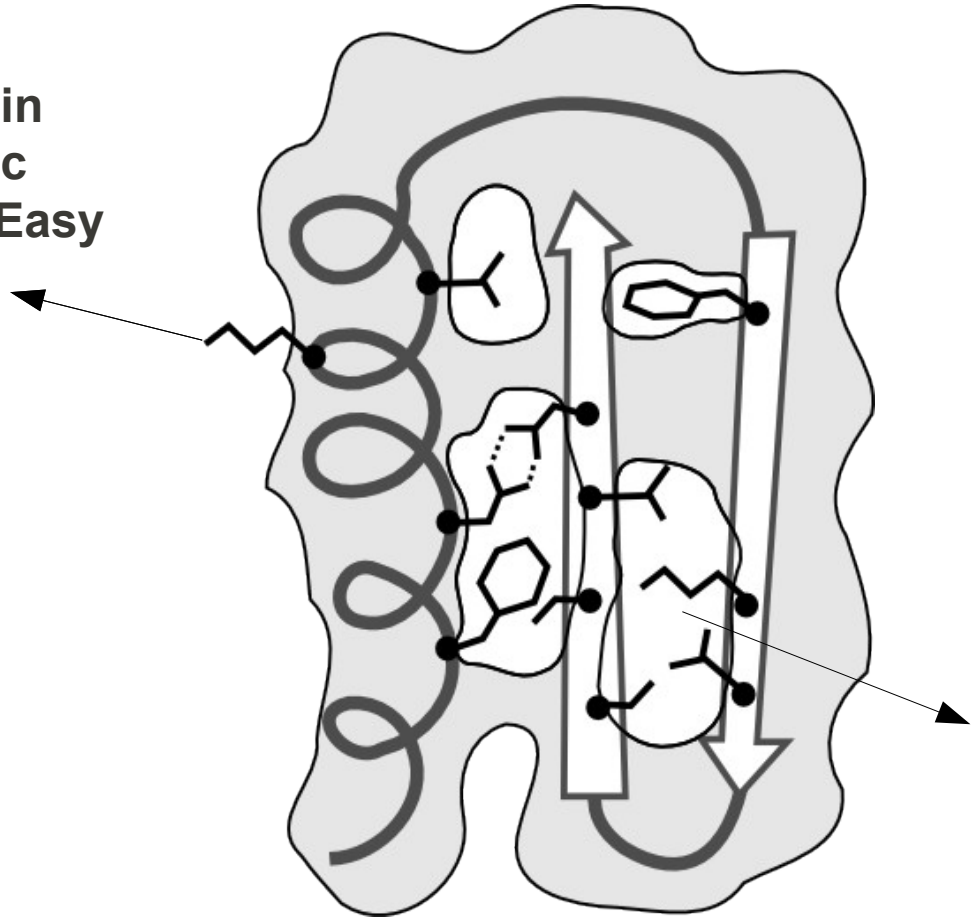
Combinatorial search space



possible conformations = 10 * 3 * 10 * 3 * 10 * 9 = 81000

Do all positions have the same sampling requirements?

Exposed sidechain
Many isoenergetic
Conformations - Easy



Buried sidechain
Fewer isoenergetic
Conformations - Hard

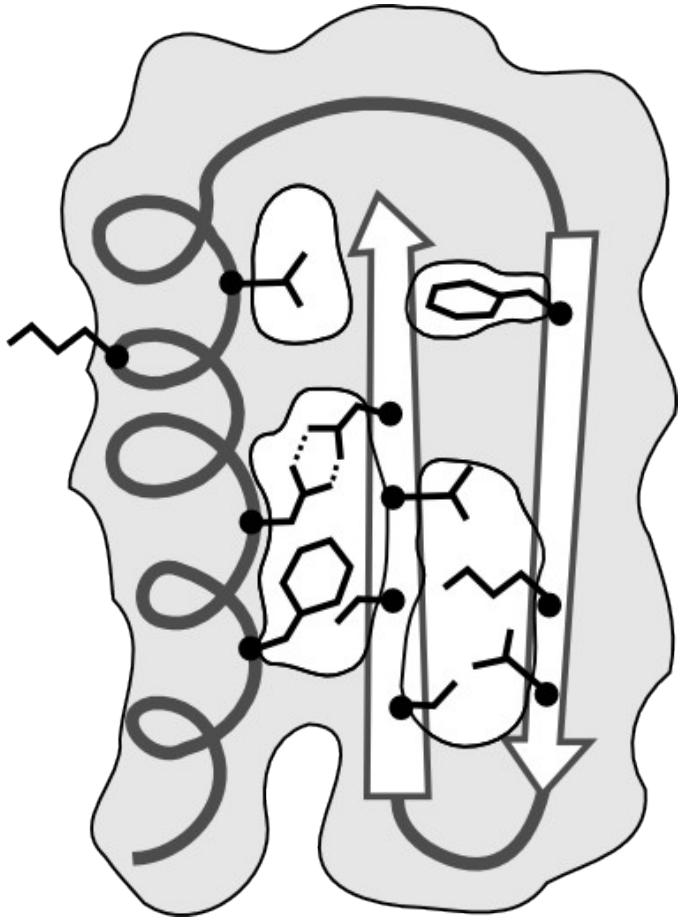
Smaller/Better search space with distributed sampling

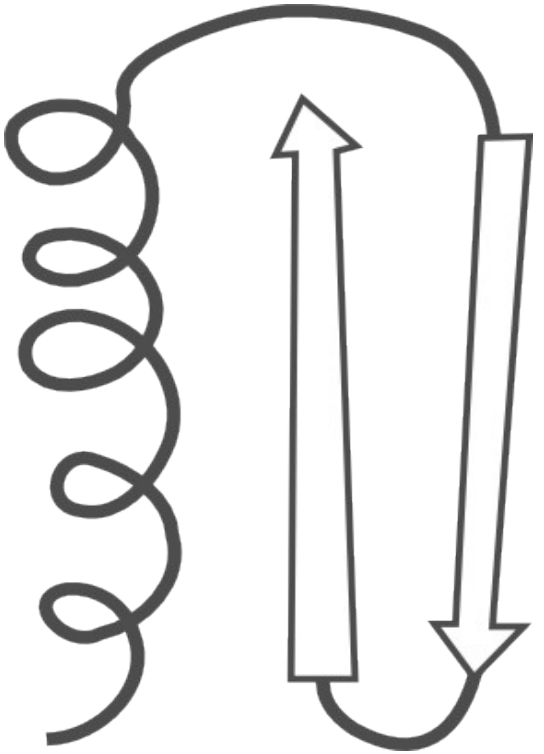
TRP	VAL	TRP	VAL	TRP	GLU	
10	3	10	3	10	9	81000
5	2	15	5	2	9	13500 (Faster)
7	3	17	7	3	11	82467 (Better)

By moving sampling from the easy positions to the hard ones, we could be more efficient (fast) and/or achieve better energies

But, can we predict if a position is easy or hard?

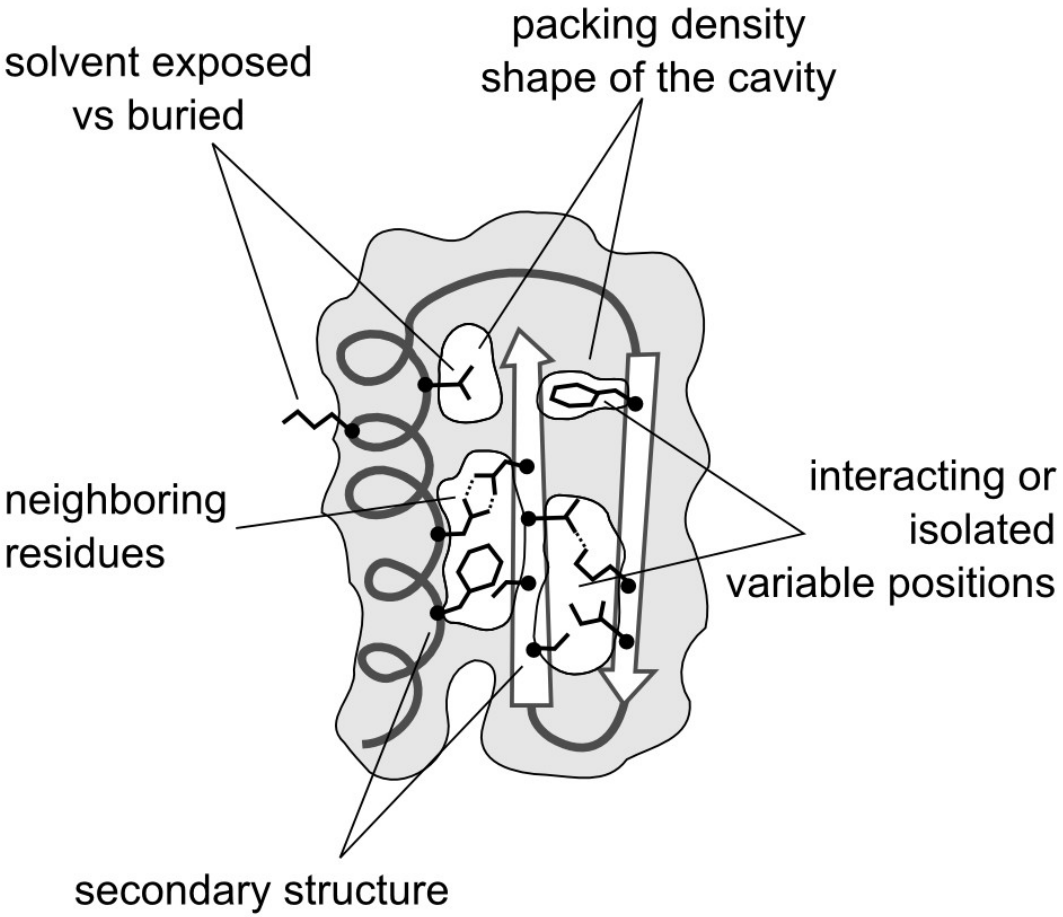
SIDECHAIN OPTIMIZATION





Use Machine Learning

- Information in the backbone - pattern recognition problem?
- Use machine learning to predict requirements based on features of the backbone


solvent exposed
vs buried
packing density
shape of the cavity
neighboring
residues
interacting or
isolated
variable positions
secondary structure

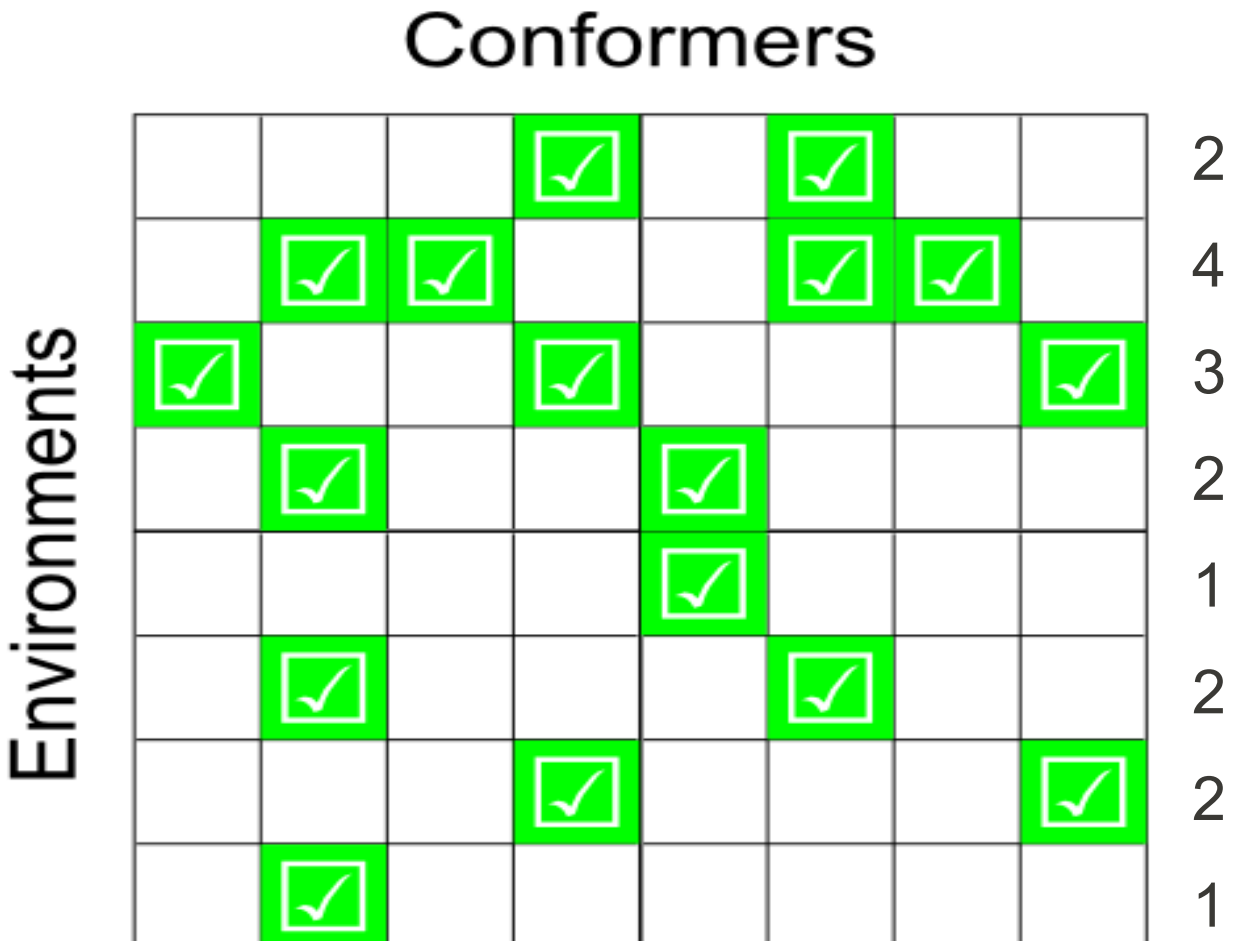
Goals

- Identify sampling requirements for each position on backbone
- Reduce run time, find better conformations (lower energy), or both

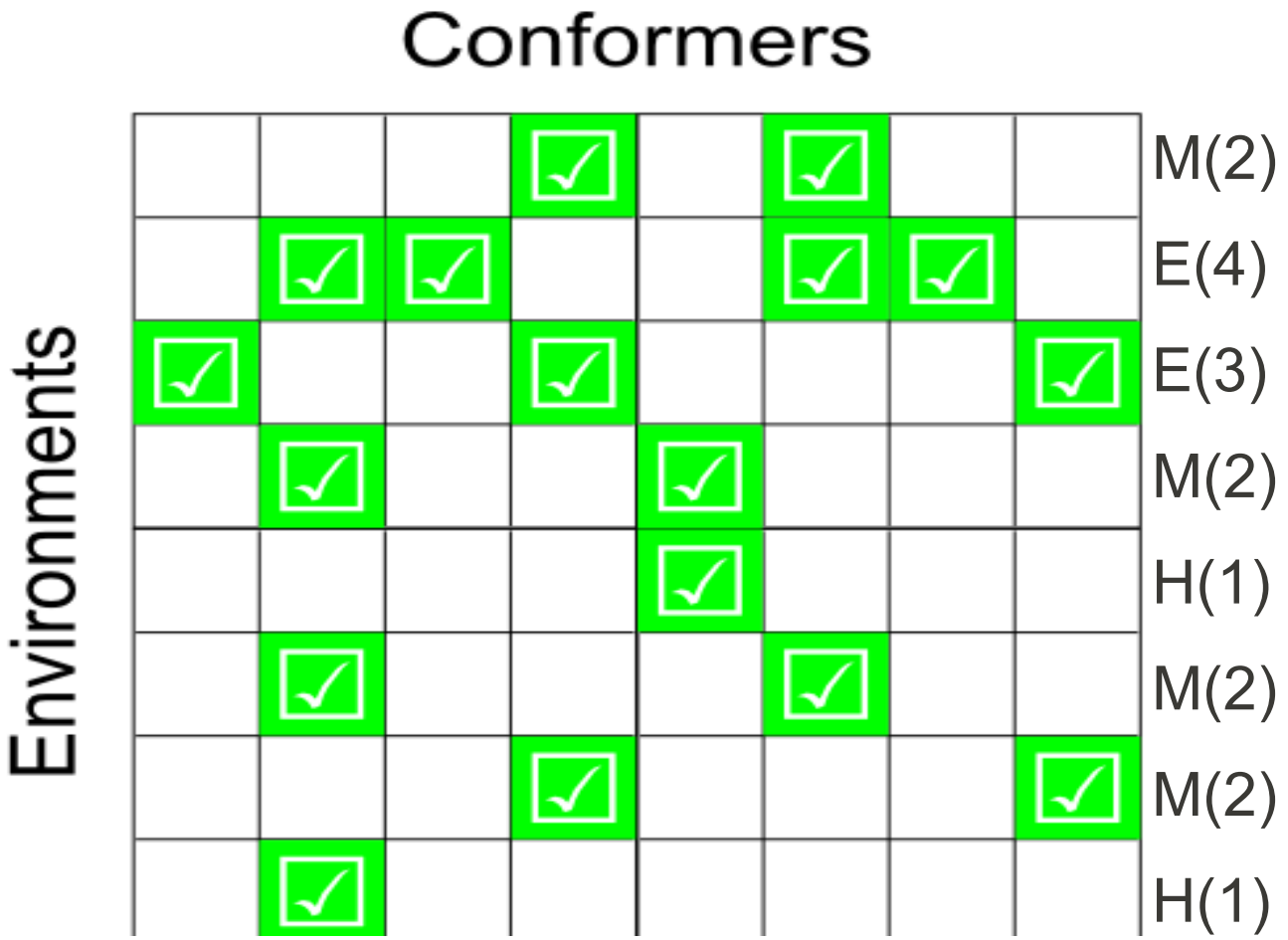
Issues

- Identify useful features from the backbone structure
- Identify most meaningful labeling strategy to label the dataset
- Devise the best machine learning strategy to predict the label

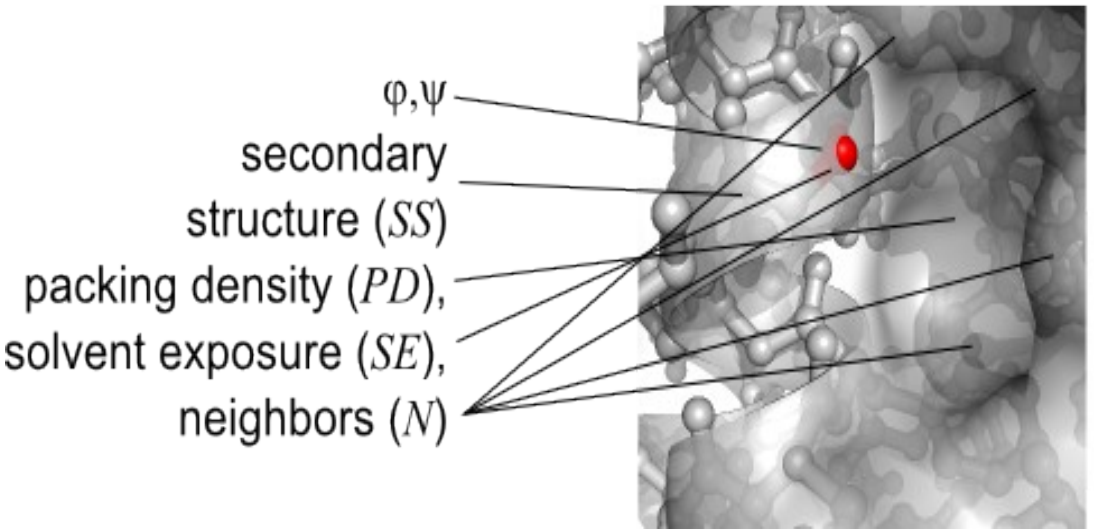
Label the dataset using the EnergyTable



Label the dataset using the EnergyTable

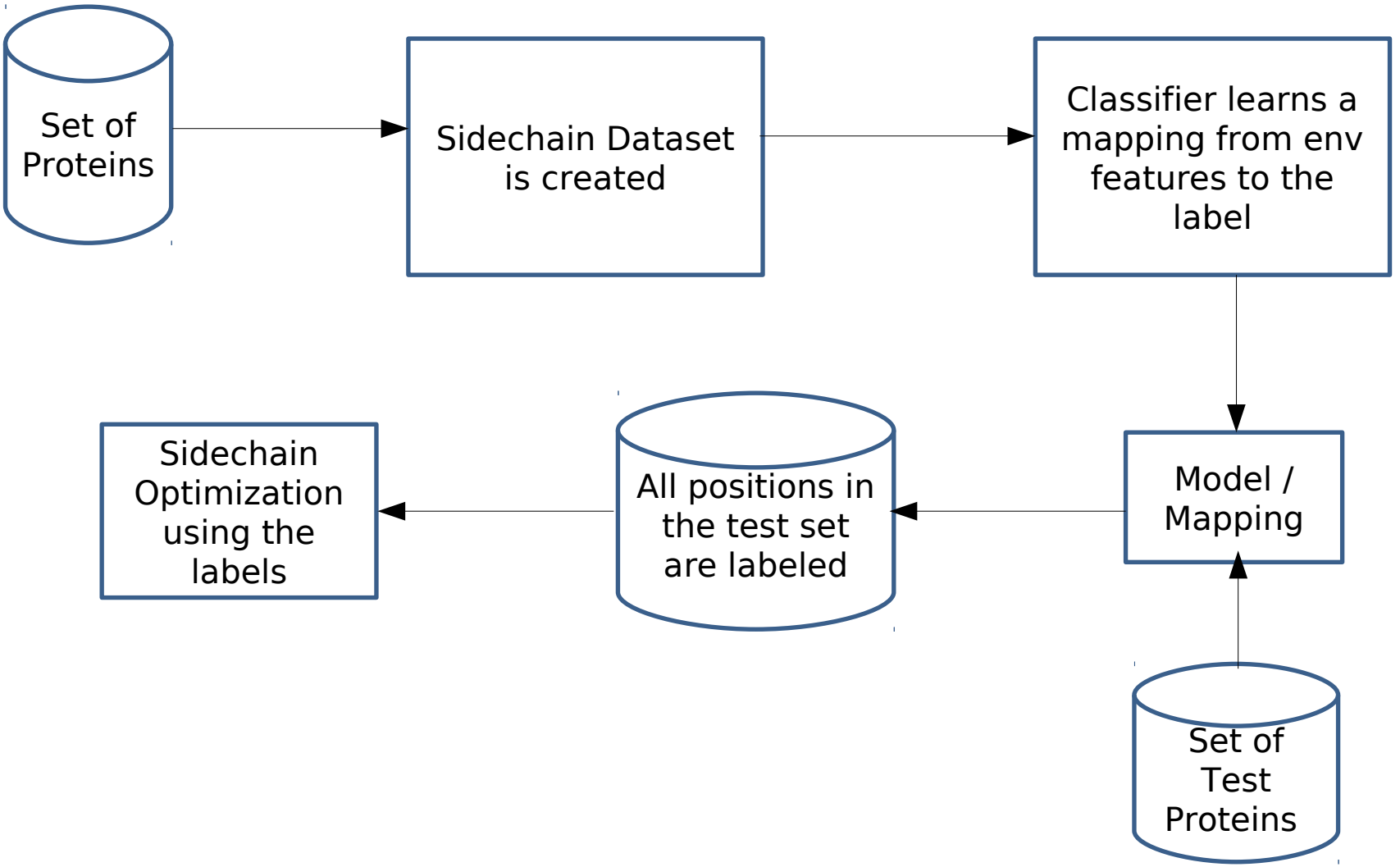


Associate each sidechain in the database
with a feature vector $\boldsymbol{X}$ and a label $\boldsymbol{Y}$

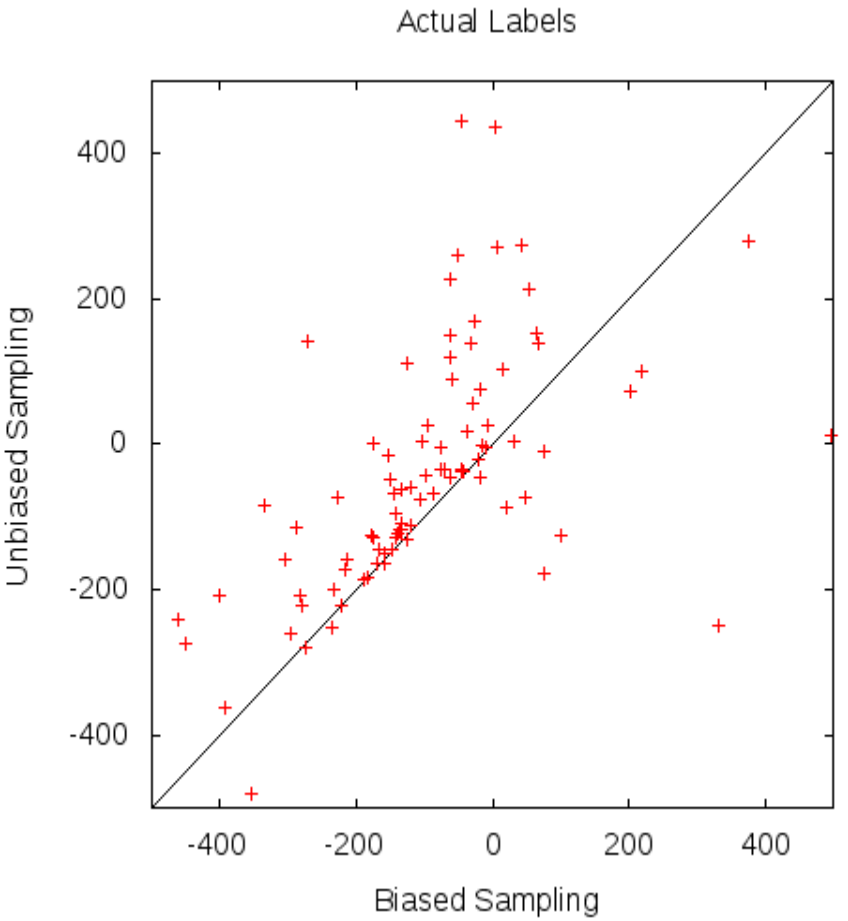


$\boldsymbol{X}$ = {φ, ψ, *SS*, *SE*, *PD*, *N*, ...}
$\boldsymbol{Y}$ = {Hard, Medium, Easy}

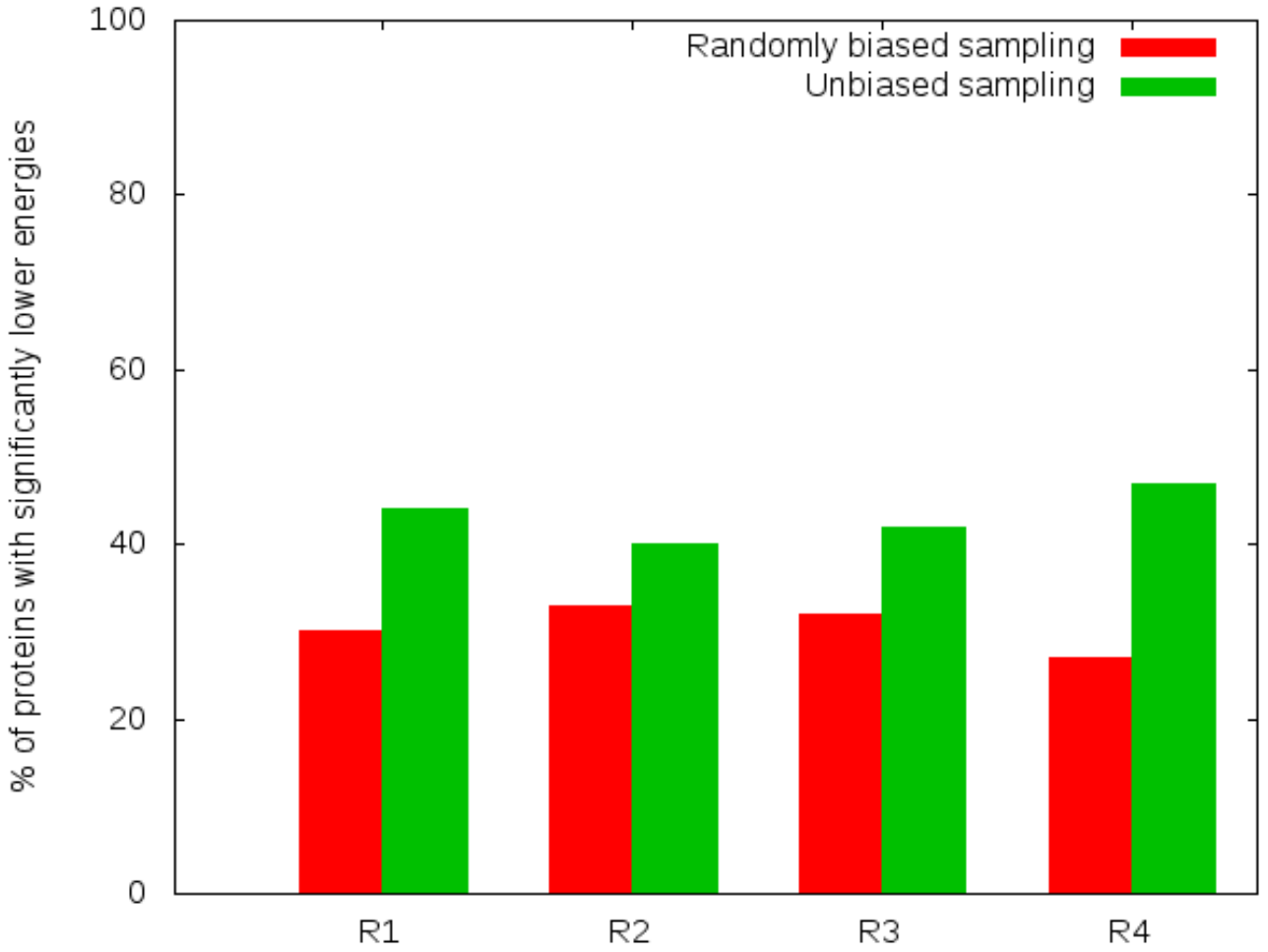
Overall Strategy



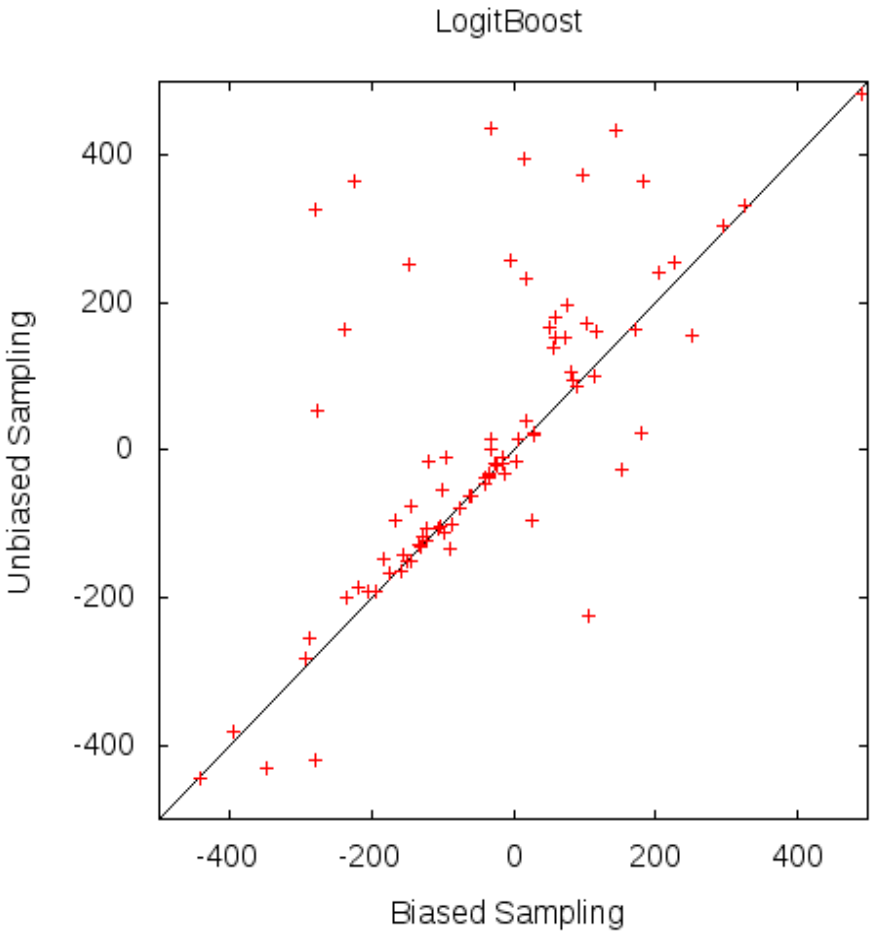
HOW EFFECTIVE ARE THE LABELS?



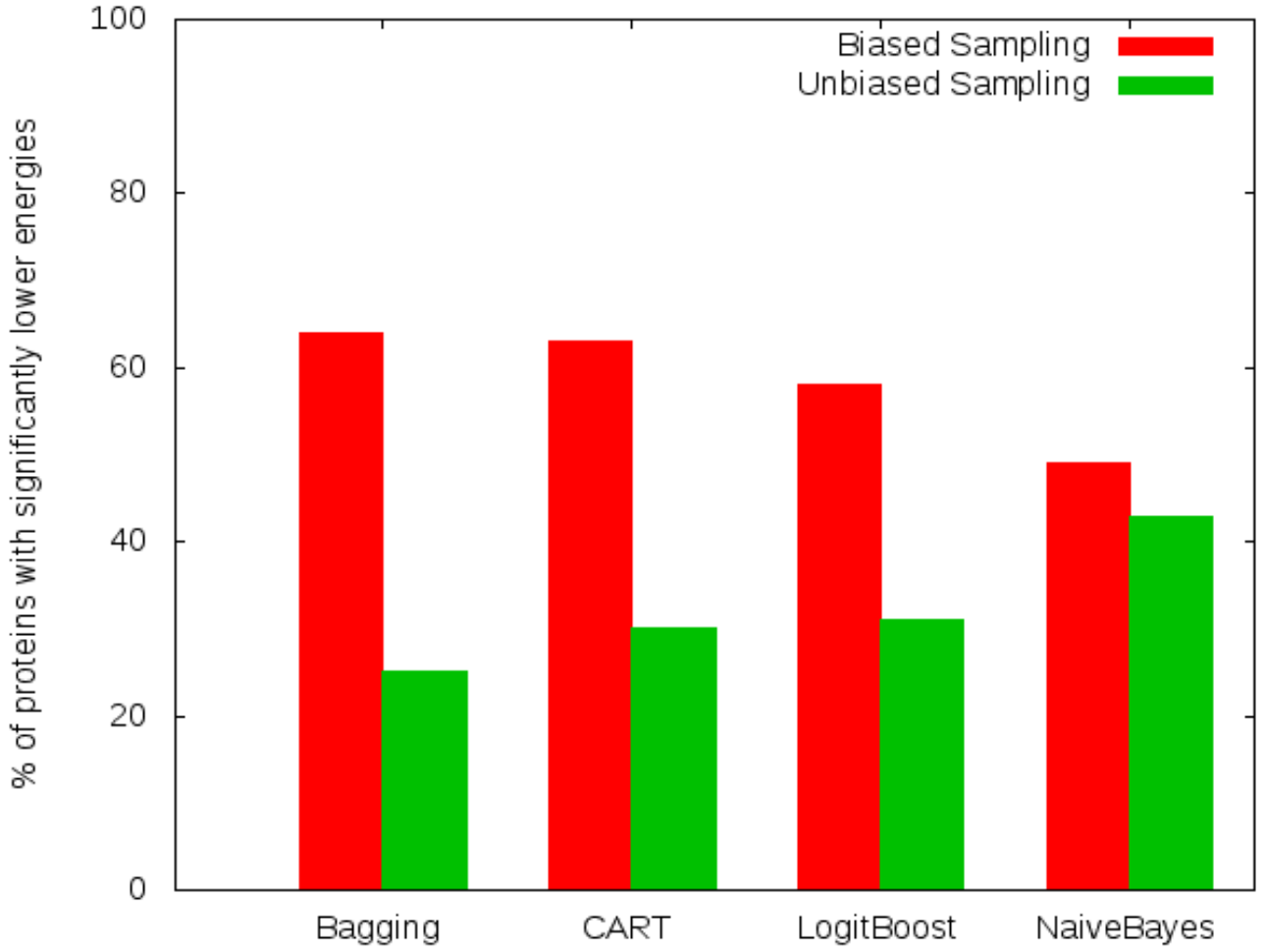
Results – Permuted labels



HOW WELL DOES MACHINE LEARNING WORK?



Results – All Classifiers



Future Research

- More features
- Alternative labeling strategies
- Optimize classifiers performance

Thank You

Alessandro Senes

Sriraam Natarajan
Ambalika Khadria
Loren LaPointe
Ben Mueller

Center of High Throughput Computing
University of Wisconsin-Madison