

Advanced machine learning for  
Innovative Drug Discovery (AIDD)

# 3D-Structure Refinement

## Simple Principles for Meaningful Structures

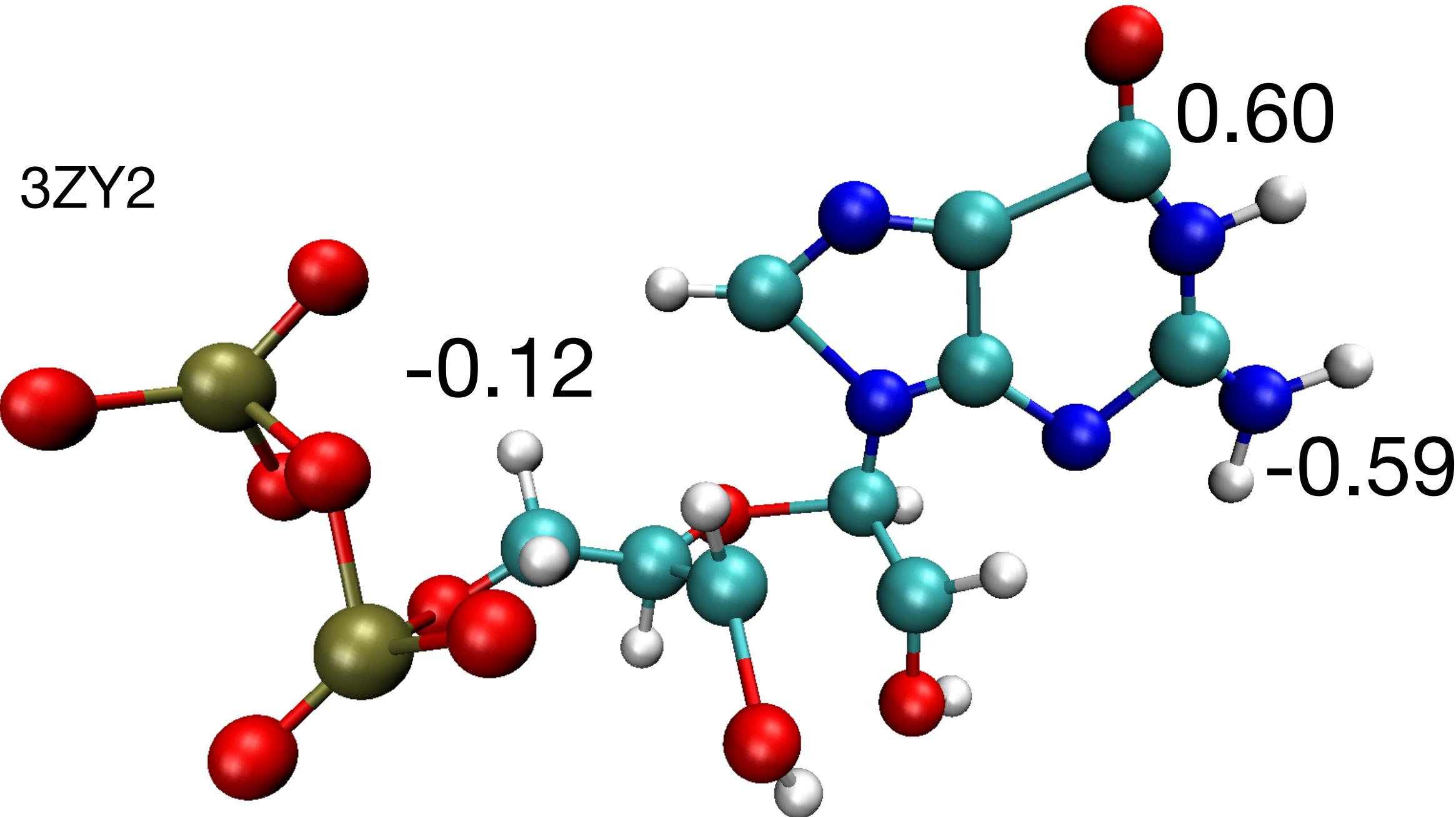
Dr. Filipe Menezes,  
Institute Structural Biology (STB), Popowicz group  
Helmholtz-Zentrum München

HELMHOLTZ  
MUNICH The logo for the Technical University of Munich (TUM) consists of the letters "TUM" in a bold, blue, sans-serif font. A thick blue horizontal bar is positioned behind the letters, and a blue curved arrow graphic is located to the right of the text.

Institute of  
Structural Biology

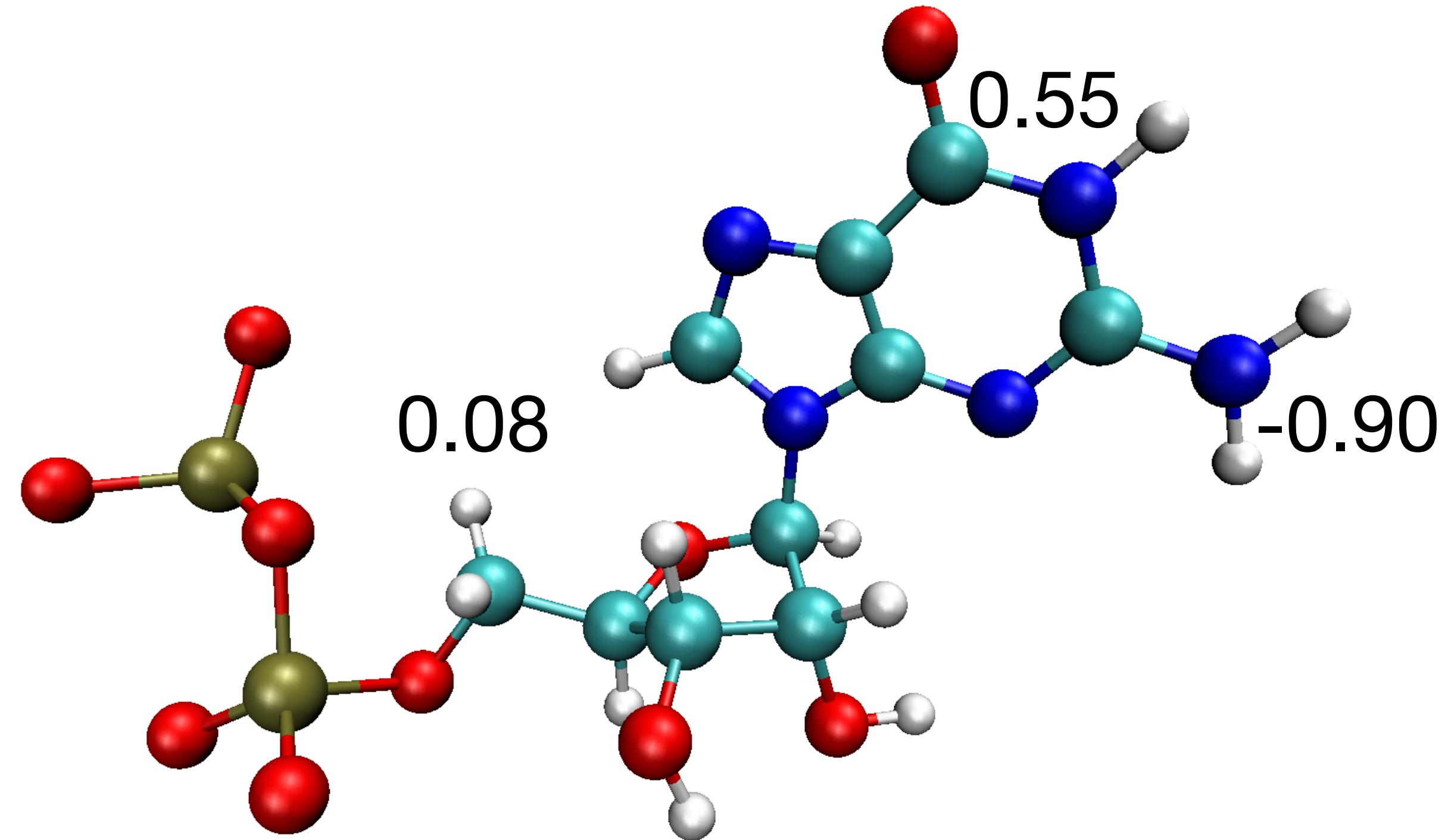
Bayerisches NMR ZentrumThe logo for the Bavarian NMR Center, featuring the text "Bayerisches NMR Zentrum" in a blue sans-serif font next to a blue circular arrow graphic.

# Why do we need good structures?



$$\Delta_{solv}G = - 566.32 \text{ kcal/mol}$$

$$\mu = 62.8 \text{ e\AA}$$

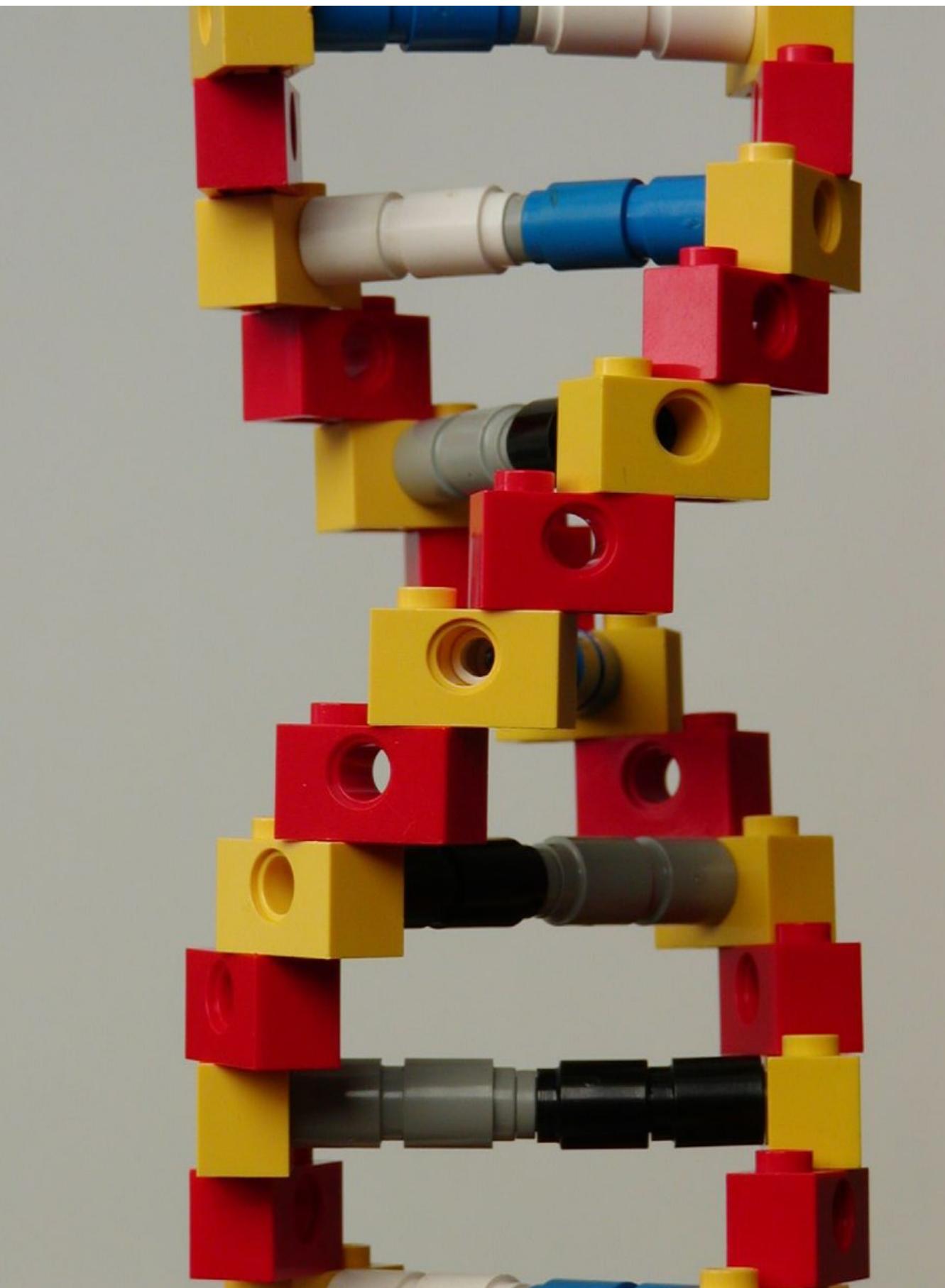


$$\Delta_{solv}G = - 513.39 \text{ kcal/mol}$$

$$\mu = 51.5 \text{ e\AA}$$

# Part 1

## Assembling Molecules



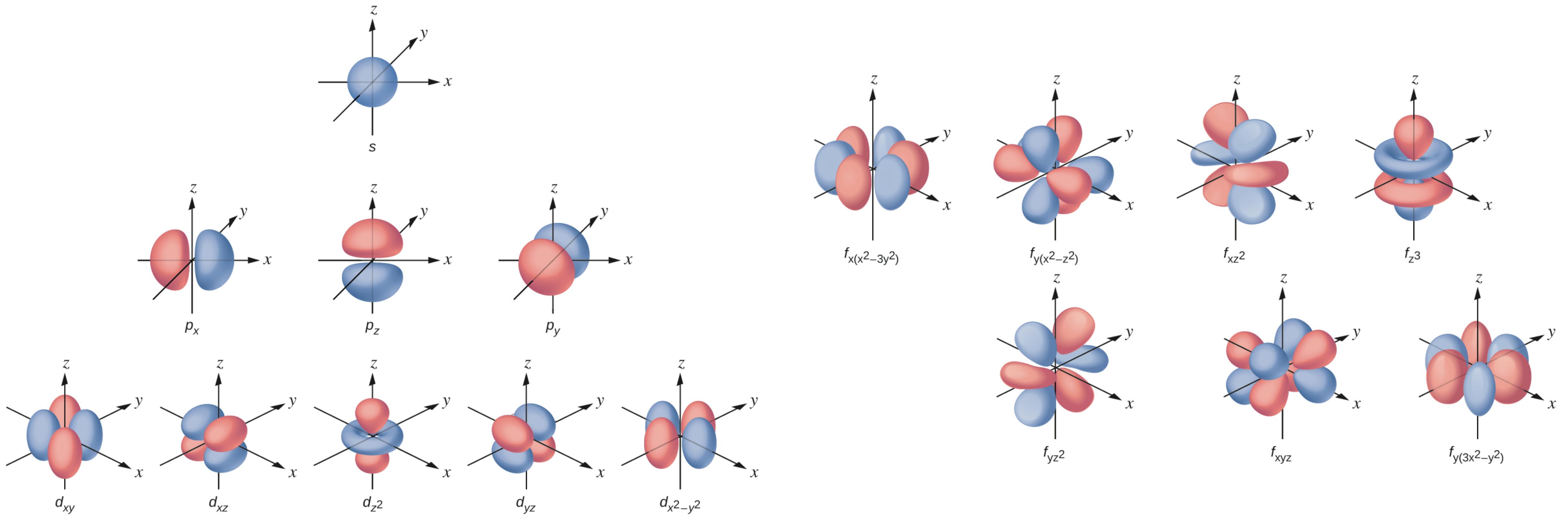
[Link](#)

# The Schrödinger Equation



$$(\hat{T} + \hat{V})\Psi = E\Psi$$

Erwin Schrödinger



# Base Principle

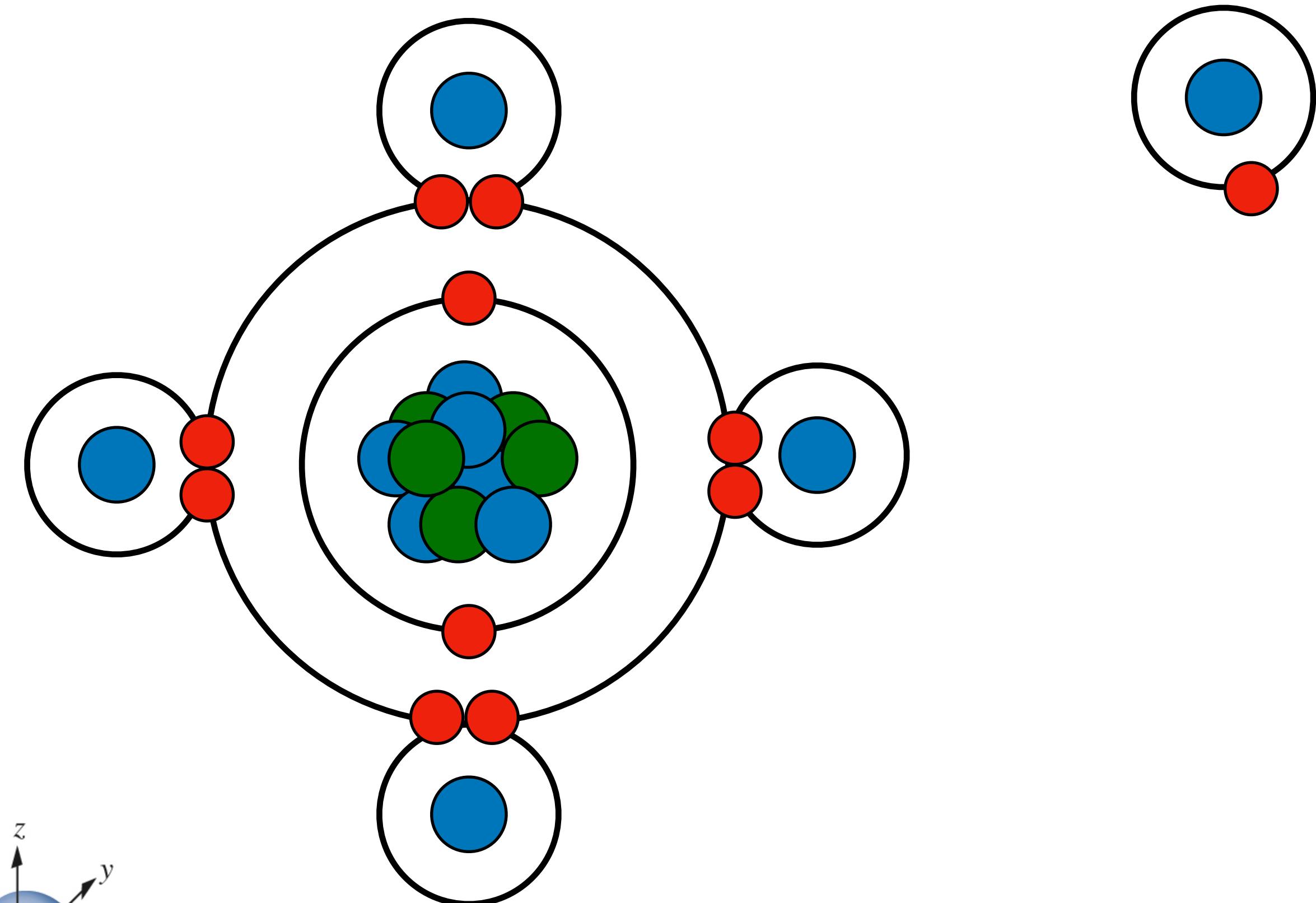
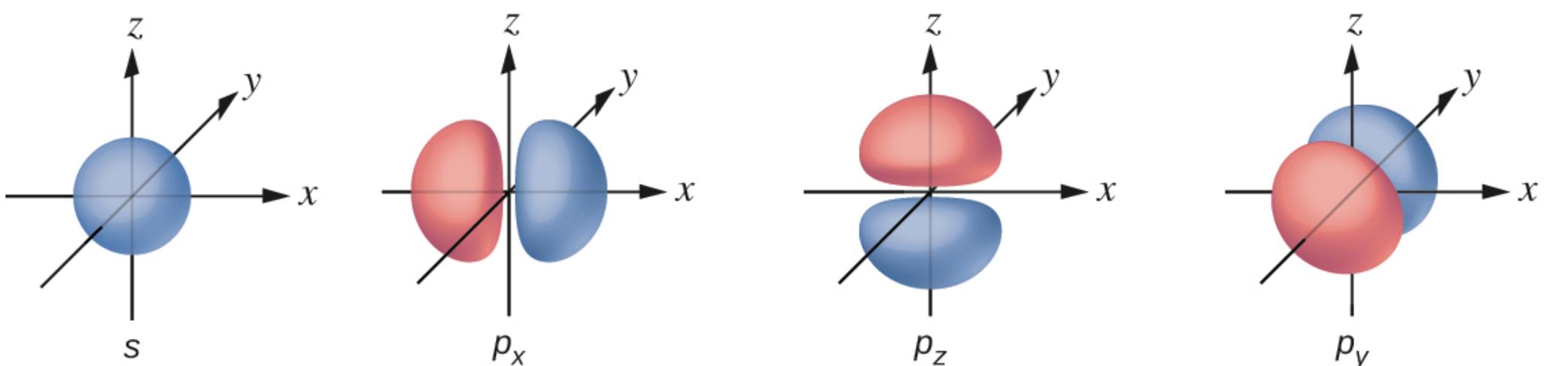
Atoms have OCD (obsessive compulsive disorder)



Brian Greene

## Sharing is Caring

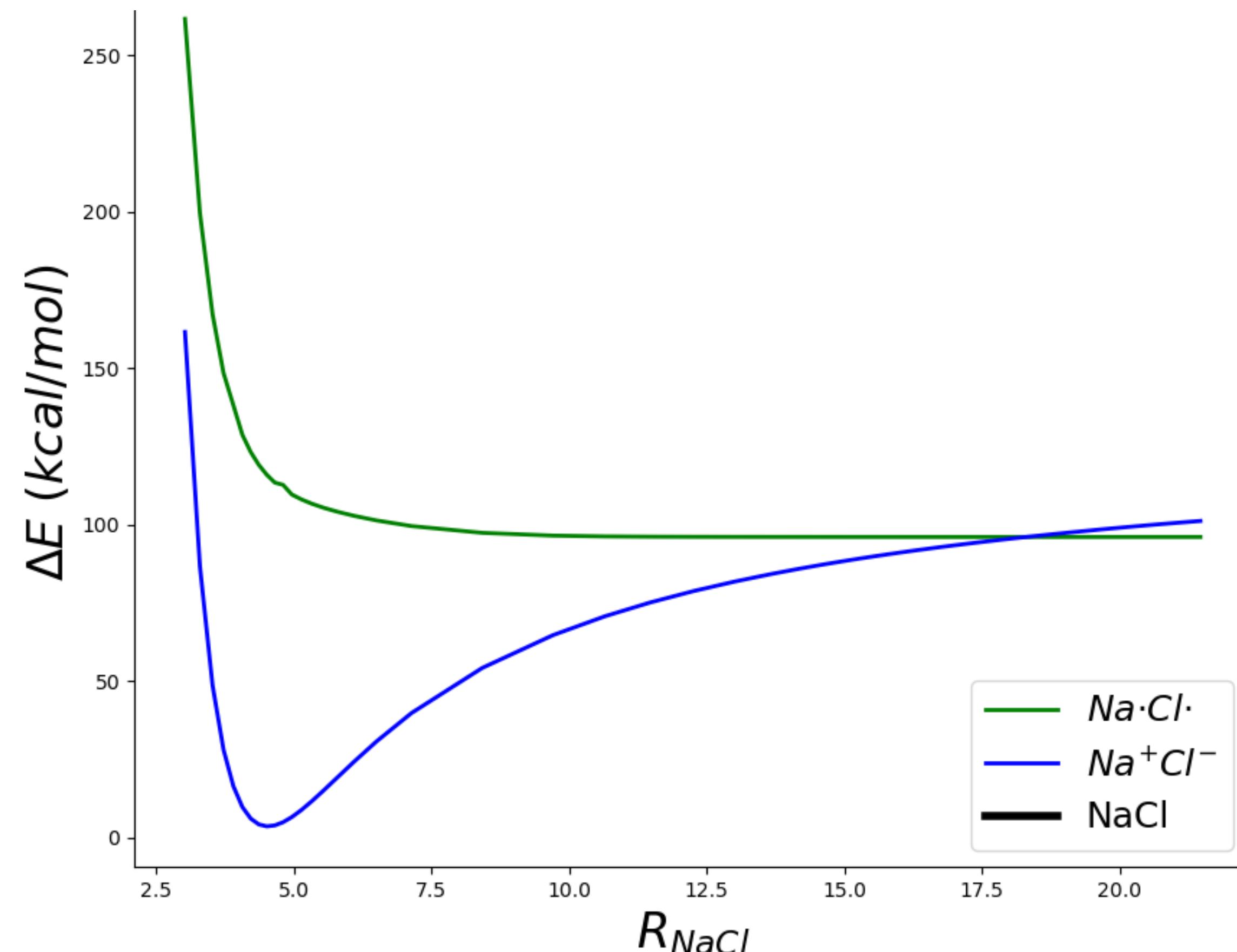
rather share electrons than having incomplete shells



- Proton
- Neutron
- Electron

# Base Principle

## Atoms have OCD (obsessive compulsive disorder)



They will do anything to avoid unpaired electrons

- “Philanthropist atoms”:

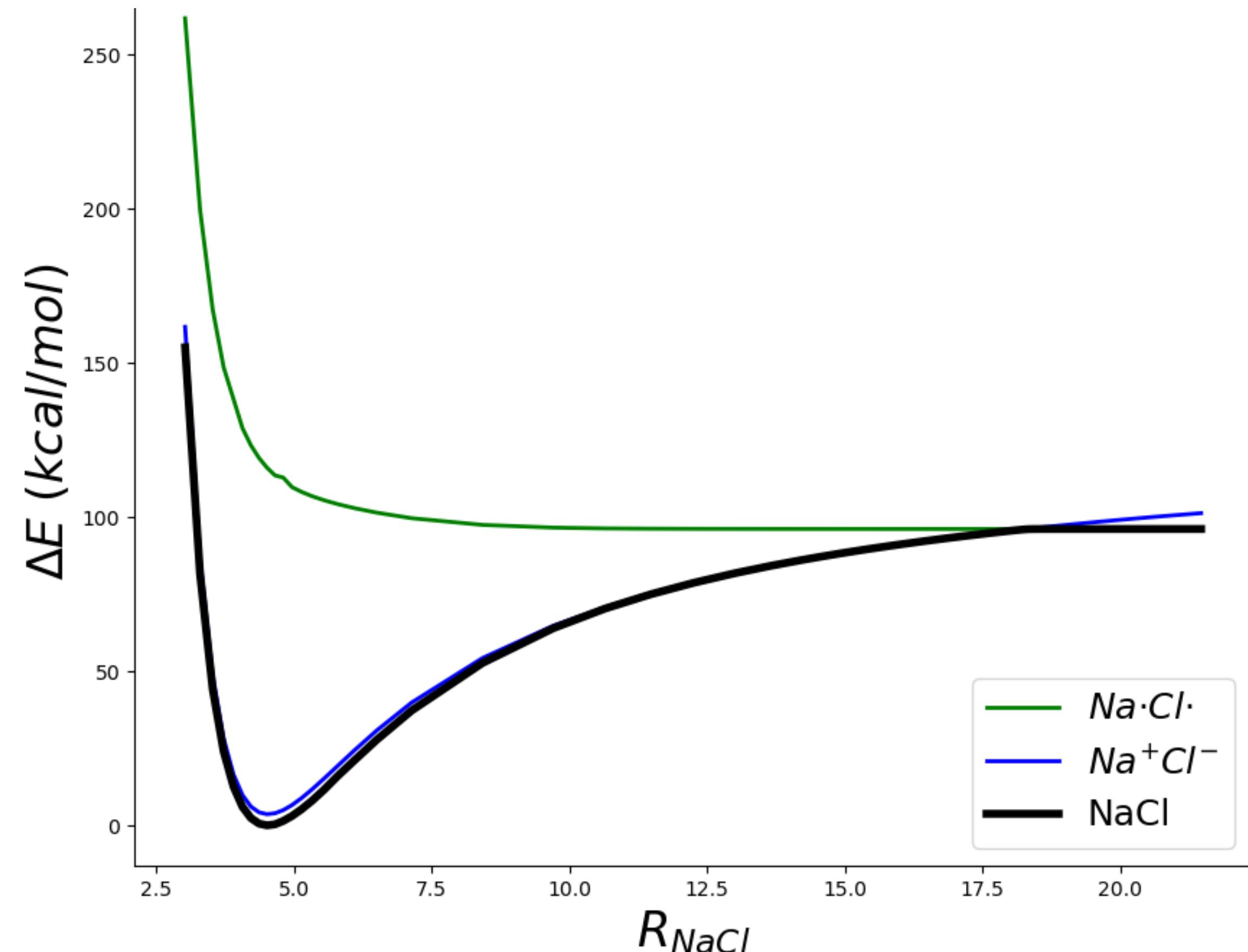


- “Looter atoms”:



# Base Principle

Atoms have OCD (obsessive compulsive disorder)

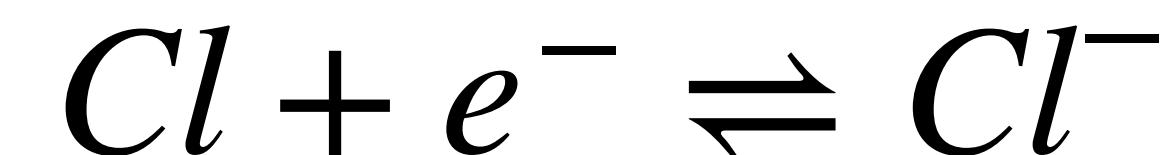


They will do anything to avoid unpaired electrons

- “Philanthropist atoms”:

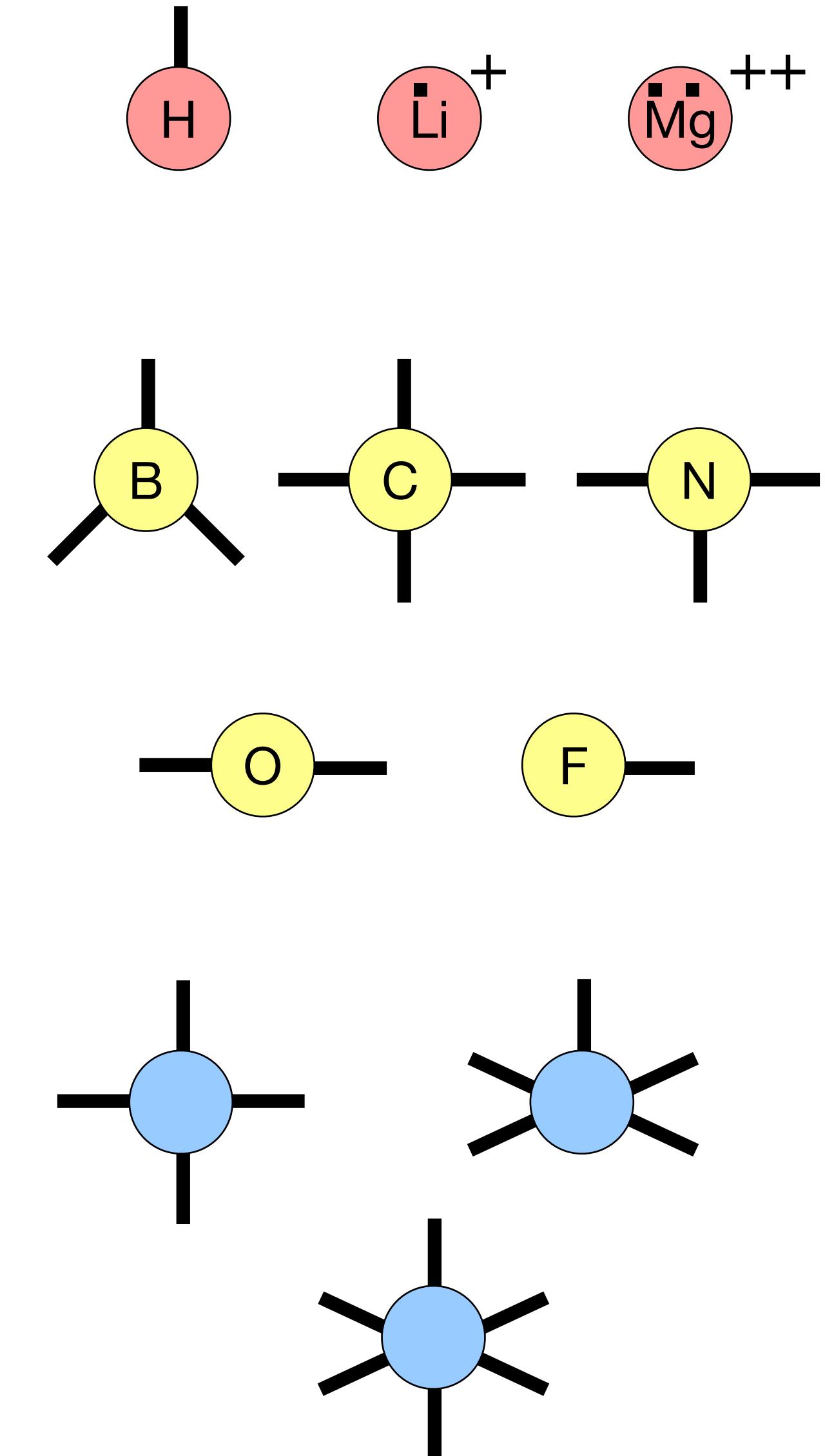


- “Looter atoms”:

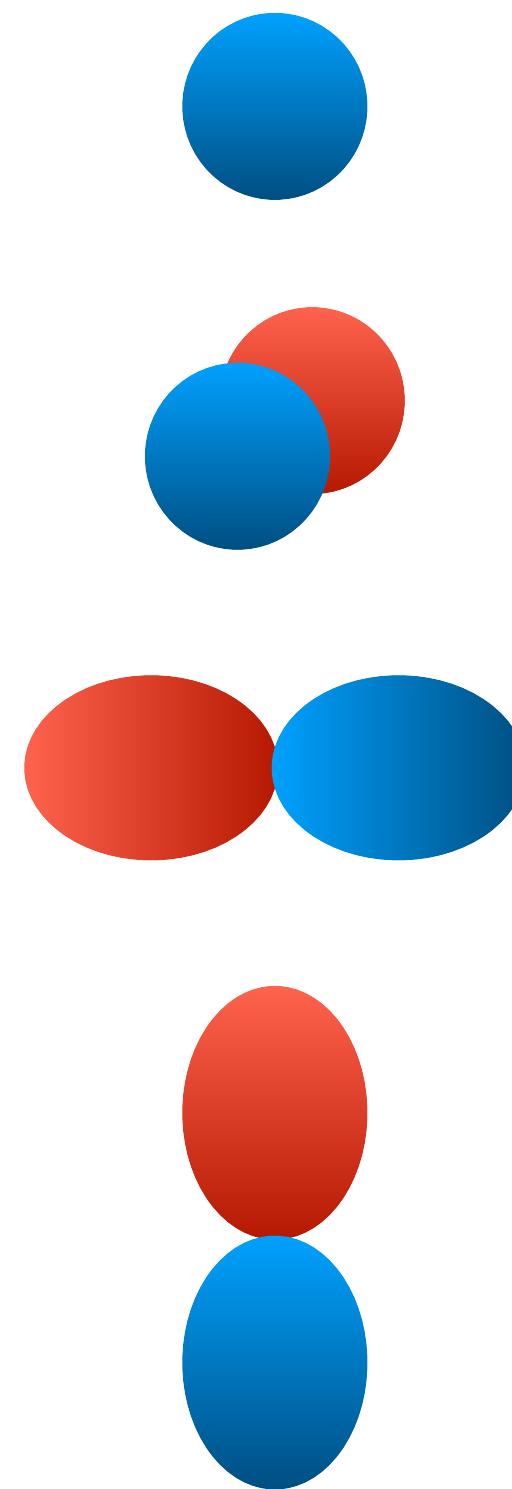
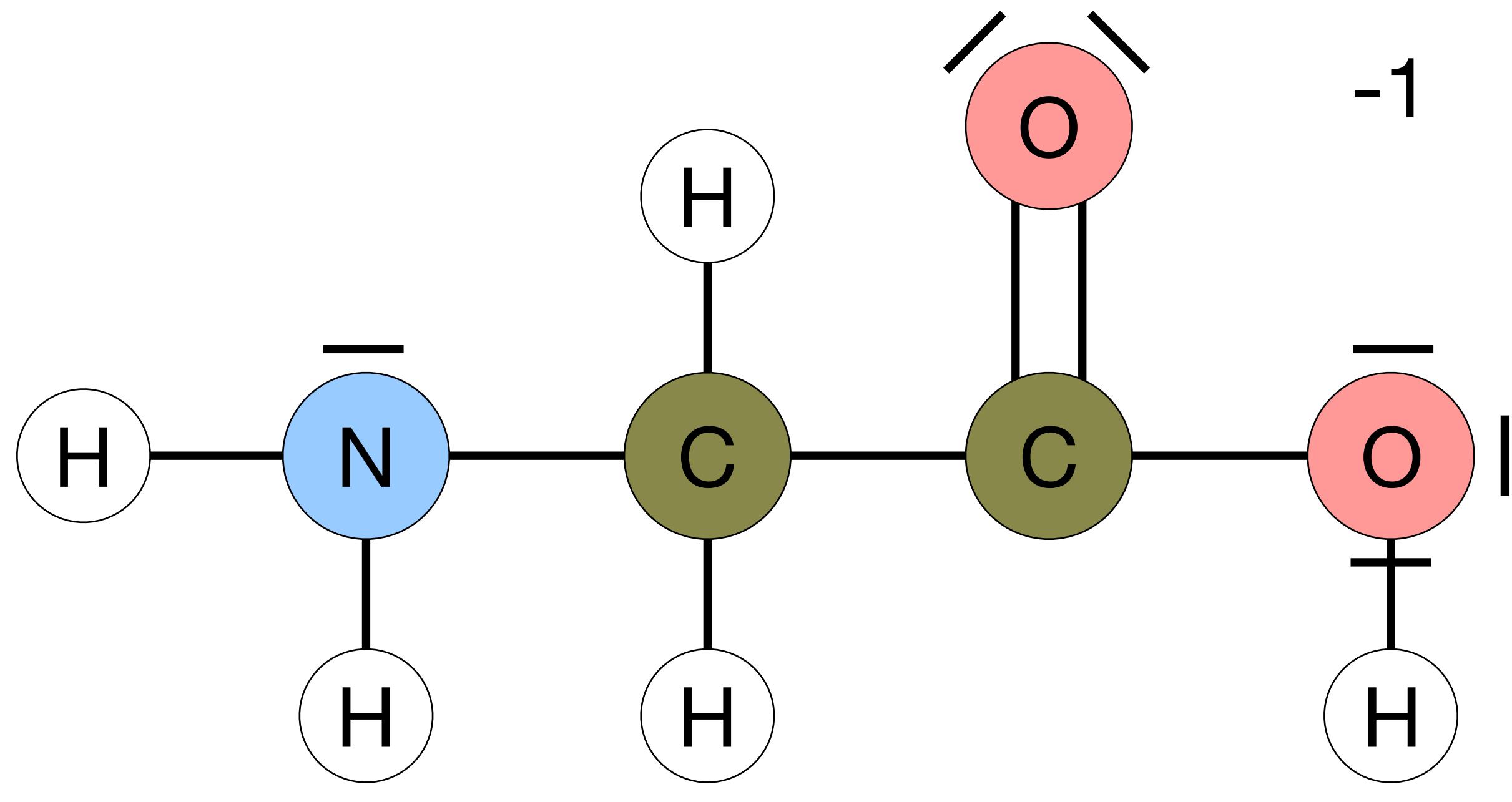


# Octet And 18-Electron Rules

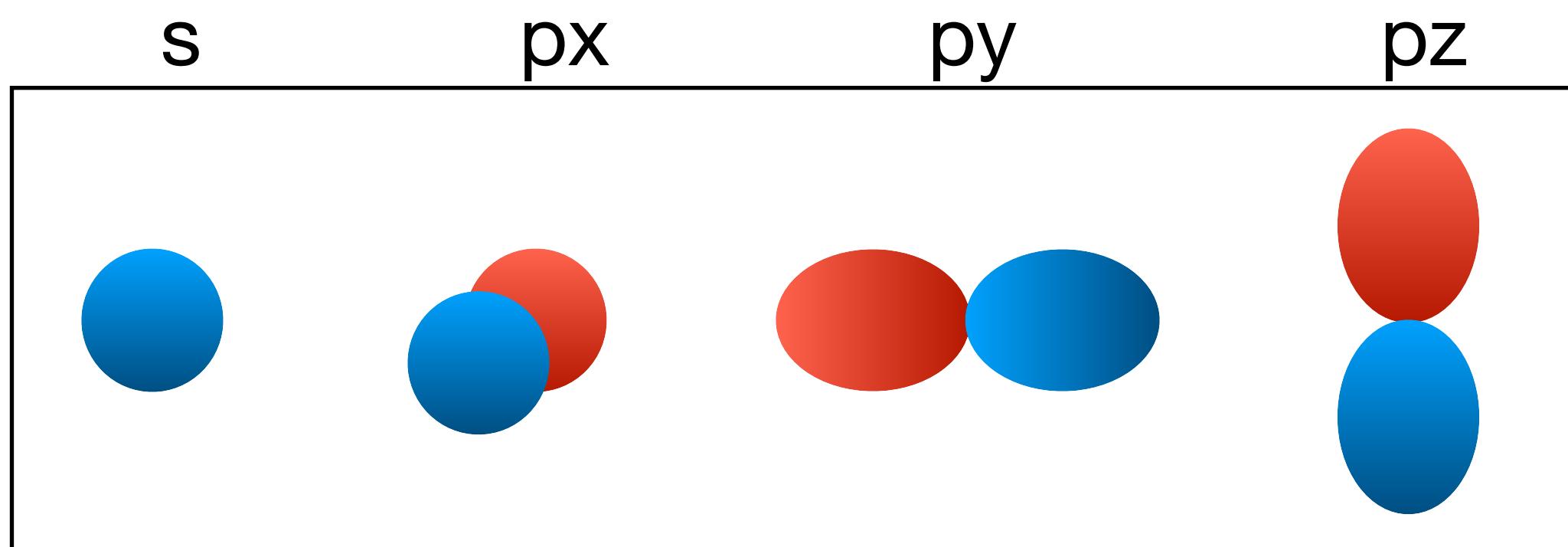
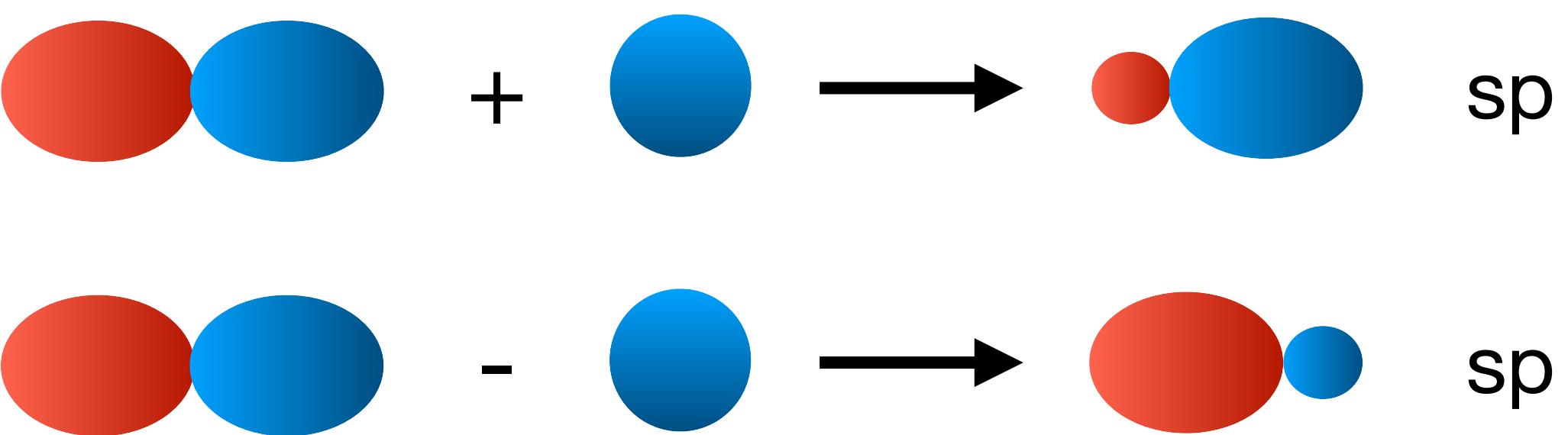
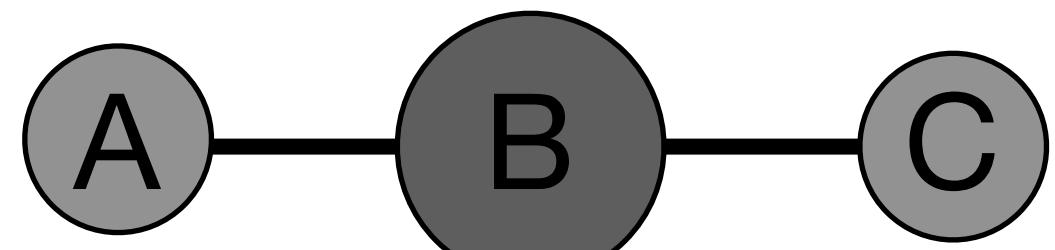
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Group →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Period ↓	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	H															He			
2	Li	Be											B	C	N	O	F	Ne	
3	Na	Mg											Al	Si	P	S	Cl	Ar	
4	K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	
5	Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe	
6	Cs	Ba	*	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	Rn	
7	Fr	Ra	*	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118
	*	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb				
	*	89	90	91	92	93	94	95	96	97	98	99	100	101	102				



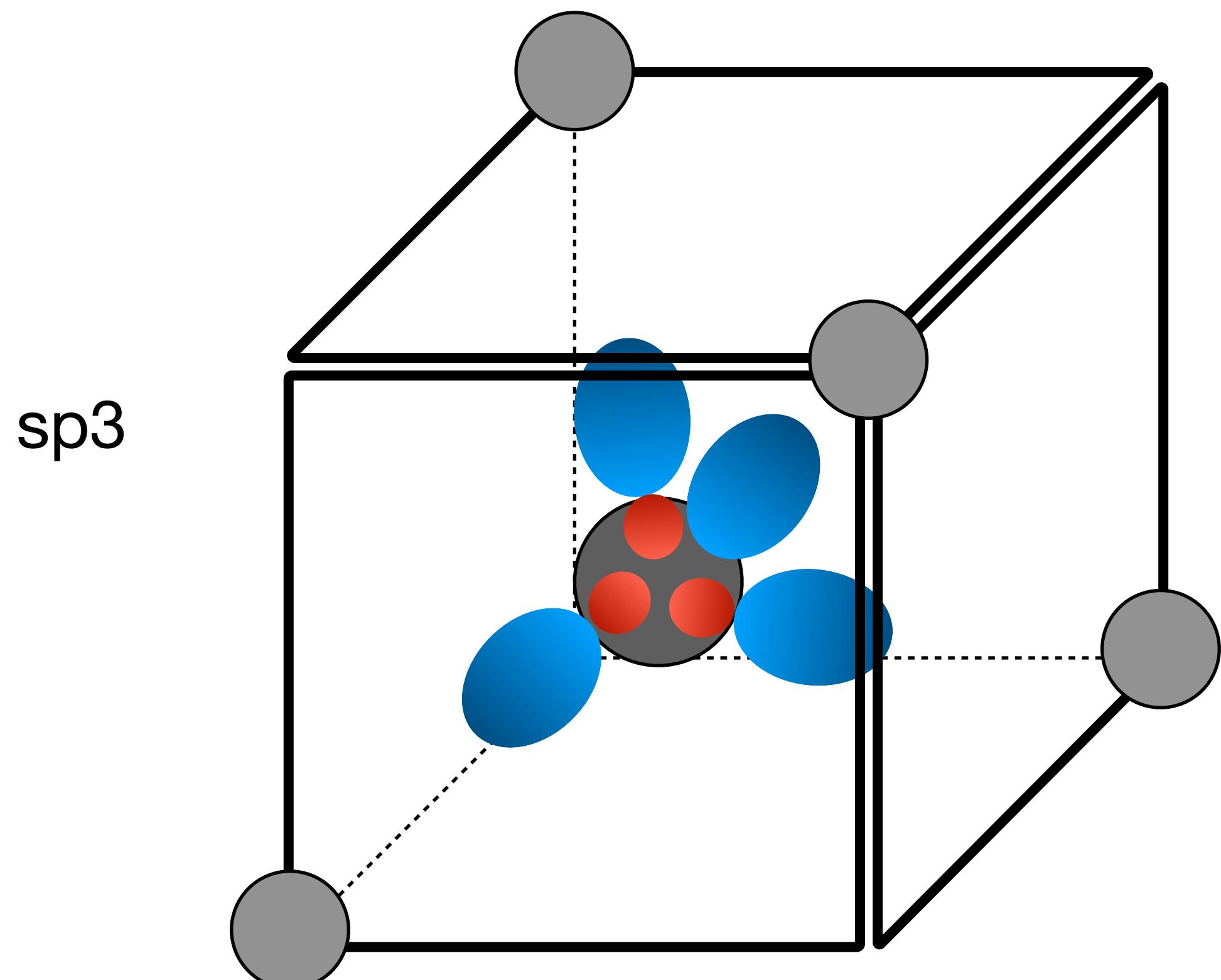
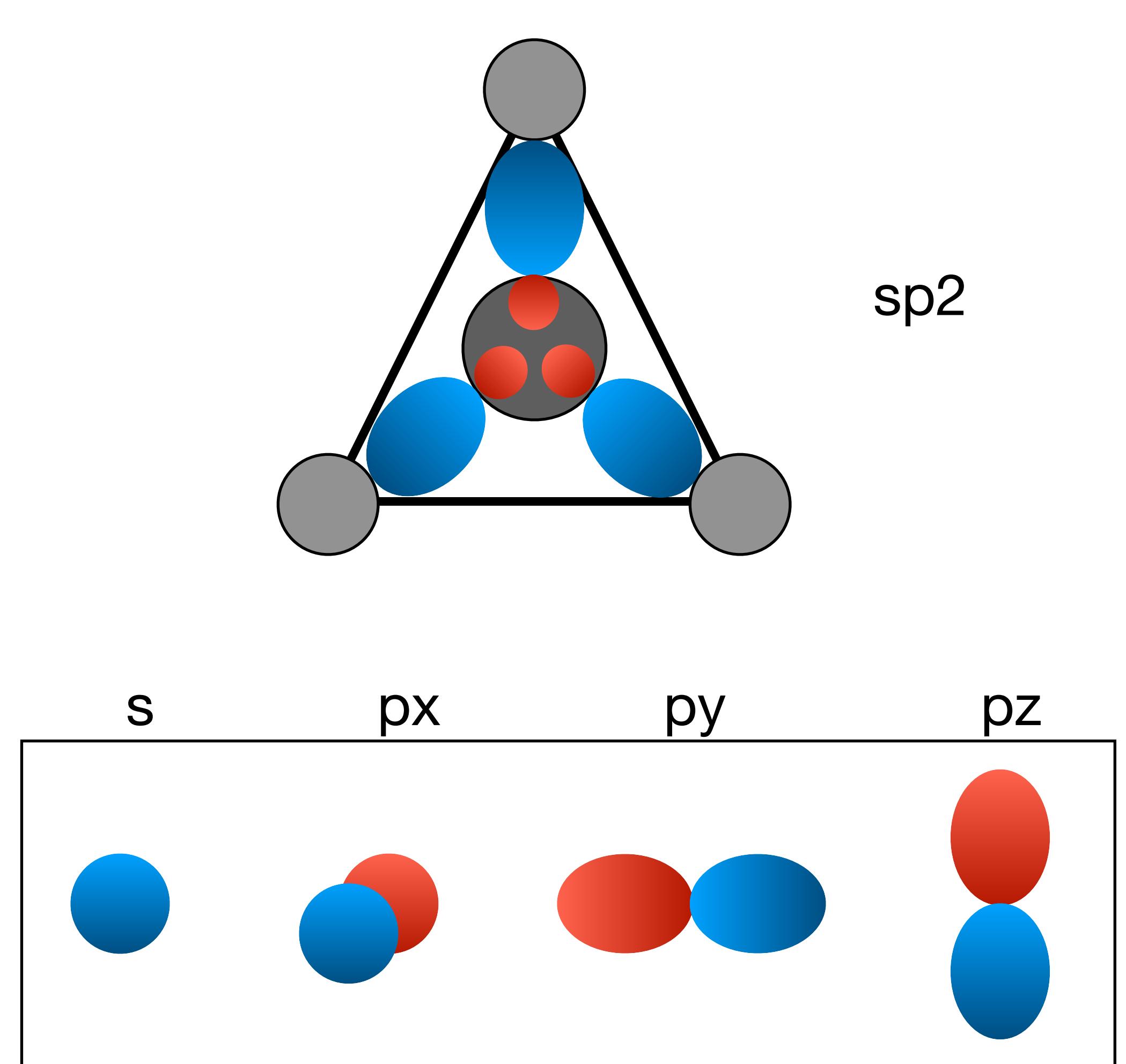
# Lewis Structures



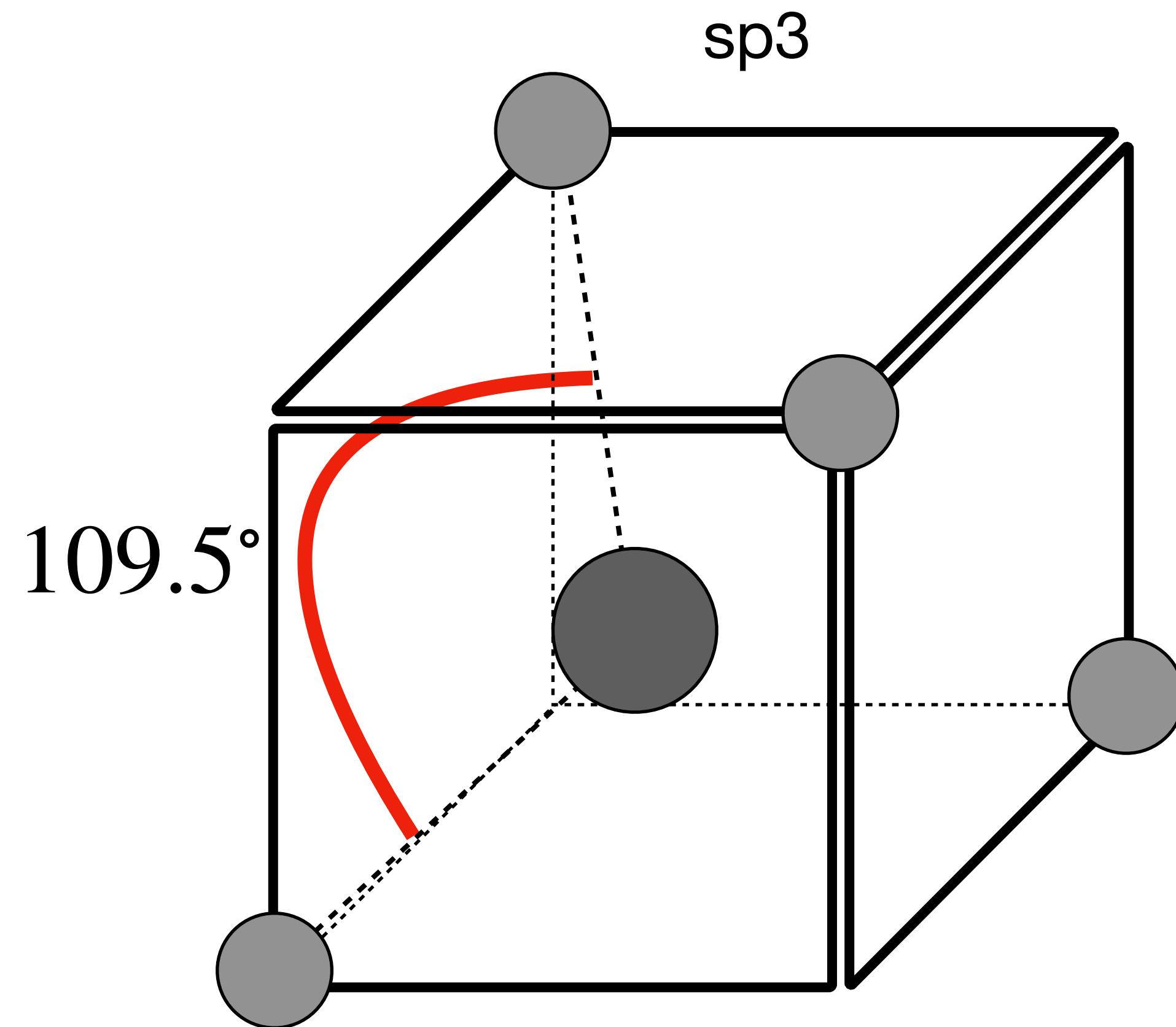
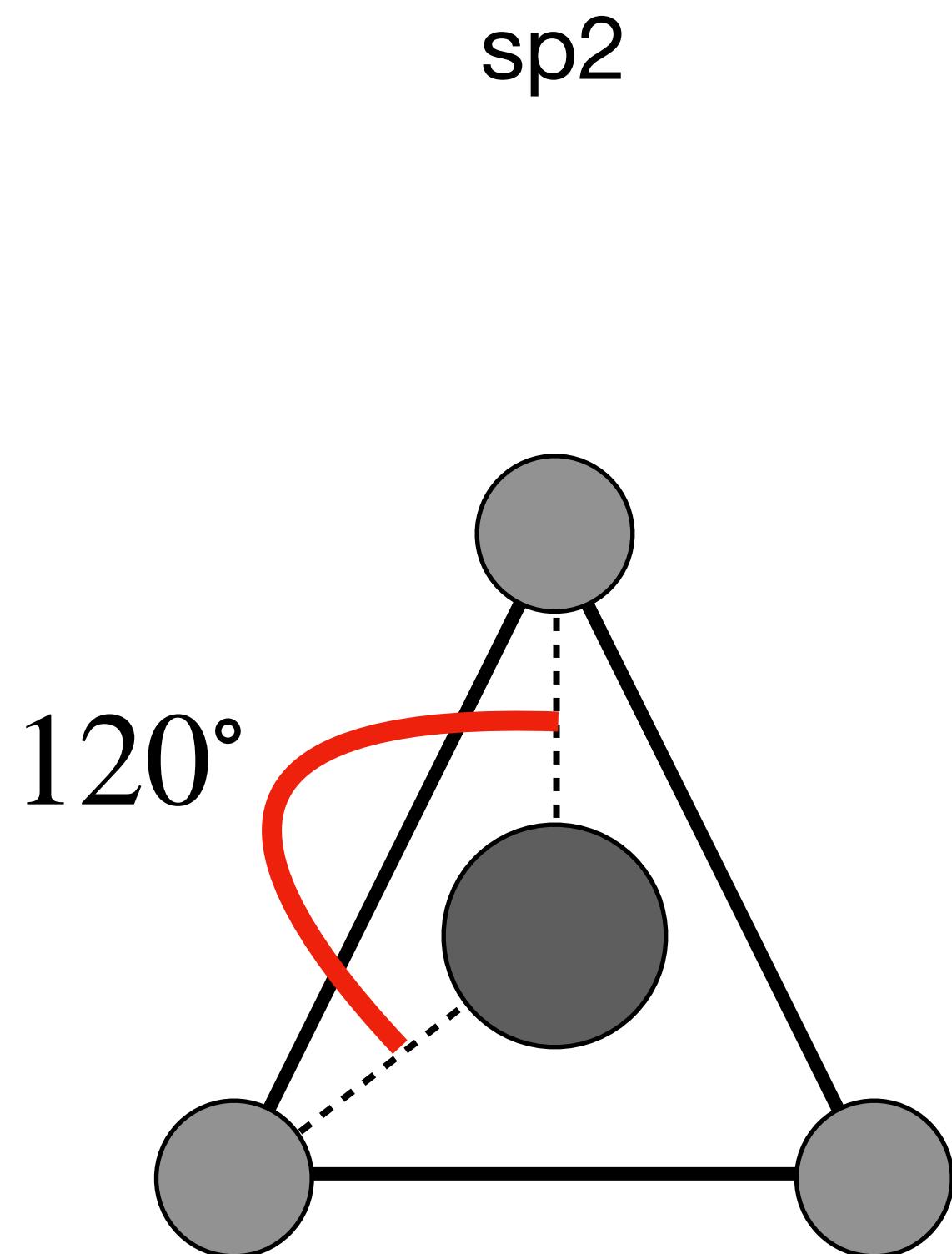
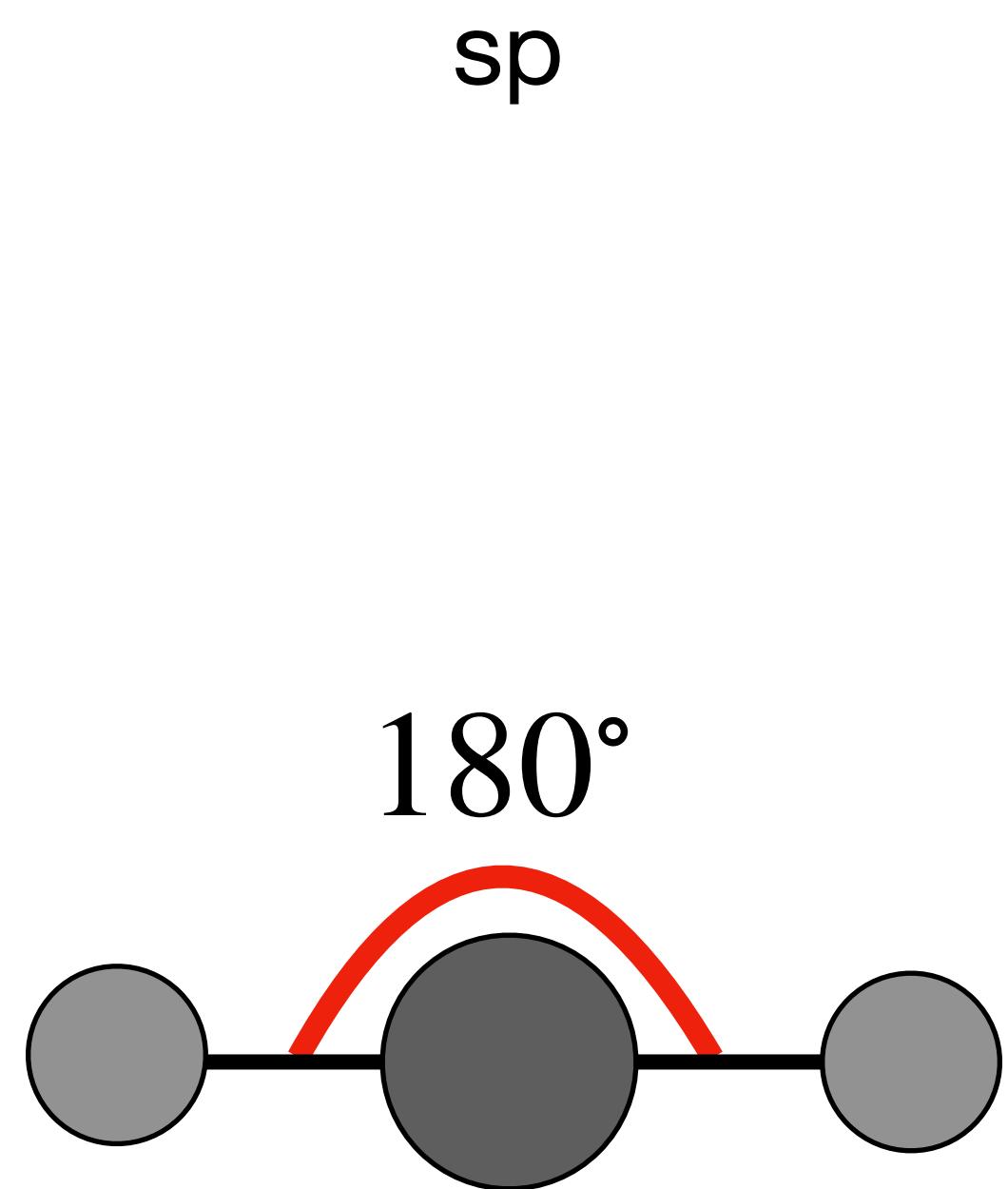
# Orbital Hybridization



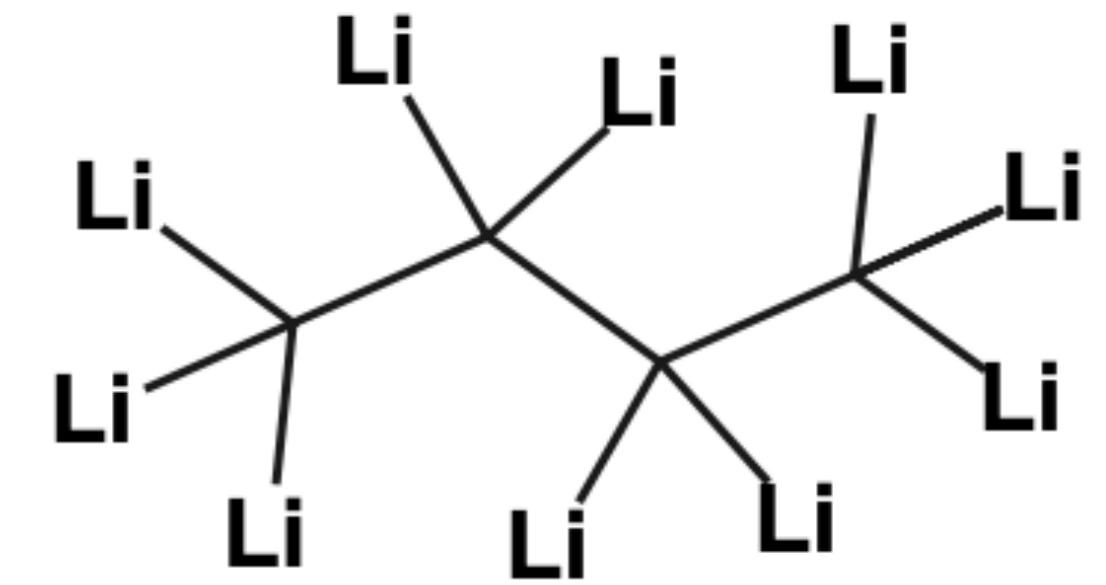
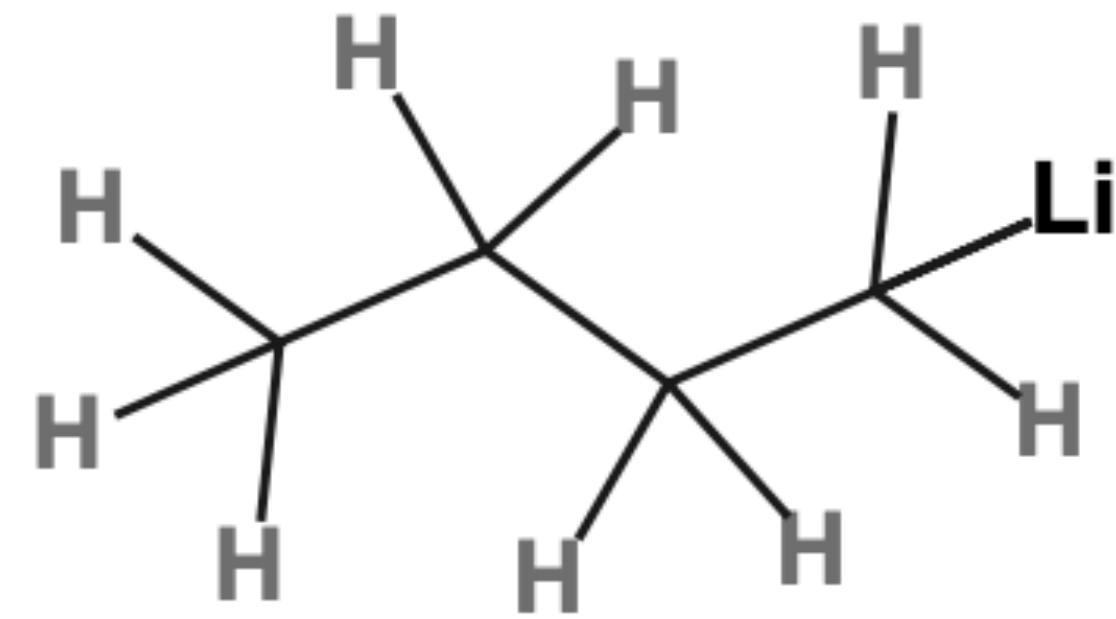
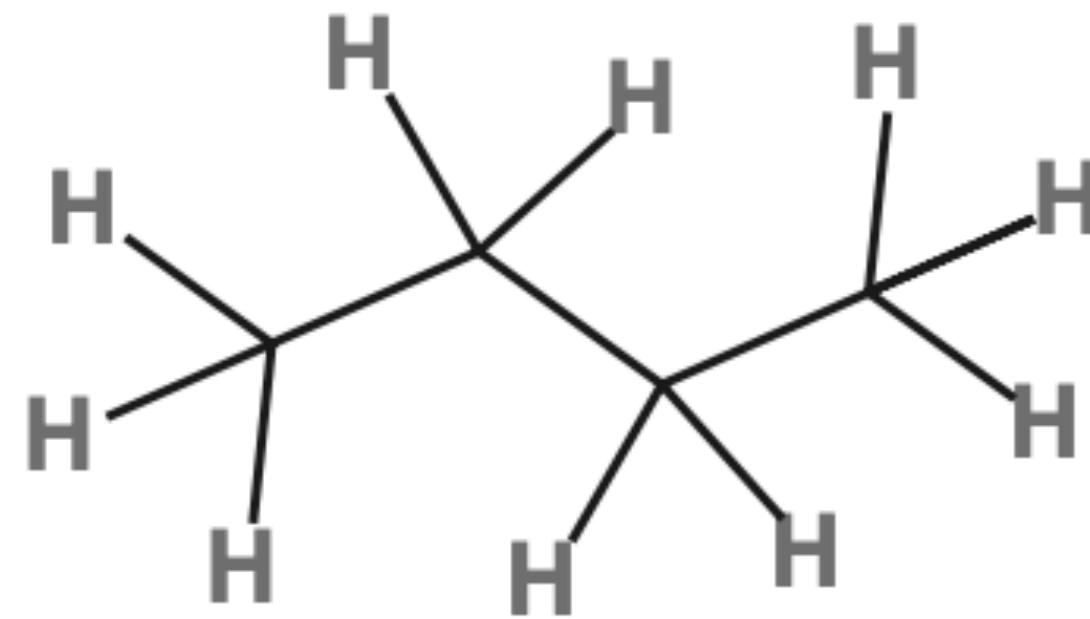
# Orbital Hybridization



# Orbital Hybridization



# To Make It Complicated: Atoms are picky with whom they hang out



[Link](#)

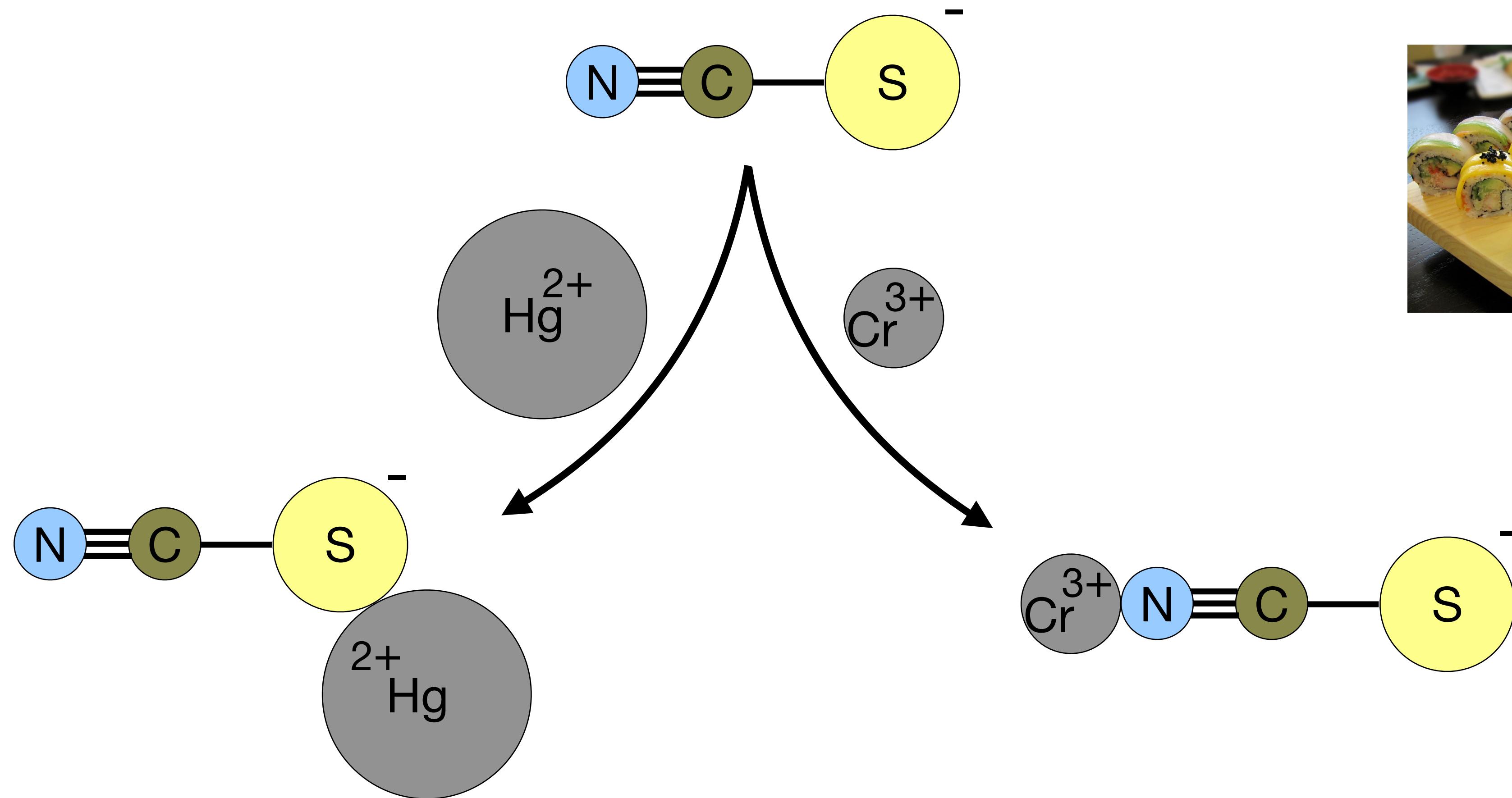


[Link](#)



# Hard-Soft Acid-Base

Atoms are picky with whom they hang out

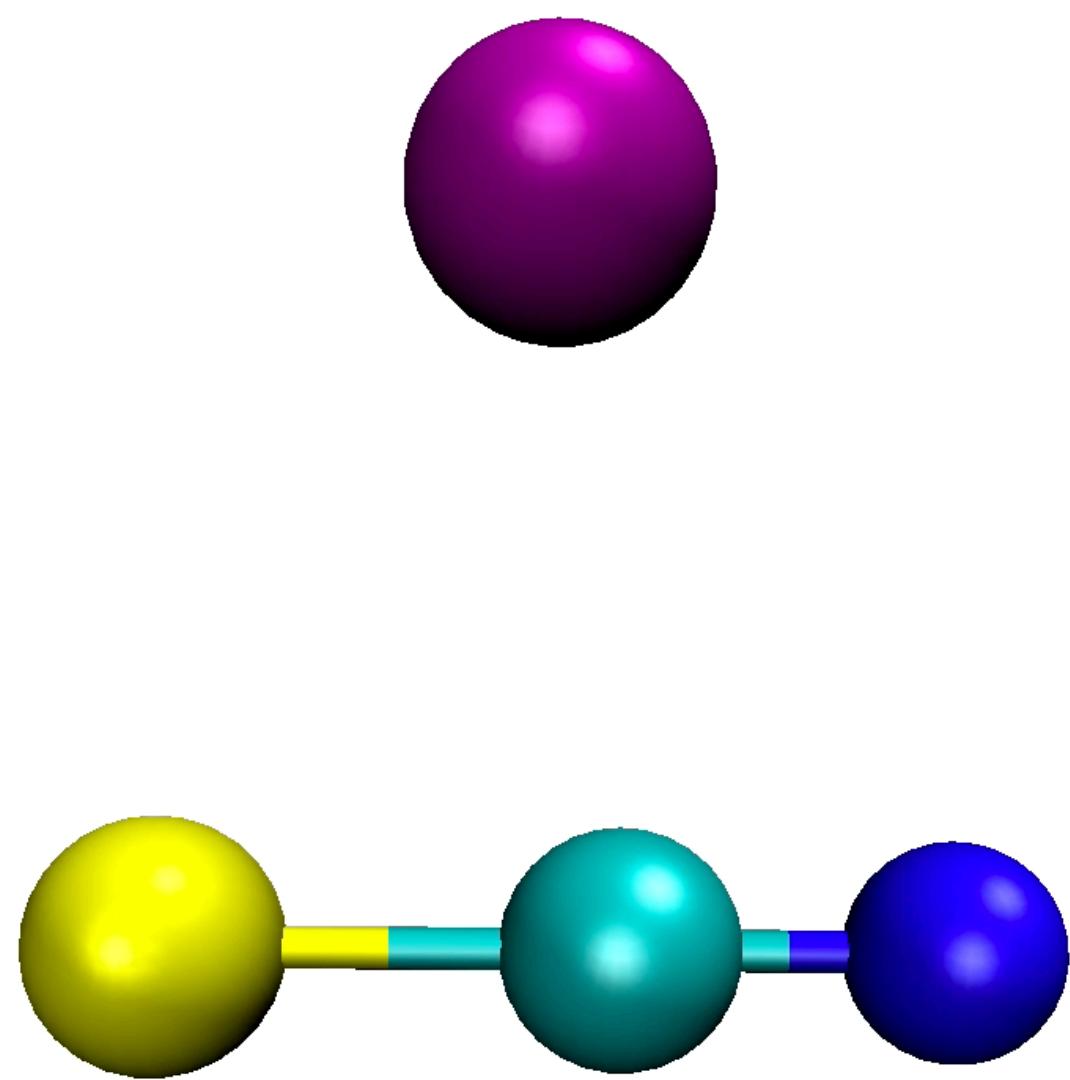


[Link](#)

# Hard-Soft Acid-Base

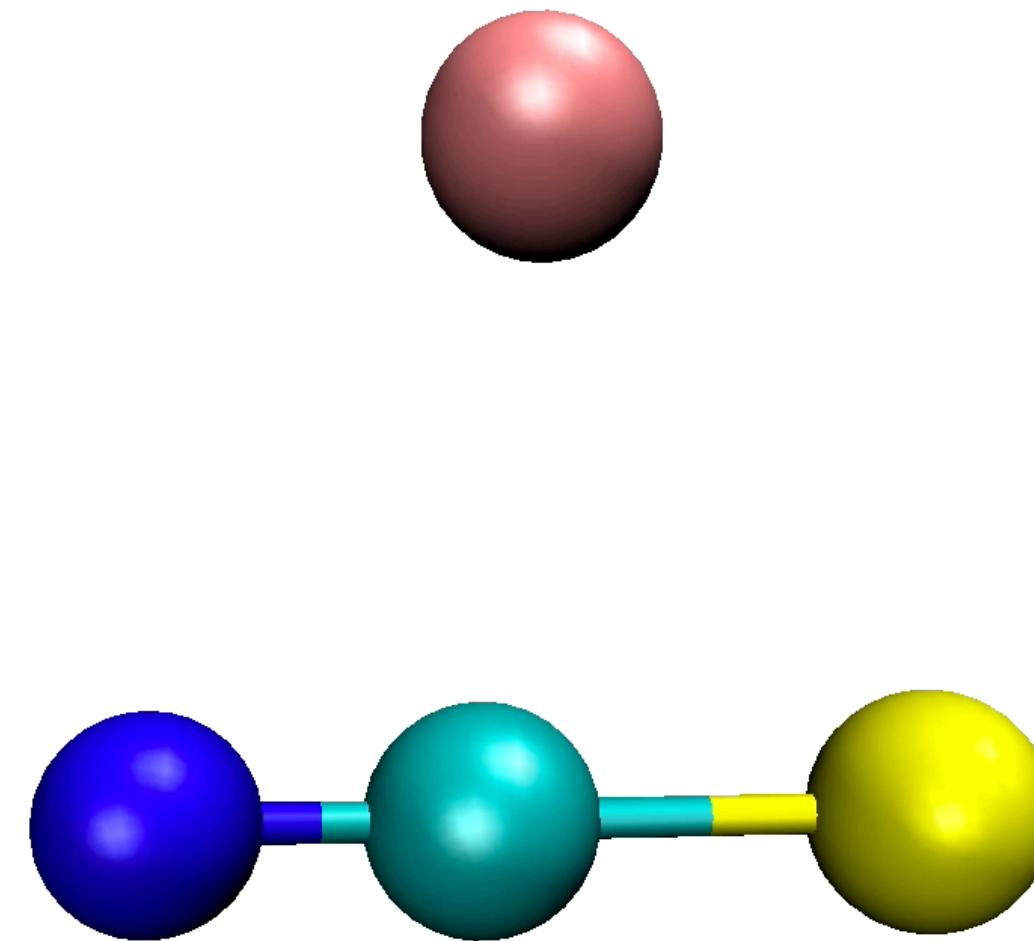
Atoms are picky with whom they hang out

---



$$(s_N | s_{Hg}) = 0.18$$

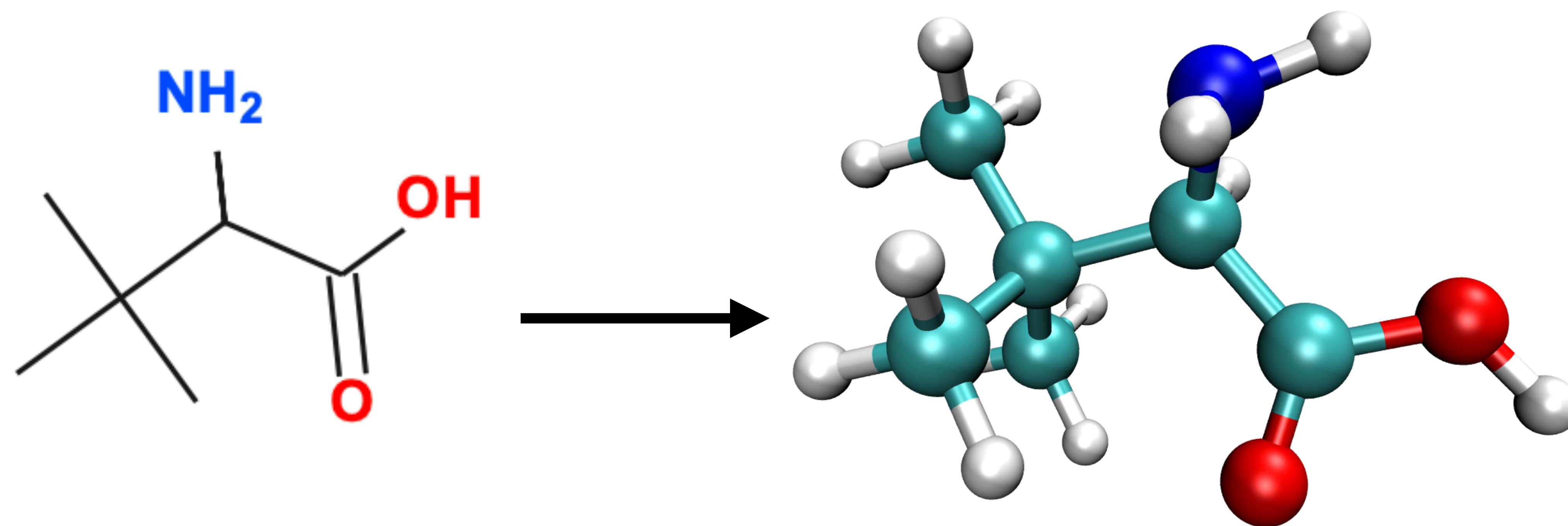
$$(s_S | s_{Hg}) = 0.22$$



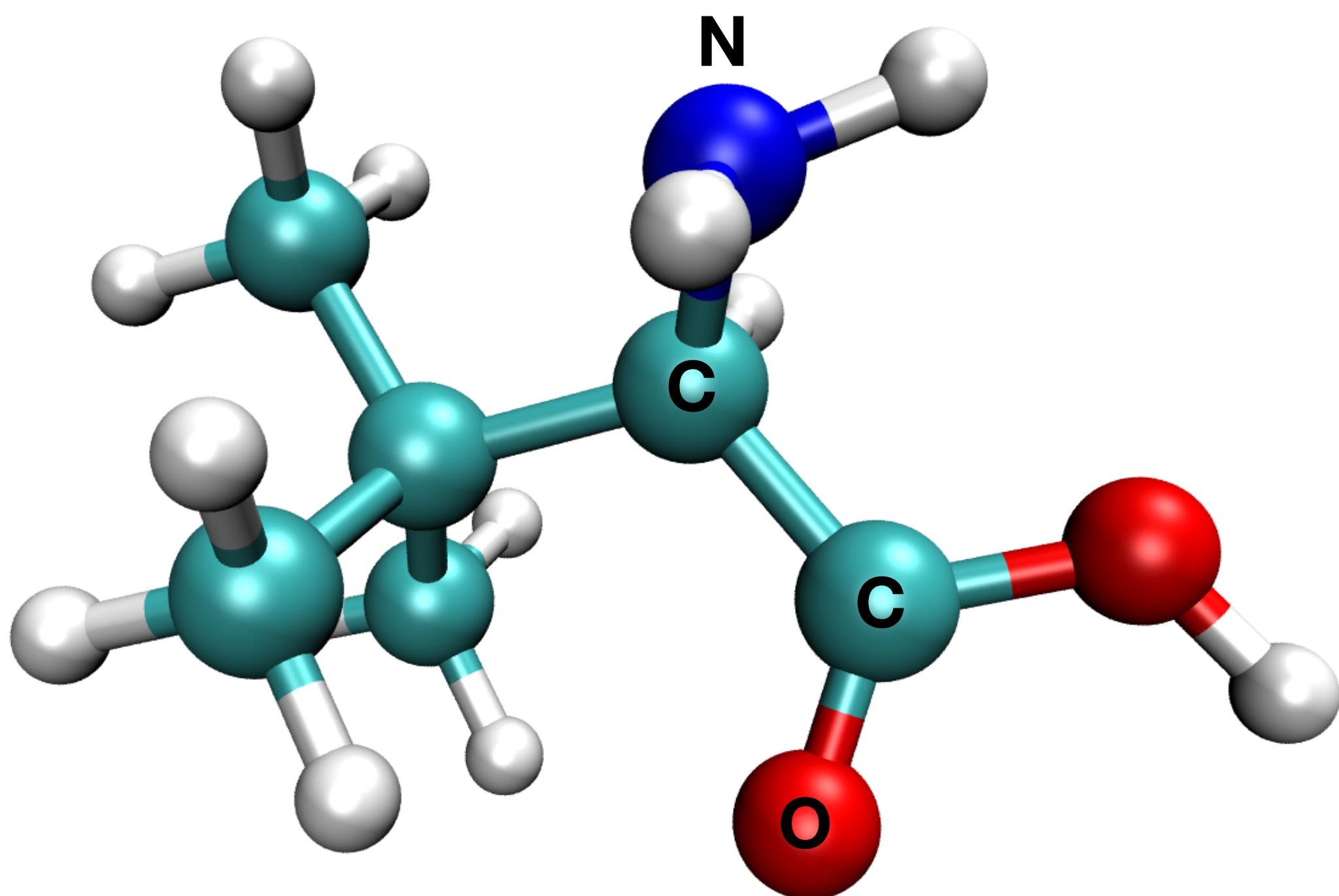
$$(s_N | s_{Cr}) = 0.30$$

$$(s_S | s_{Cr}) = -$$

# From 2D to 3D

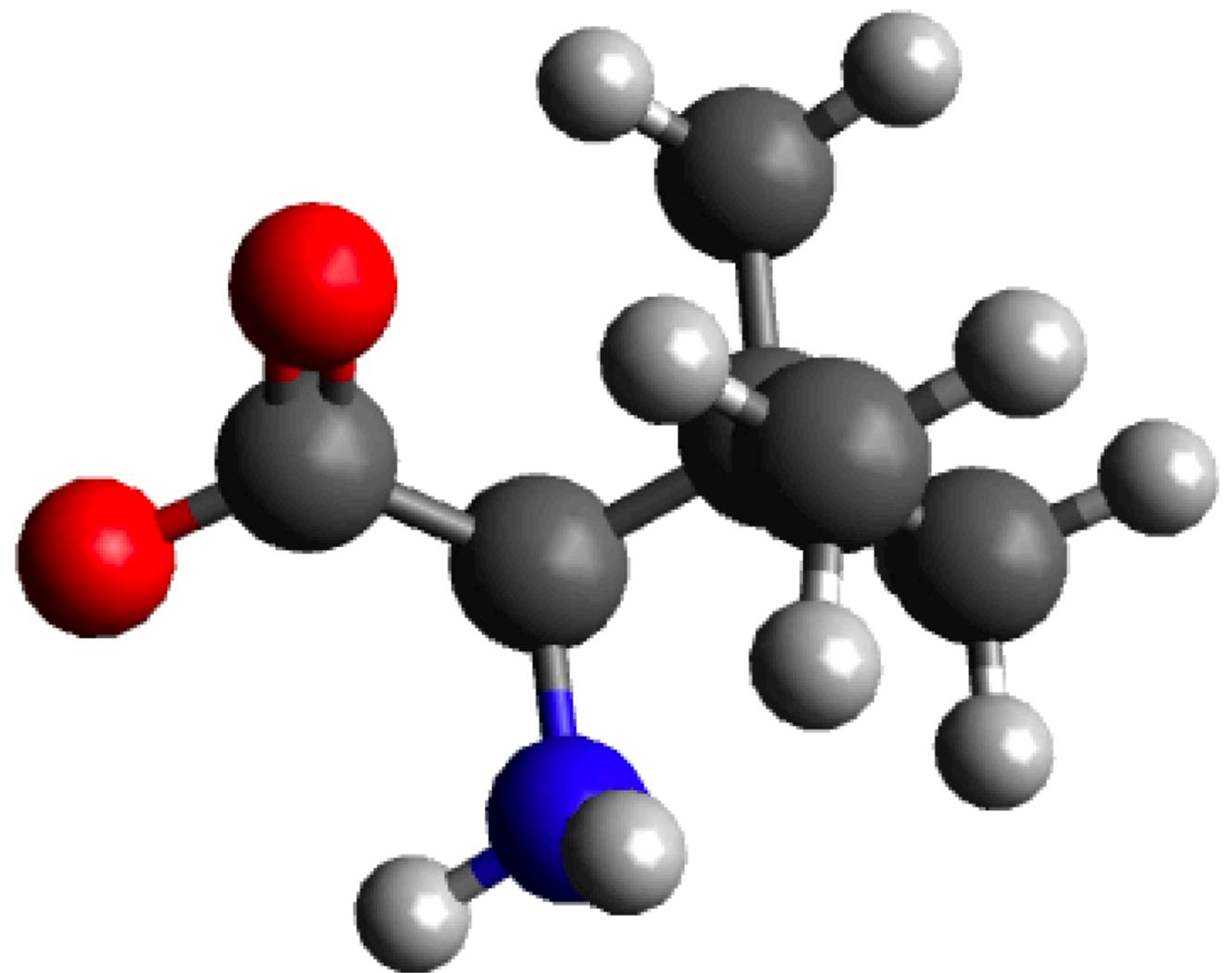


# Z-Matrix



El	#	#	#	#			
C	1						
O	1	1.20					
C	1	1.55	2	120			
N	3	1.45	1	109.5	2	120	
...							

# Z-Matrix



C	1					
O	1	1.20				
C	1	1.55	2	120.0		
N	3	1.45	1	109.5	2	120.0
C	3	2.50	1	150.0	2	360.0
C	5	1.55	3	35.0	1	0.0
H	5	1.10	3	90.0	1	230.0
H	5	1.10	3	90.0	1	130.0
H	5	1.10	3	150.0	1	0.0
C	6	1.55	5	109.5	3	240.0
C	6	1.55	5	109.5	3	120.0
H	10	1.10	6	109.5	5	60.0
H	10	1.10	6	109.5	5	290.0
H	10	1.10	6	109.5	5	180.0
H	11	1.10	6	109.5	5	290.0
H	11	1.10	6	109.5	5	180.0
H	11	1.10	6	109.5	5	60.0
H	3	1.10	1	109.5	2	240.0
O	1	1.35	2	120.0	3	180.0
H	4	1.10	3	109.5	1	290.0
H	4	1.10	3	109.5	1	60.0
H	19	0.10	1	120.0	2	0.0

# Automated 3D-Structure Generation



Open-Source Cheminformatics  
and Machine Learning

c1ccccc1



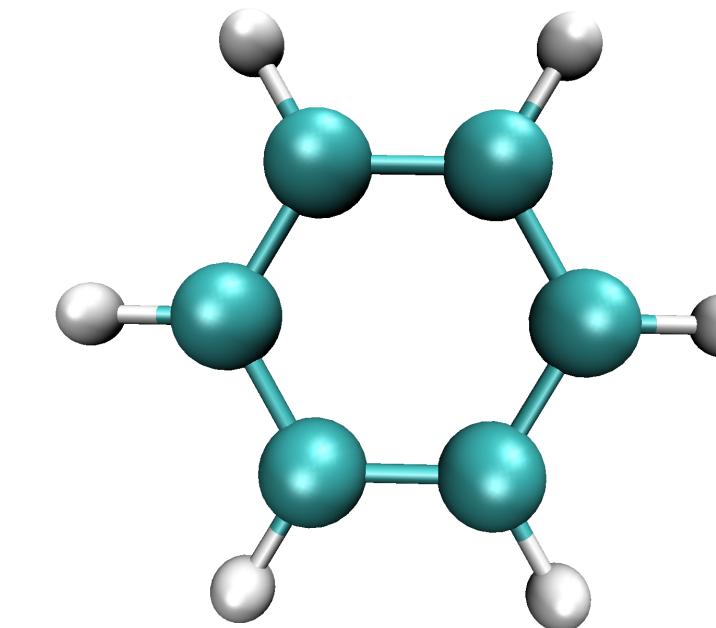
Use rules above to connect atoms



(use fragments to simplify)

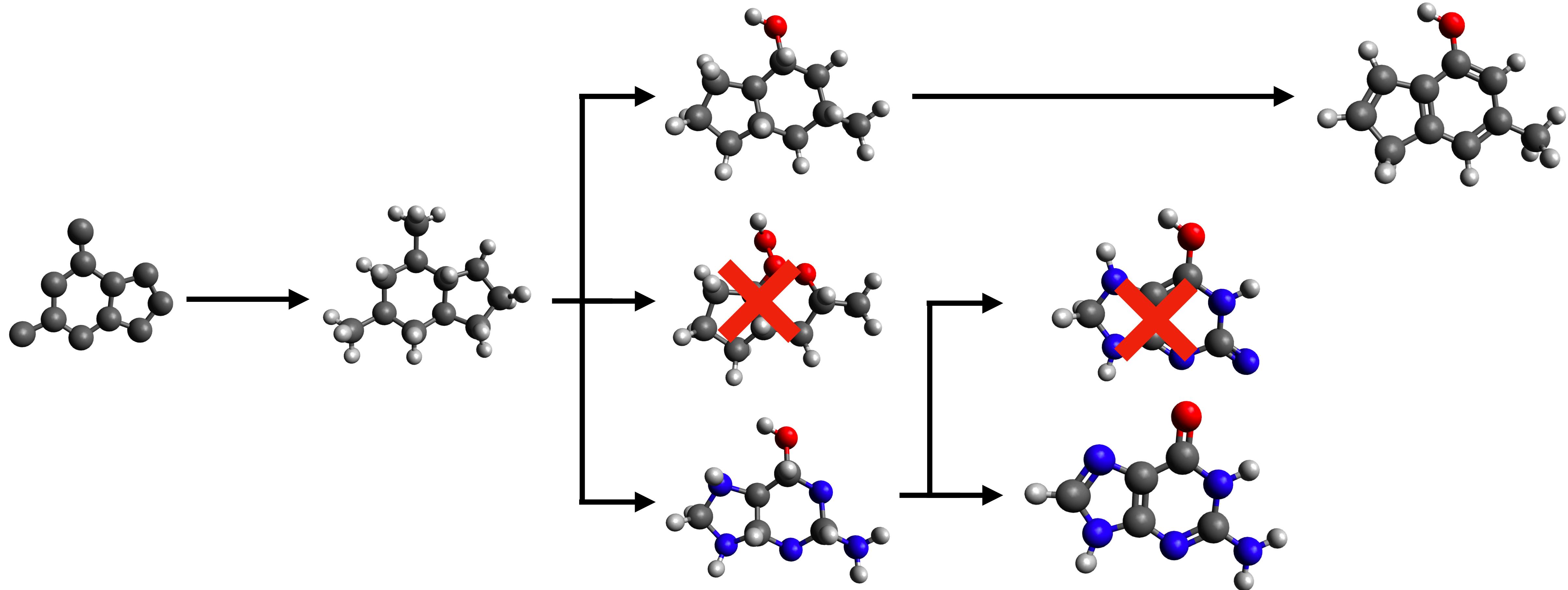


Optimize the Geometry (FF or QM) →

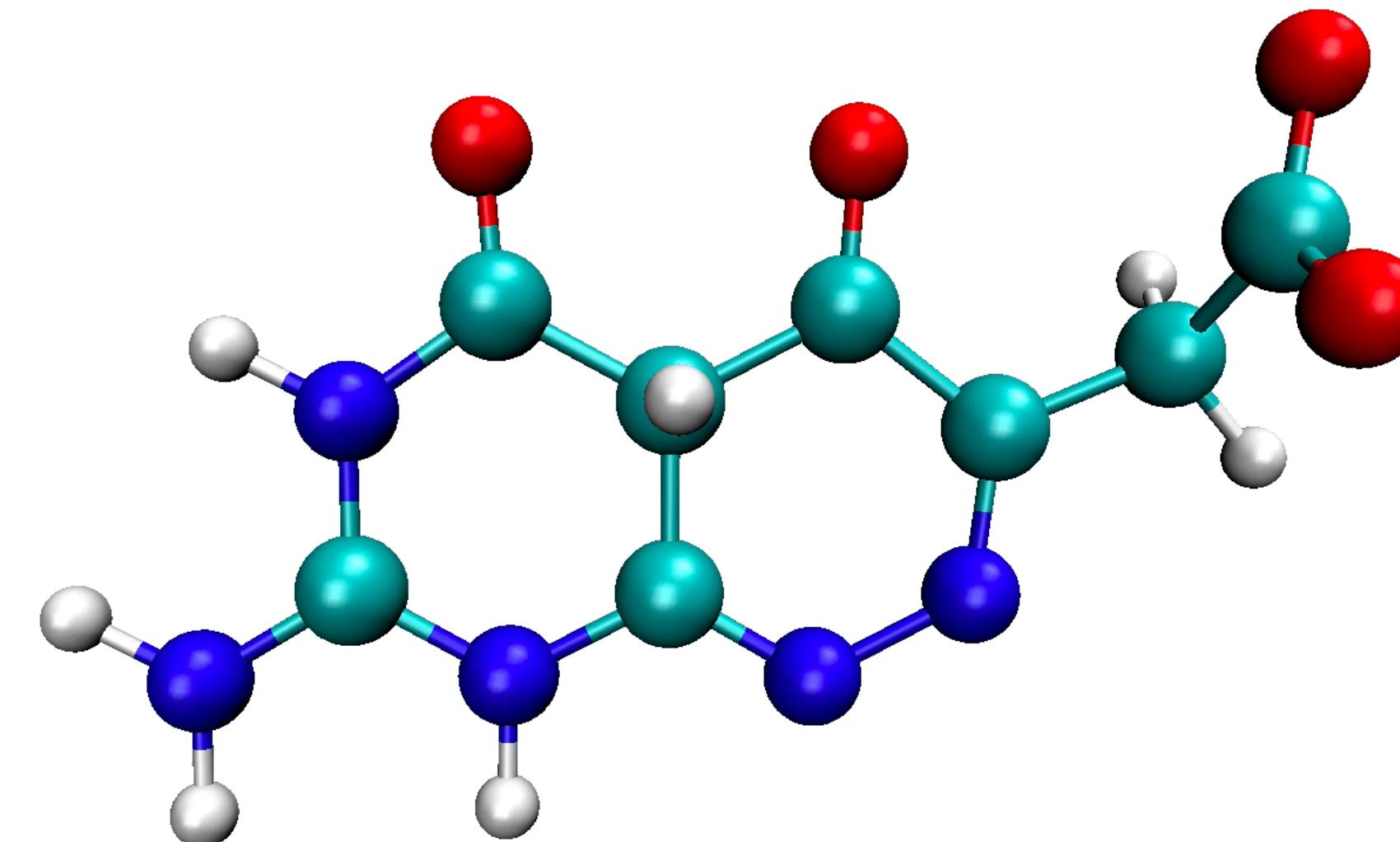
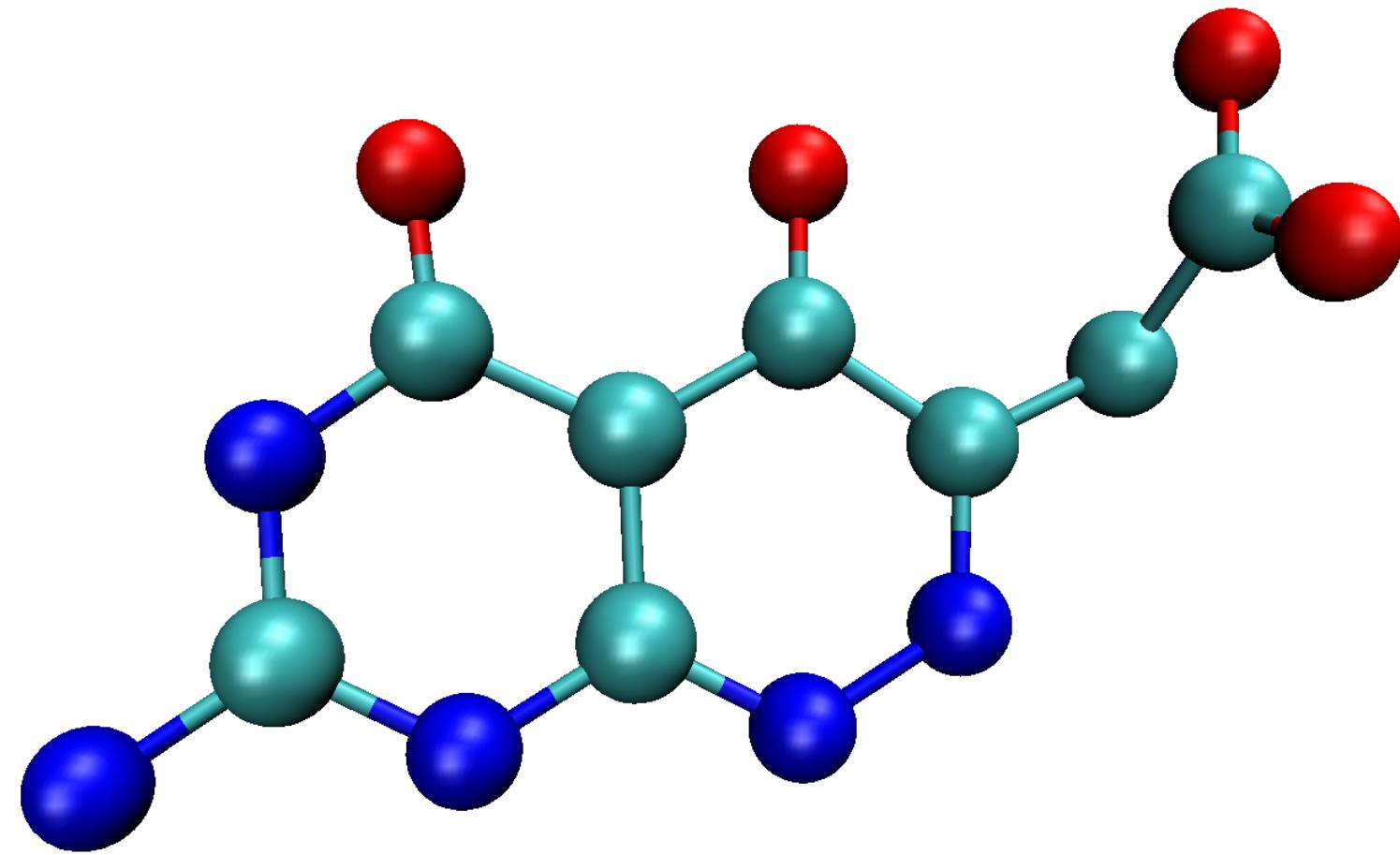


# Automated 3D-Structure Generation

QUANTUM-MACHINE.ORG

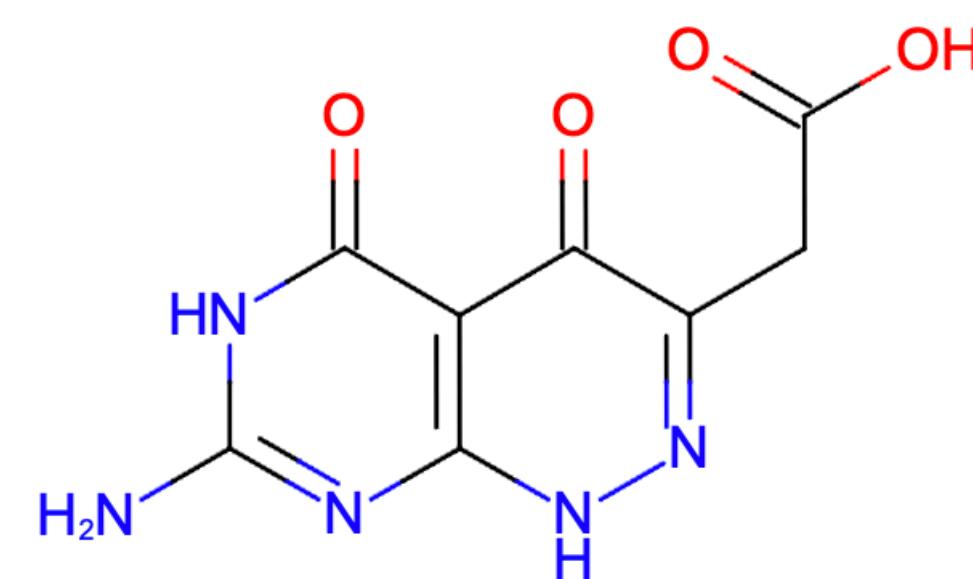


# What can go wrong... will go wrong

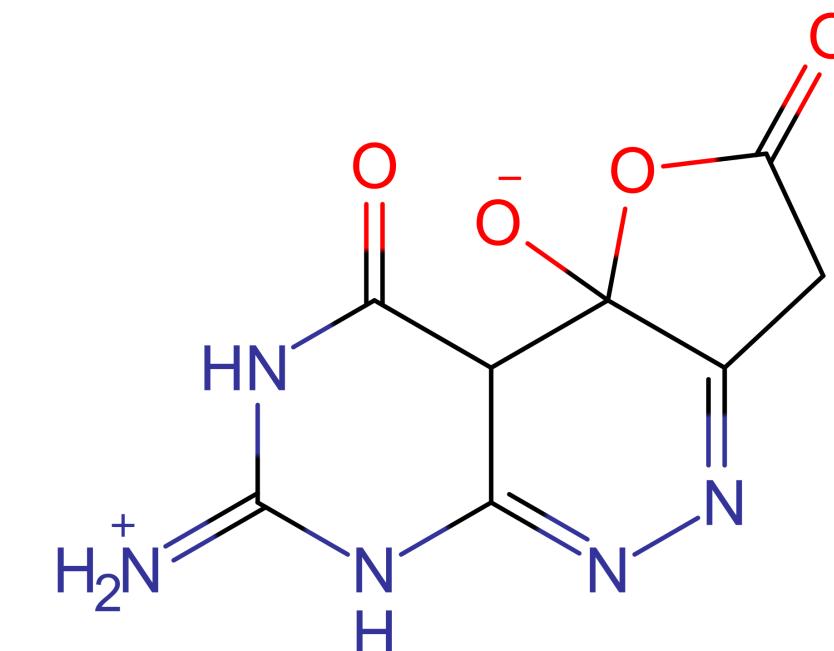


PDBbind

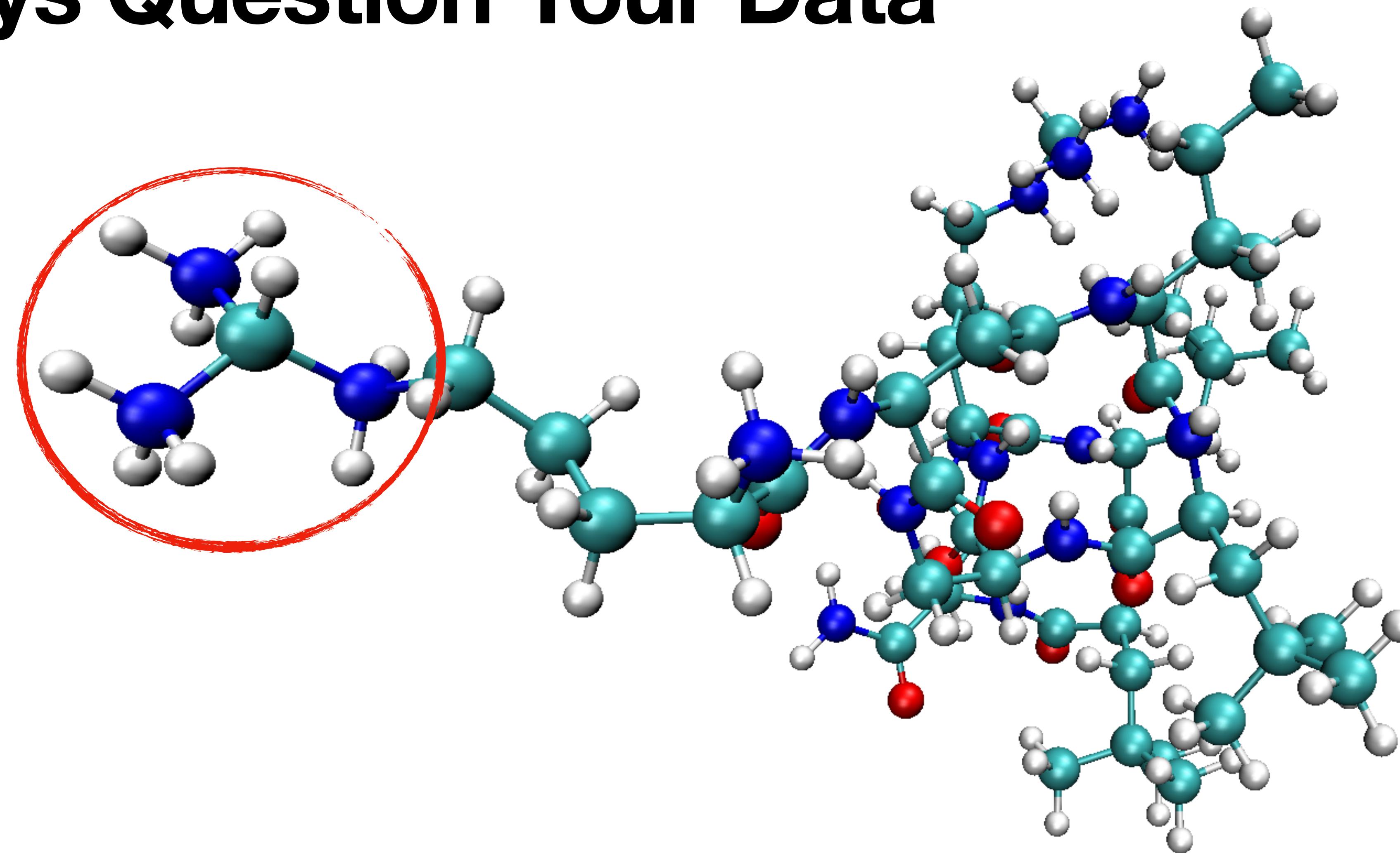
PDB <https://www.rcsb.org/structure/4DAI>



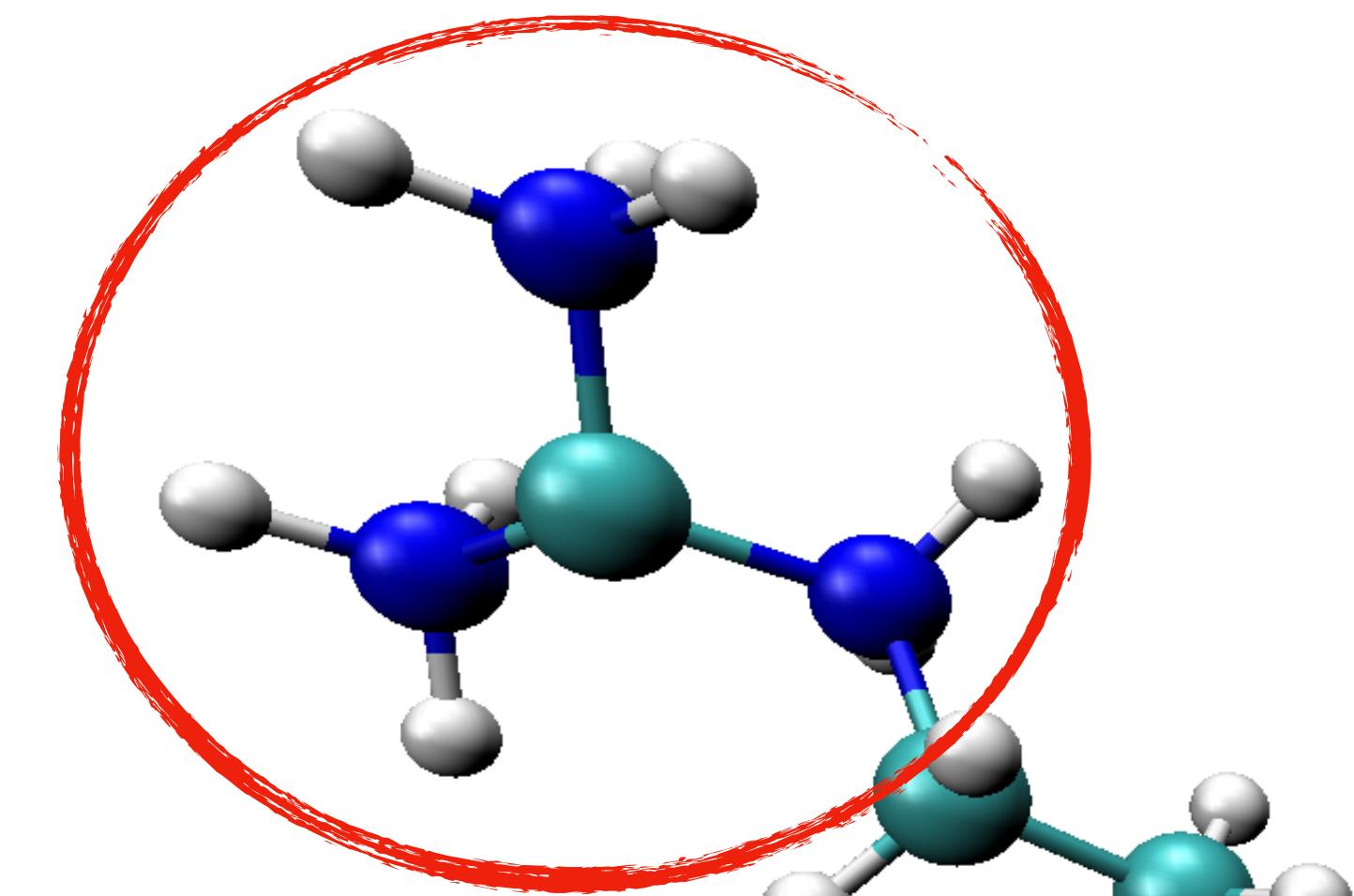
4DAI



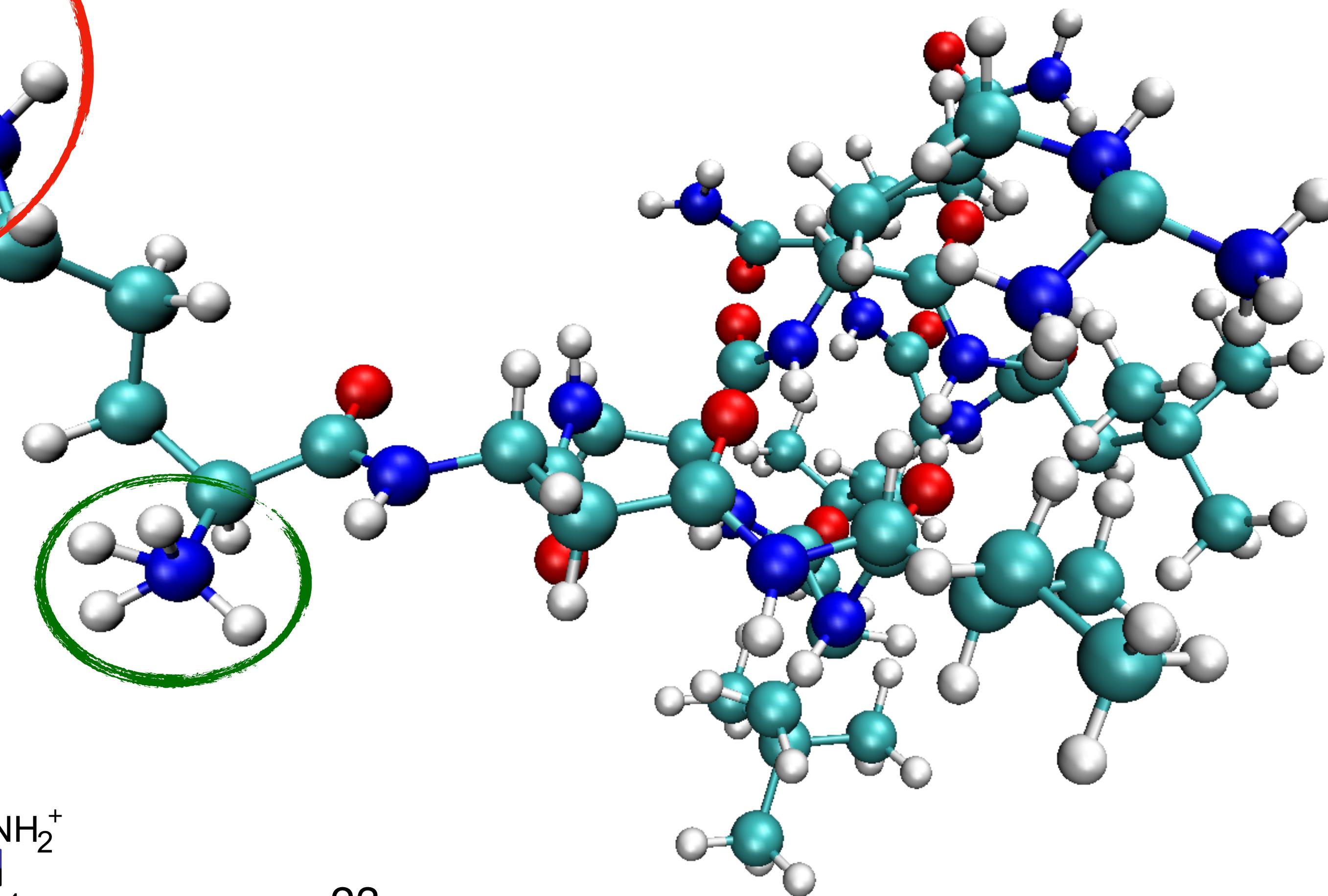
# Always Question Your Data



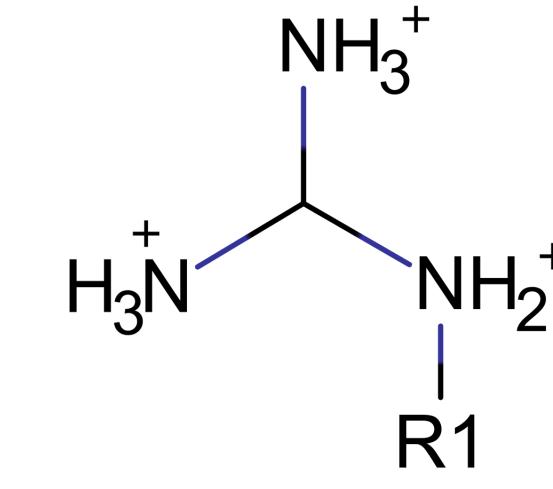
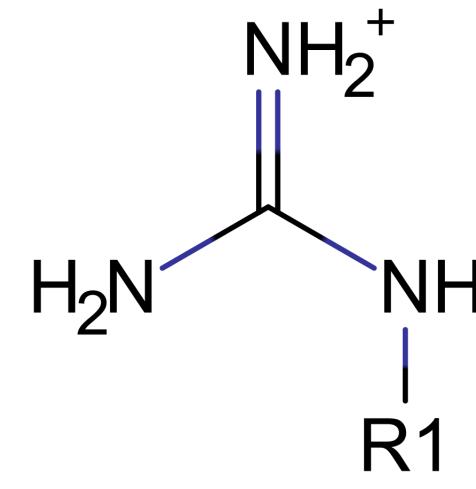
# So fix it with widely used software



Explicit violation of octet rule!

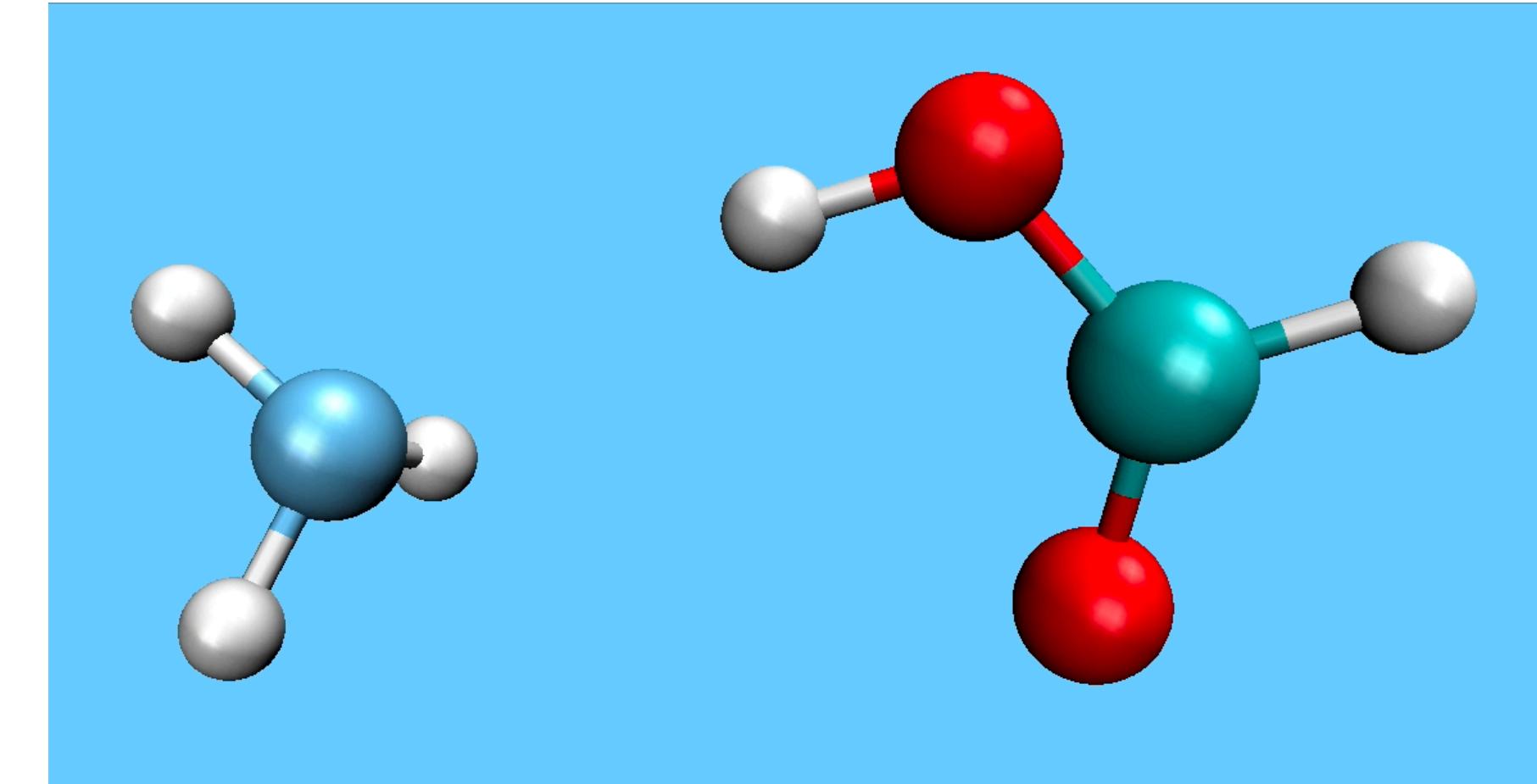
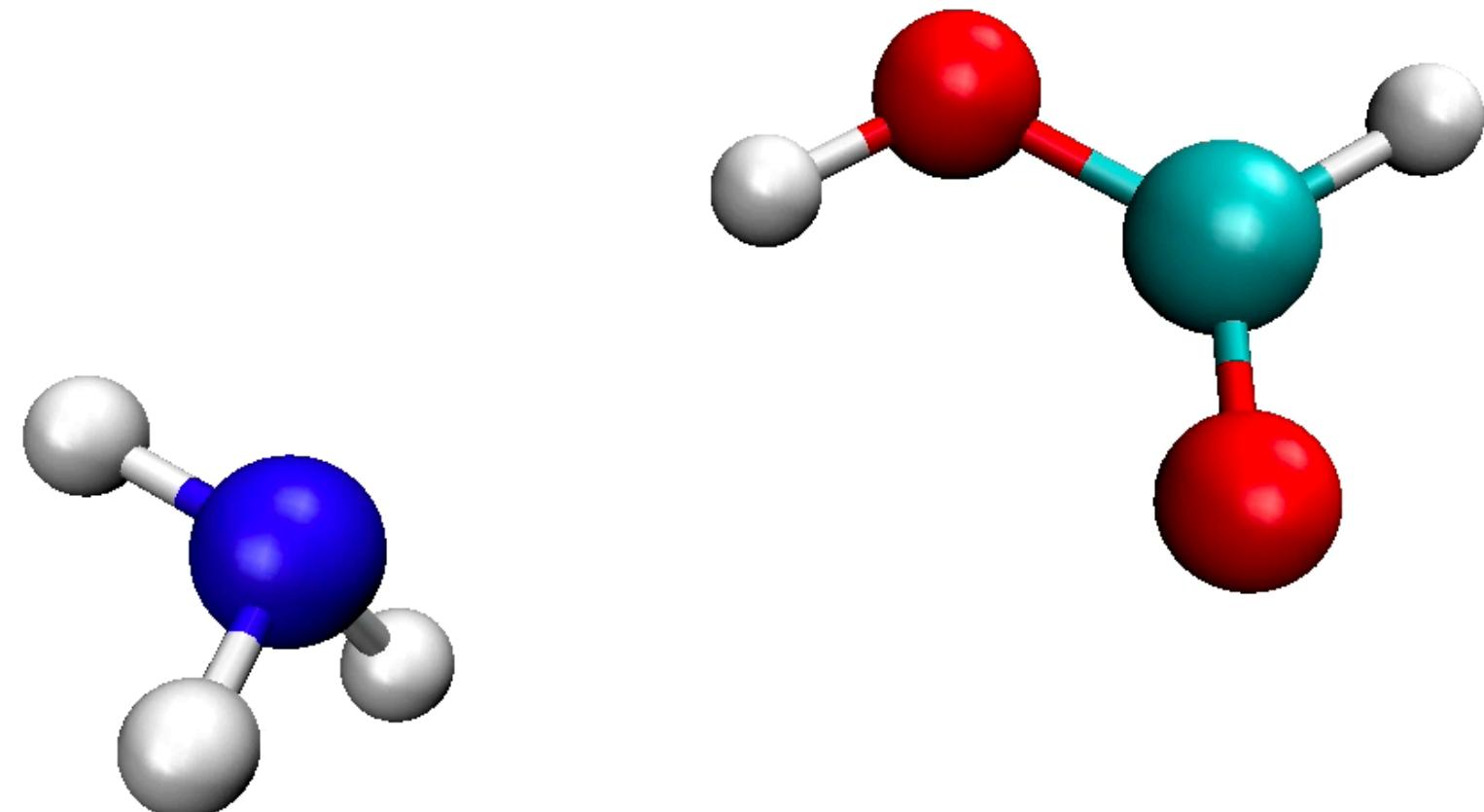


5GTR



# Tipps and Tricks

- Follow the rules above and the structure will be reasonable (at least in gas).
- Some groups like to be charged. Use pKa and pKb tables.



- Rather consistently wrong, than sometimes (inconsistently) correct.

# Tipps and Tricks

- Use Databases for double-checking

The screenshot shows the "SDBS Compounds and Spectral Search" page. It features a search form with fields for Compound Name, Molecular Formula, Molecular Weight, CAS Registry No., SDBS No., and various spectral parameters like IR Peaks, <sup>13</sup>C NMR Shift, and MS Peaks. The interface is in Japanese and includes a yellow header bar with the AIST logo.

**Spectral Database for Organic Compounds SDBS**

**SDBS Compounds and Spectral Search**

**Compound Name:**  match partial

**Molecular Formula:**   
C, H, then the other elements are alphabetical order, "%,\*" for the wild card

**Molecular Weight:**  to   
Numbers between left and right columns  
Up to the first place of a decimal point

**CAS Registry No.:**   
"%,\*" for the wild card.

**SDBS No.:**   
"%,\*" for the wild card.

**Atoms:** C(Carbon)  to  H(Hydrogen)  to  N(Nitrogen)  to  O(Oxygen)  to  F(Fluorine)  to  Cl(Chlorine)  to  Br(Bromine)  to  I(Iodine)  to  S(Sulfur)  to  P(Phosphorus)  to  Si(Silicon)  to

**Spectrum:** Check the spectra of your interest.  
 MS  IR  
 <sup>13</sup>C NMR  Raman  
 <sup>1</sup>H NMR  ESR

**IR Peaks(cm<sup>-1</sup>):** Allowance  ± 10  
"," or space is the separator for multiple peaks.  
Use "-", to set a range: eg. 550-750,1650-3000-

**Transmittance <**  %

**<sup>13</sup>C NMR Shift(ppm):** Allowance  ± 2.0  
"," is the separator for multiple shifts, eg. 129.3,18.4....

**No shift regions:**   
Range defined by two numbers separated by a space, eg. 110 78,...

**<sup>1</sup>H NMR Shift(ppm):** Allowance  ± 0.2

**No shift regions:**

**MS Peaks and intensities:**   
Mass and its intensity are a set of data separated by a space, eg. 110 22,...

**Sort by:** Molecular Weight  Ascending Order  Result Display type:  with Structures

**(c) National Institute of Advanced Industrial Science and Technology (AIST)**

Link

The screenshot shows the "Explore Chemistry" page of PubChem. It features a search bar, a "Try" button with example queries, and links to Draw Structure, Upload ID List, Browse Data, and Periodic Table. At the bottom, it displays statistics: 110M Compounds, 277M Substances, 293M Bioactivities, and 843 Data Sources.

**National Library of Medicine**  
National Center for Biotechnology Information

**PubChem** About Blog Submit Contact

**Explore Chemistry**

Quickly find chemical information from authoritative sources

Try covid-19 aspirin EGFR C9H8O4 57-27-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3(2)/4/h1-2H3

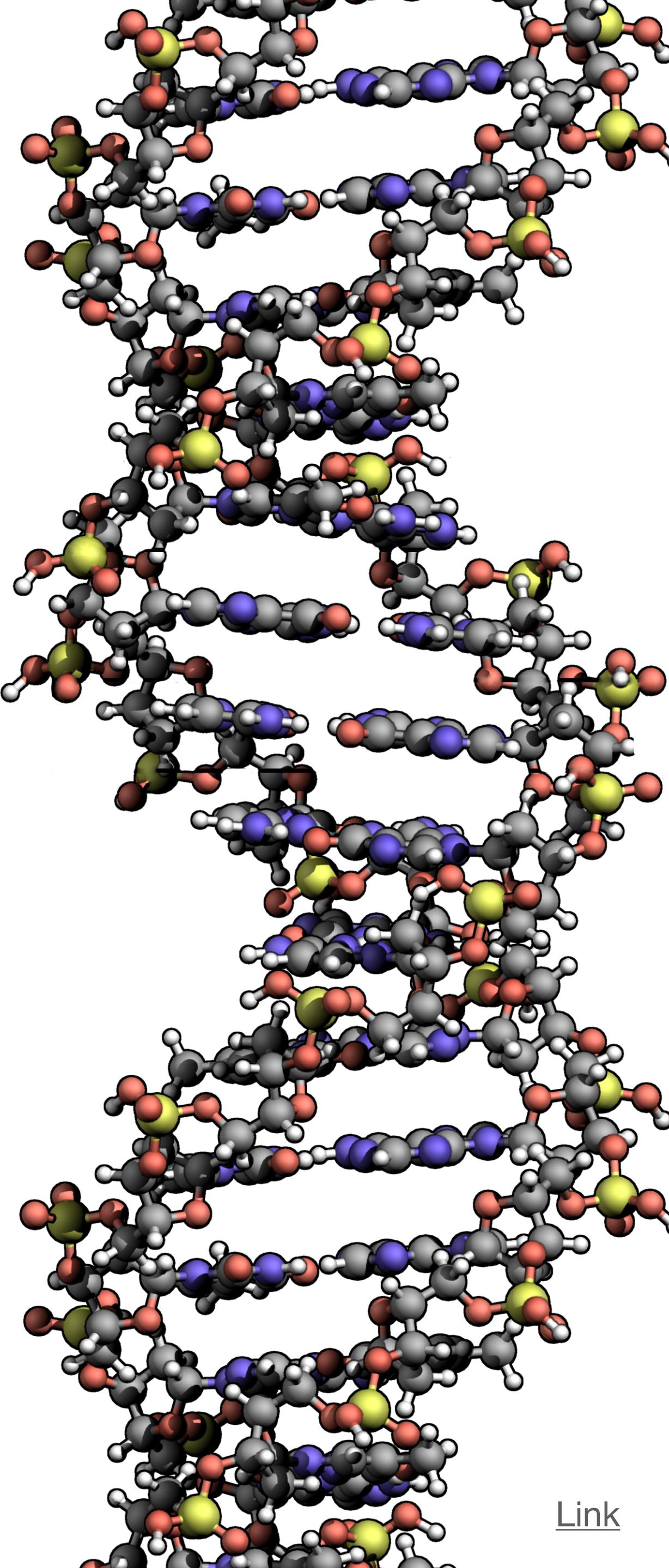
Use Entrez  Compounds  Substances  BioAssays

110M Compounds 277M Substances 293M Bioactivities 843 Data Sources

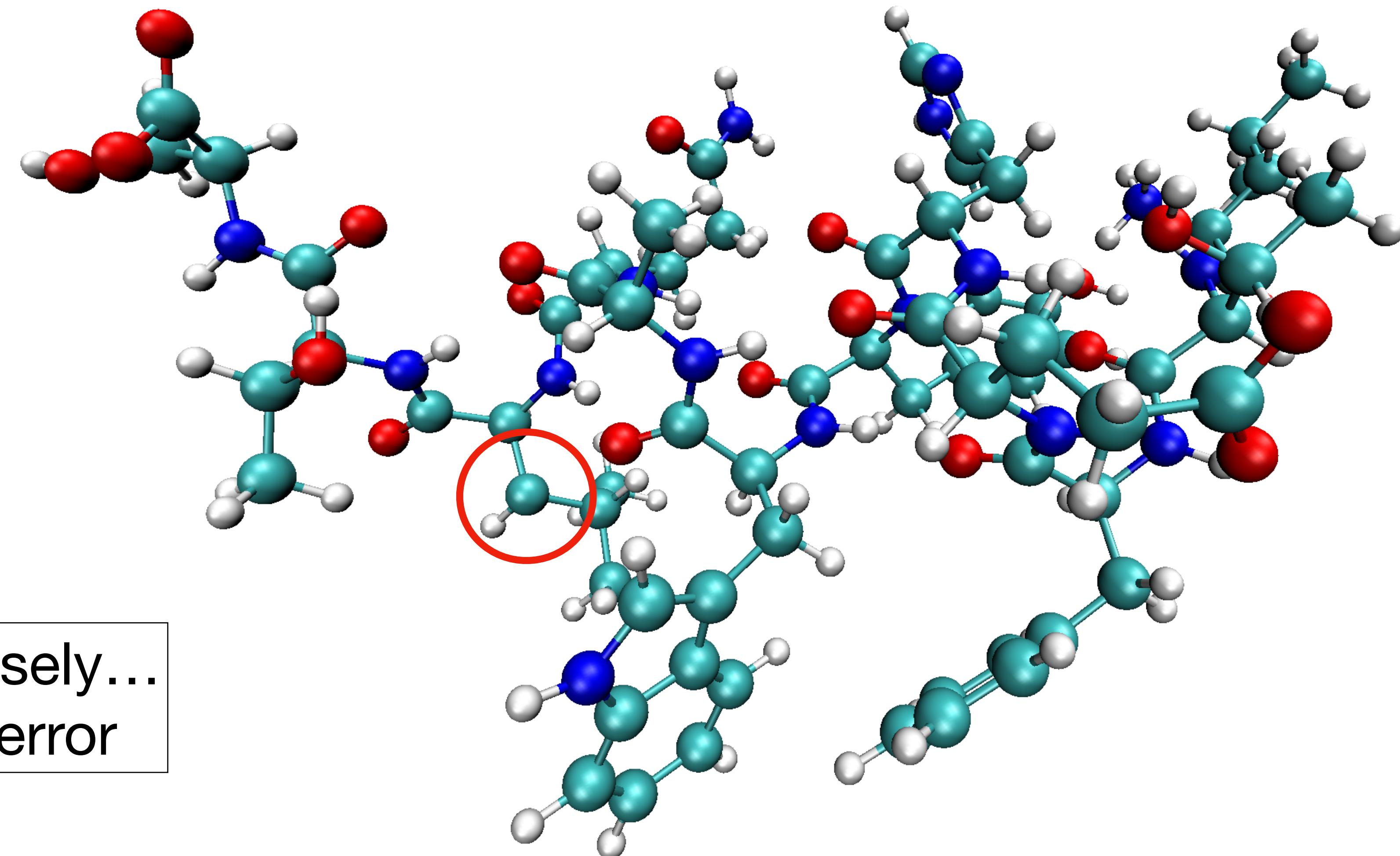
Link

# Part 2

## Structure Refinement



# How to find problems in structures 101

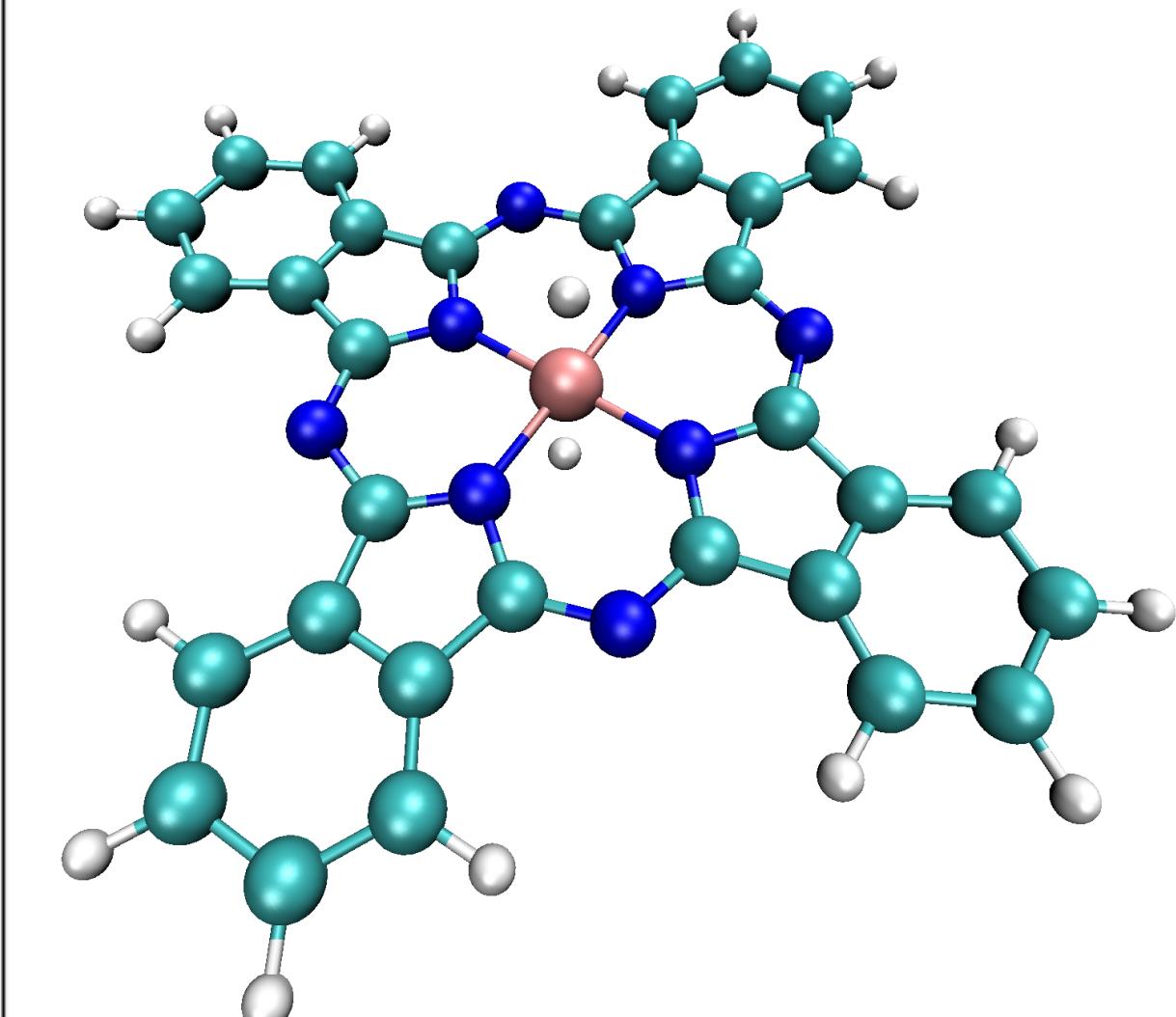
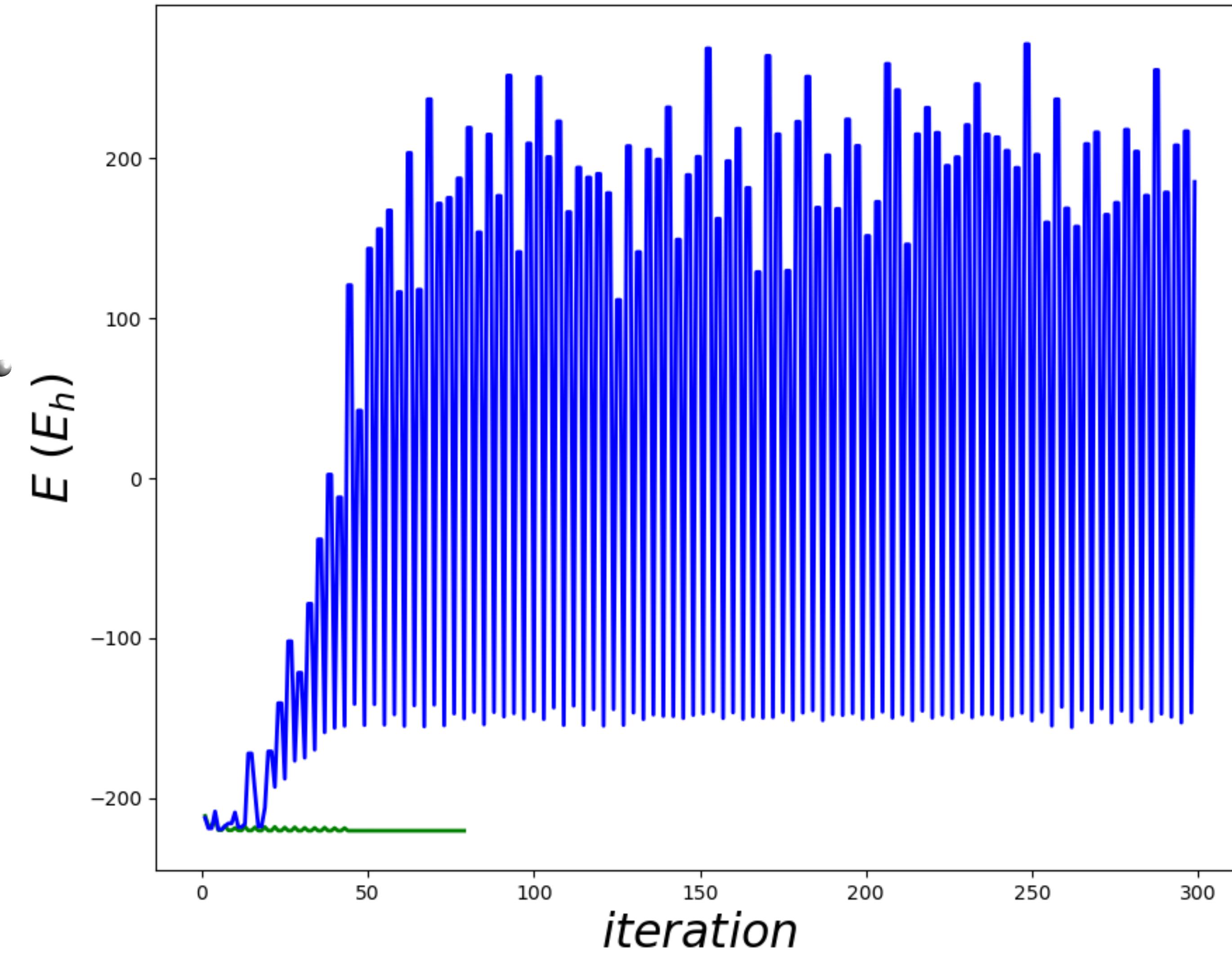
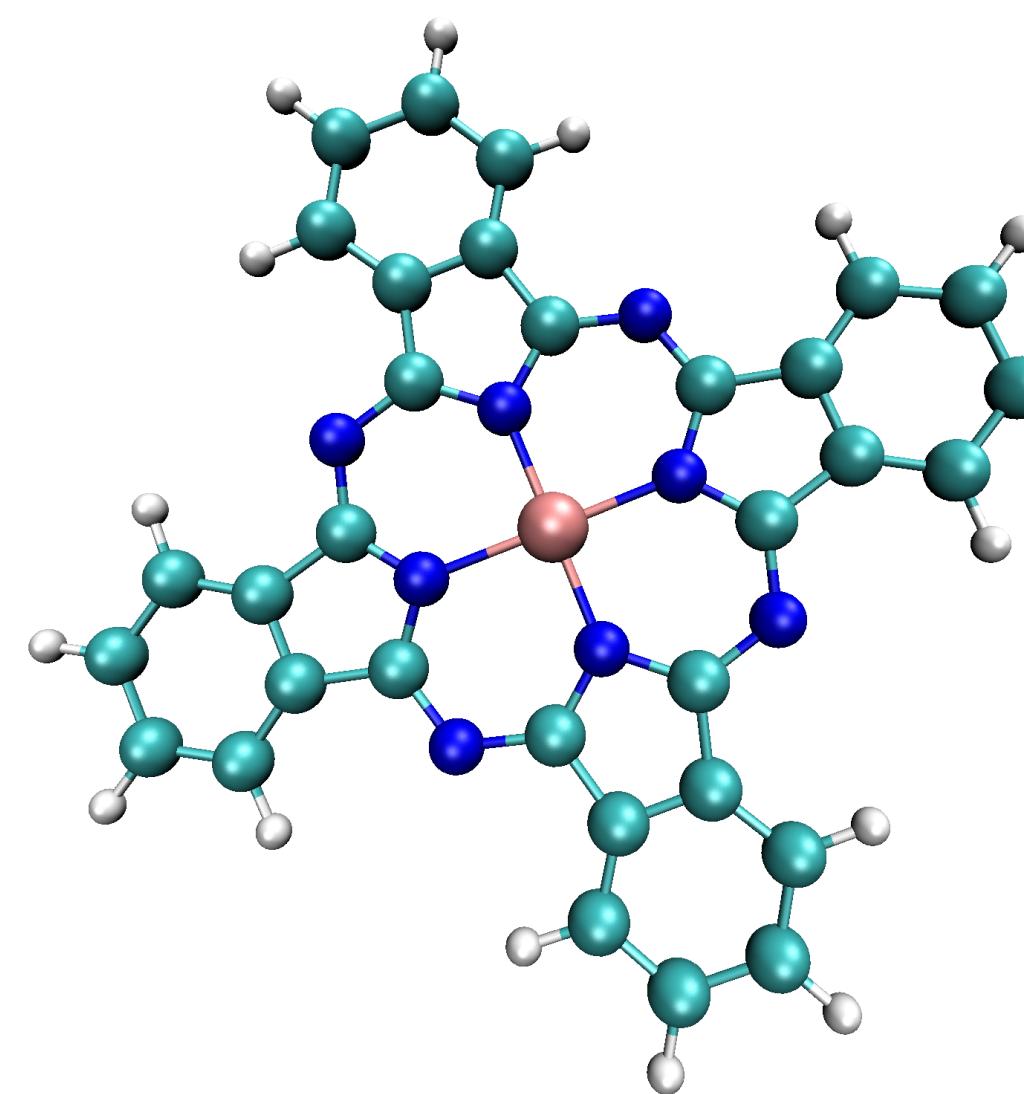


You look very closely...  
and you find the error

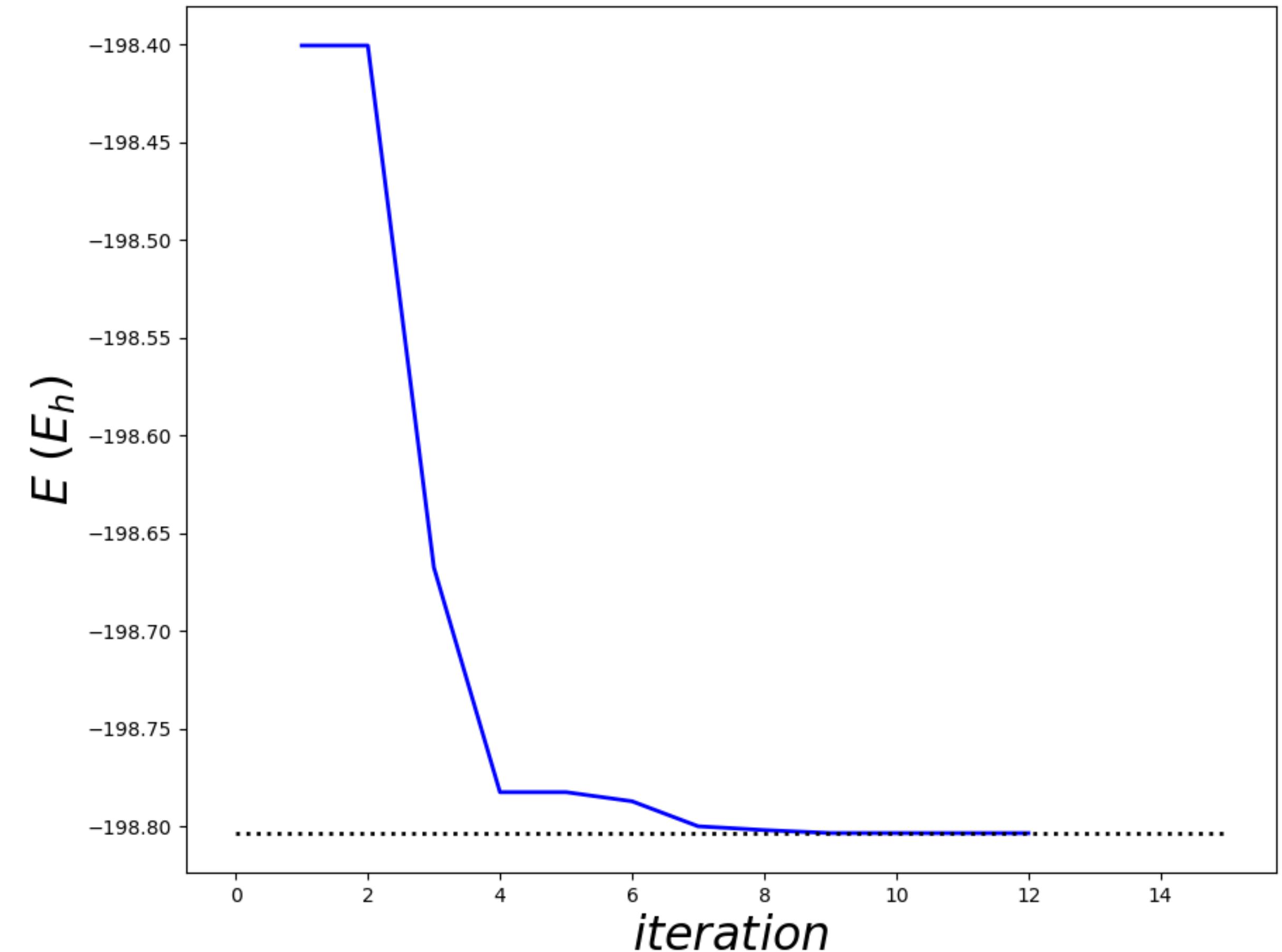
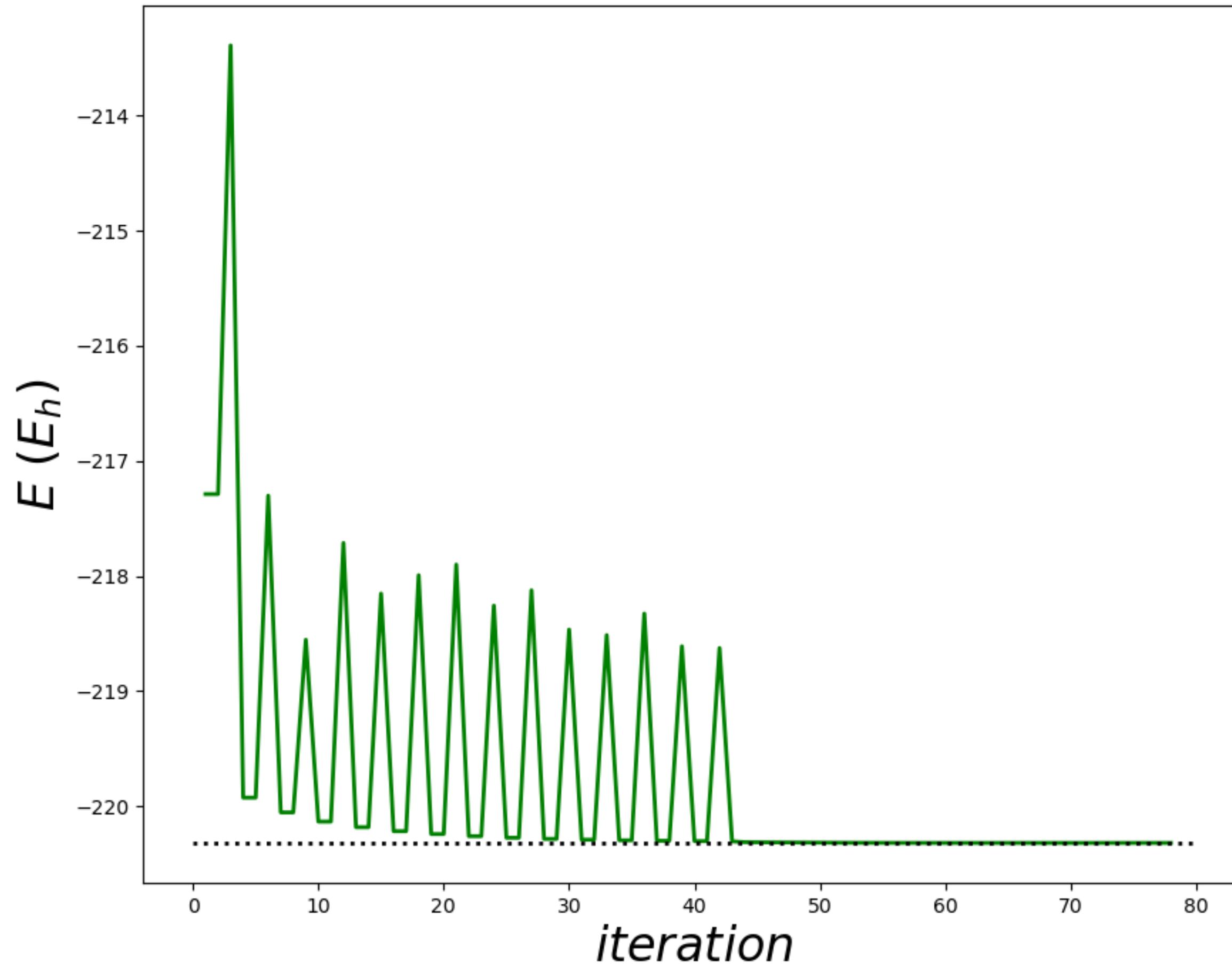
# If You don't Know, Ask Physics

- Physically wrong structures (and electron counting) affects:
  - Force Field Parametrization (fails).
  - Quantum Mechanics convergence (zero to none).
- But also good structures have problems:
  - Dancing Charge problem.
- Population Analysis.
- Geometry Optimization or Dynamics.

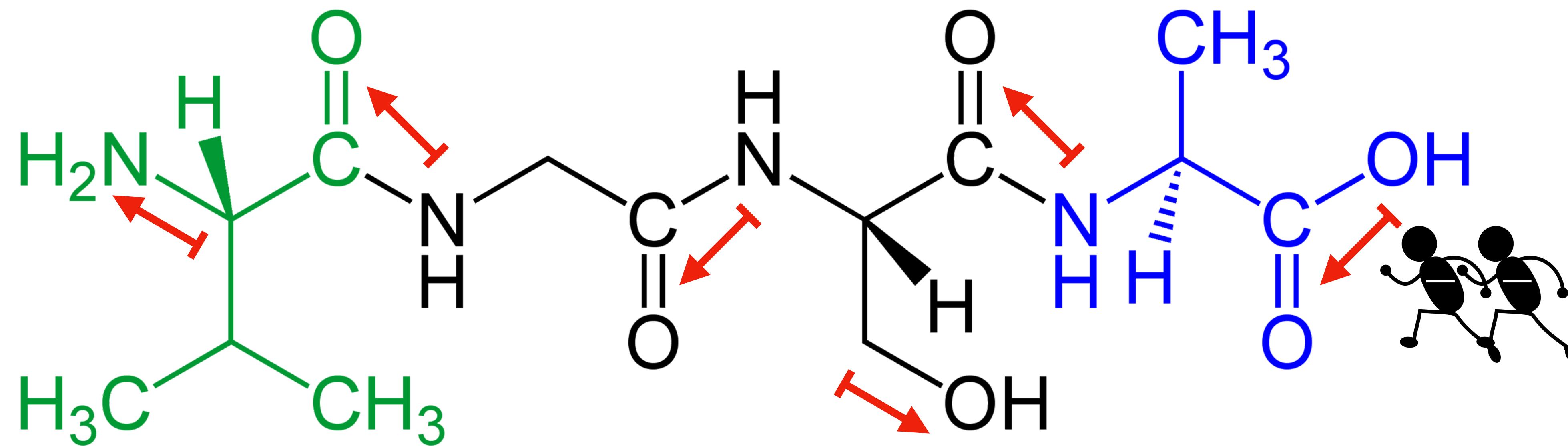
# Example of Convergence of Wavefunction



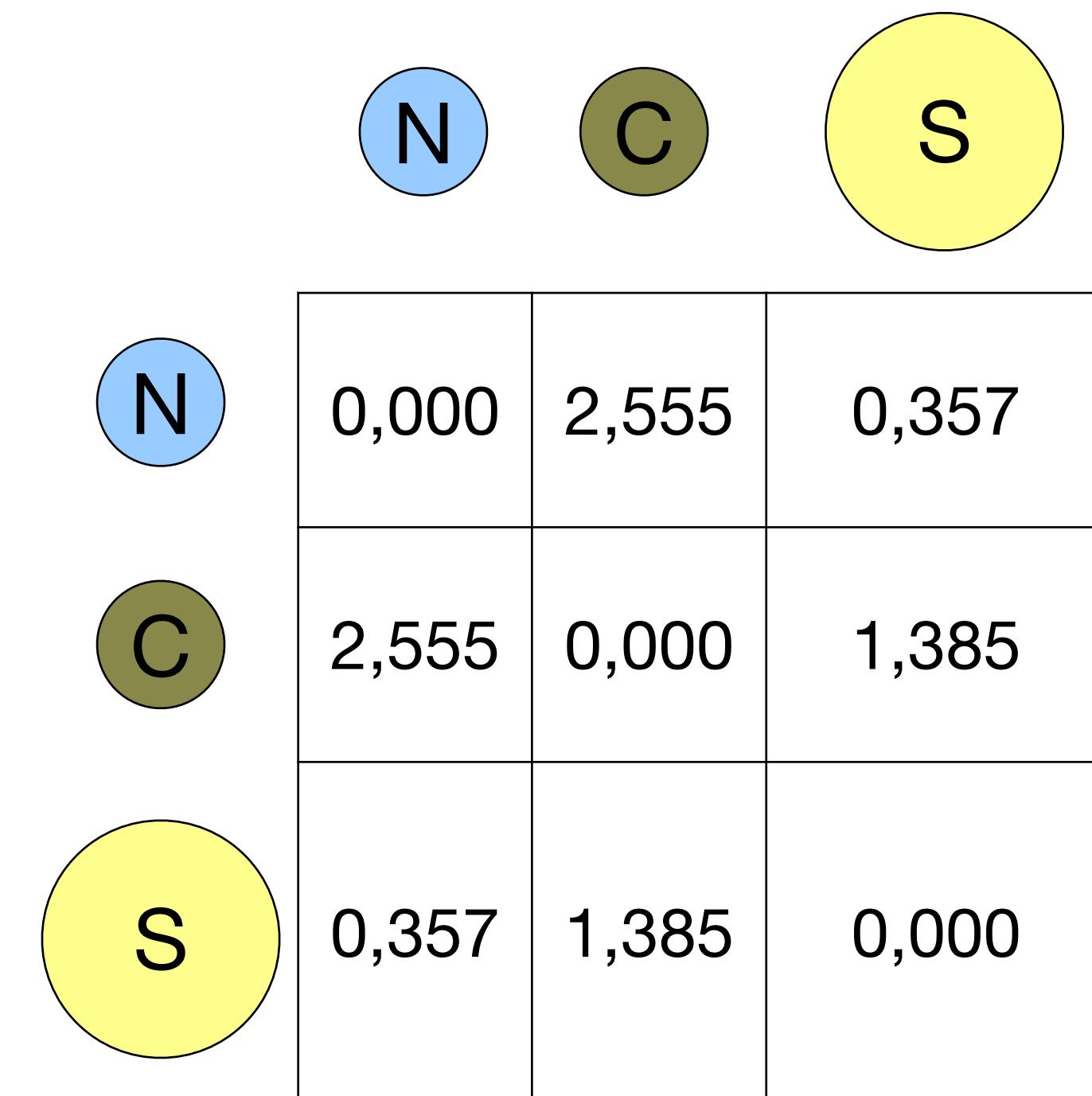
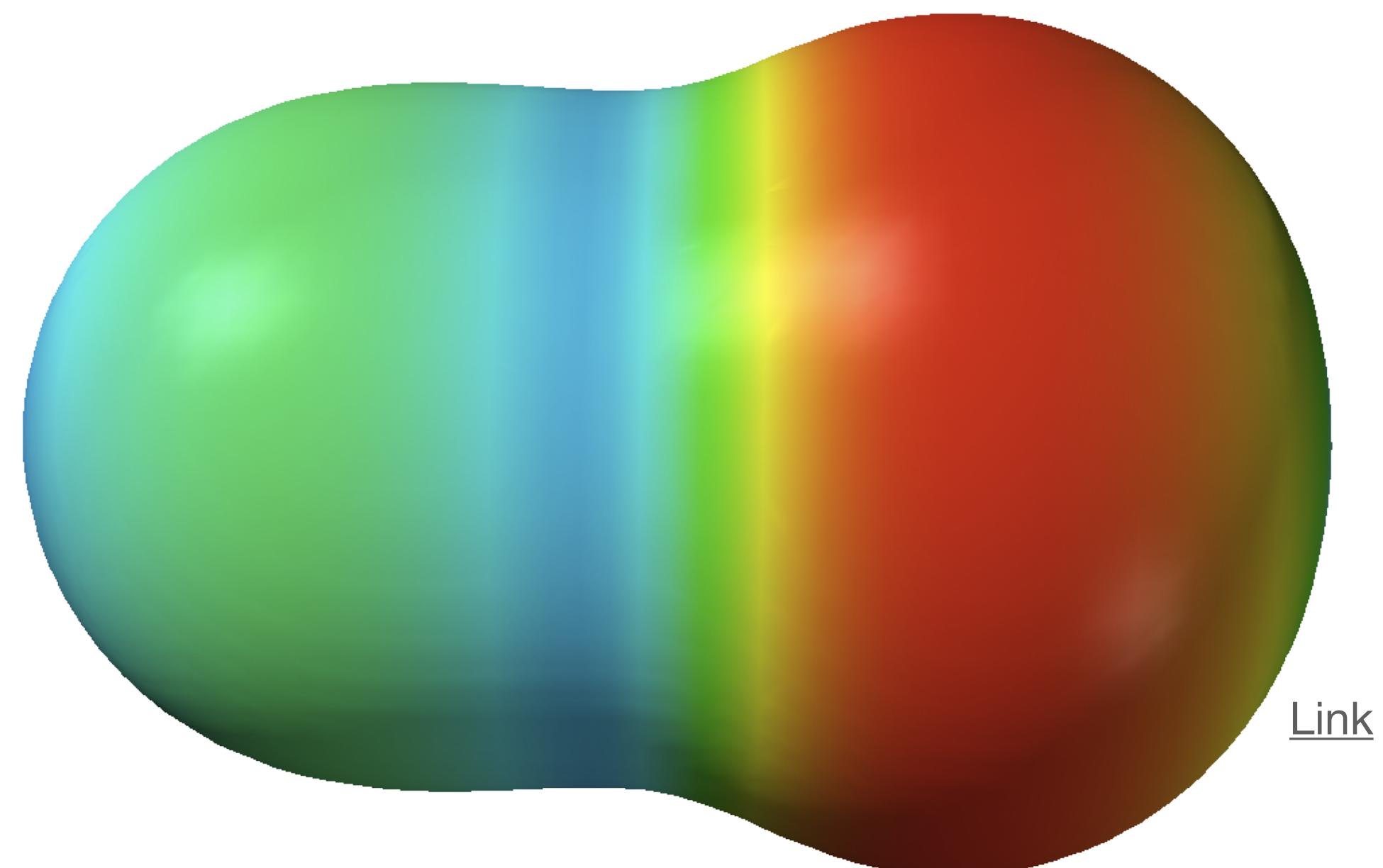
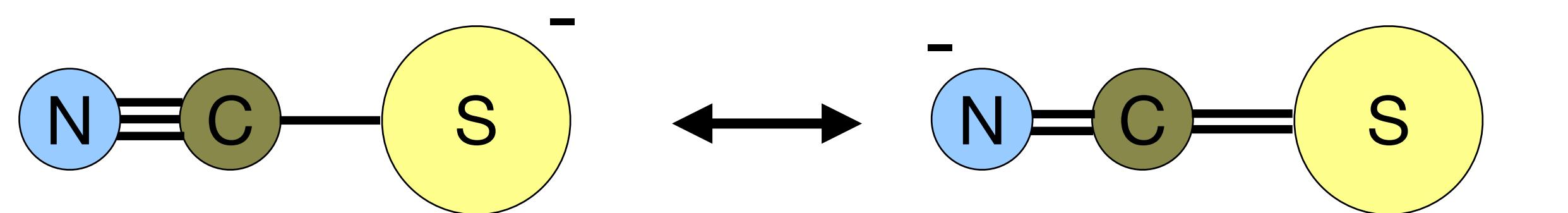
# Example of Convergence of Wavefunction



# Peptides and Dancing Charges



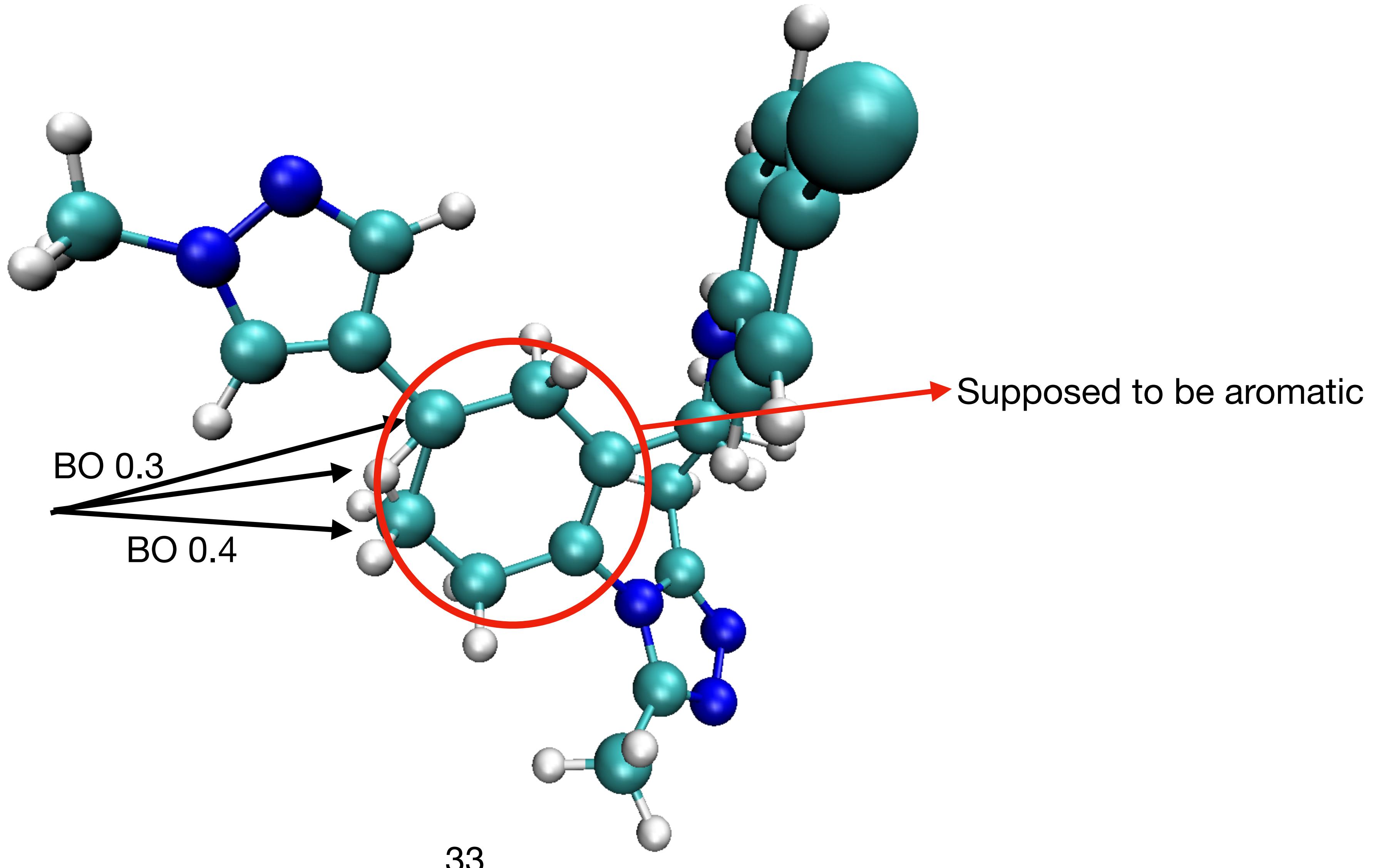
# Population Analysis and Bond Orders



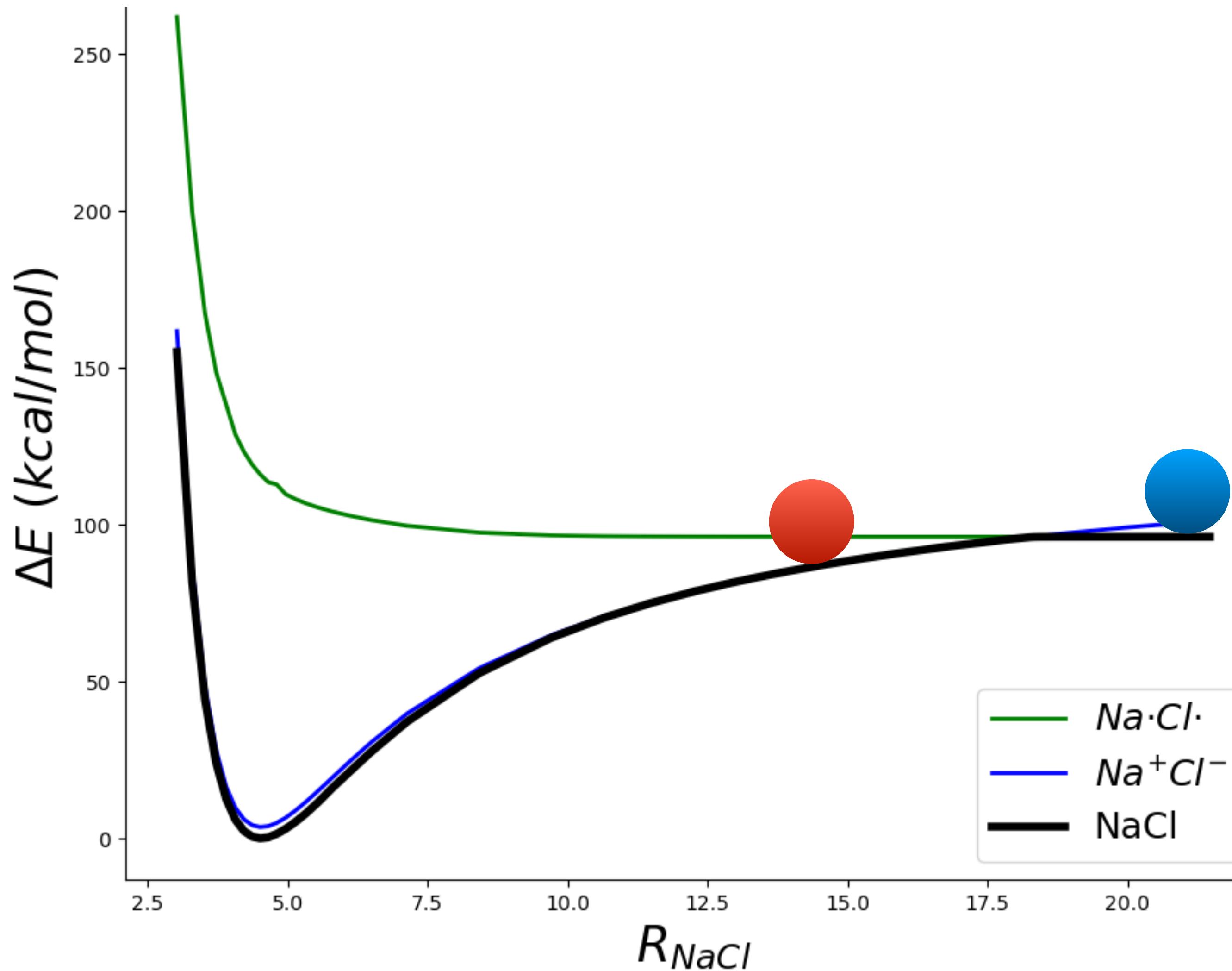
Atom population analysis table showing the atomic charges for the nitro radical ( $\text{NO}_2^-$ ) and nitro anion ( $\text{NO}_2$ ).

	N	C	S
N	0,000	2,555	0,357
C	2,555	0,000	1,385
S	0,357	1,385	0,000

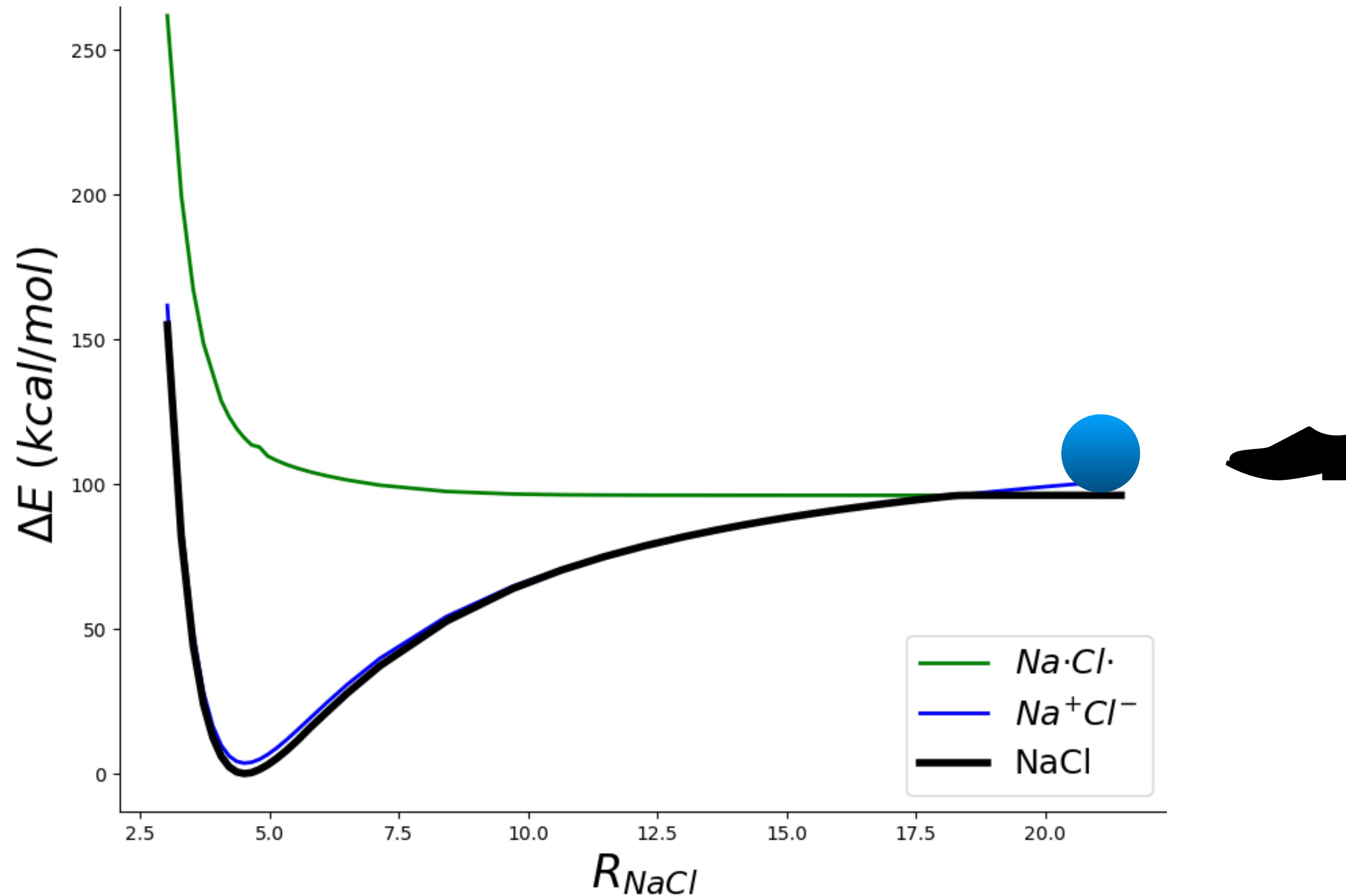
# Bond Order Matrices



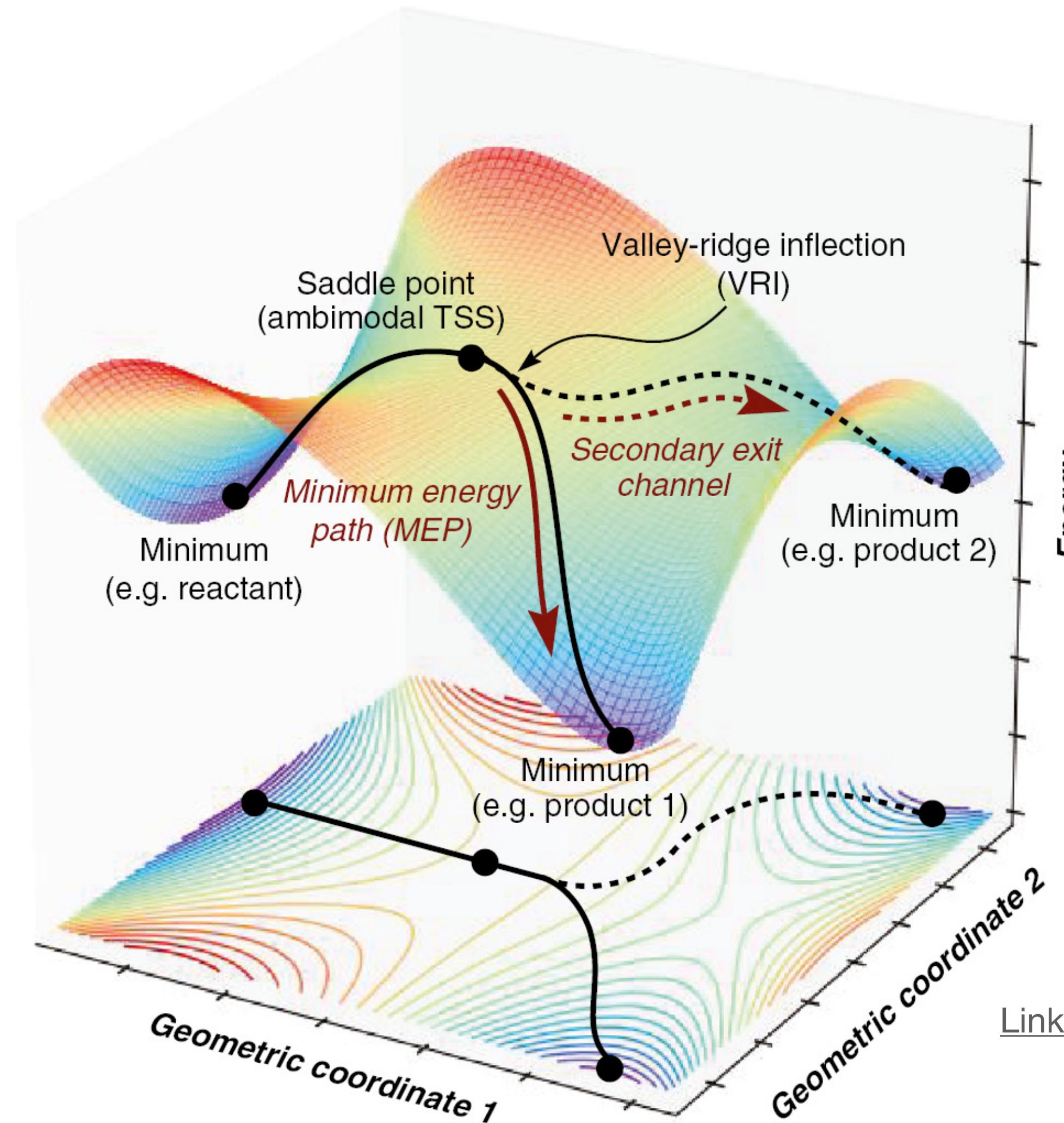
# Optg and MD



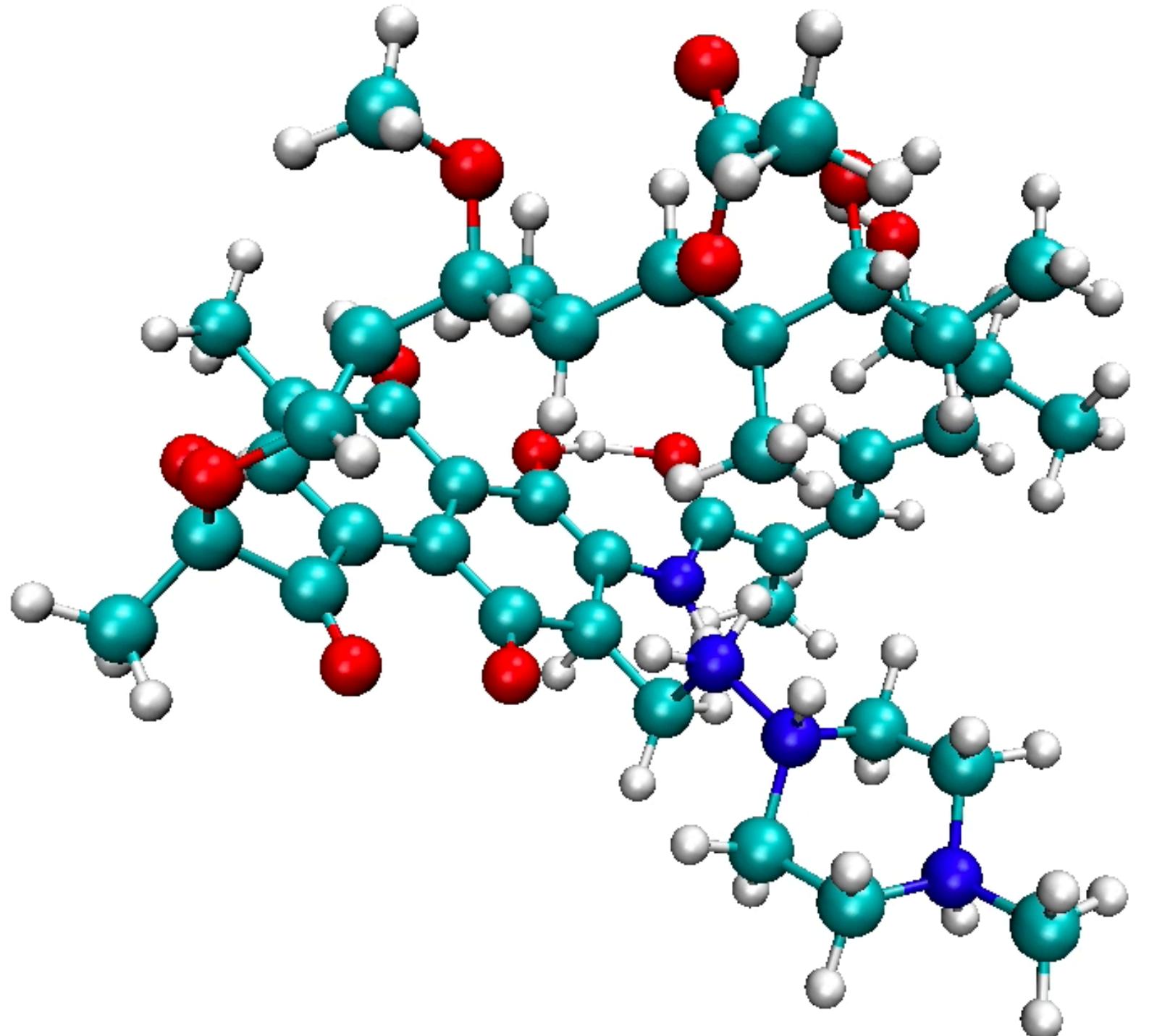
# Optg and MD



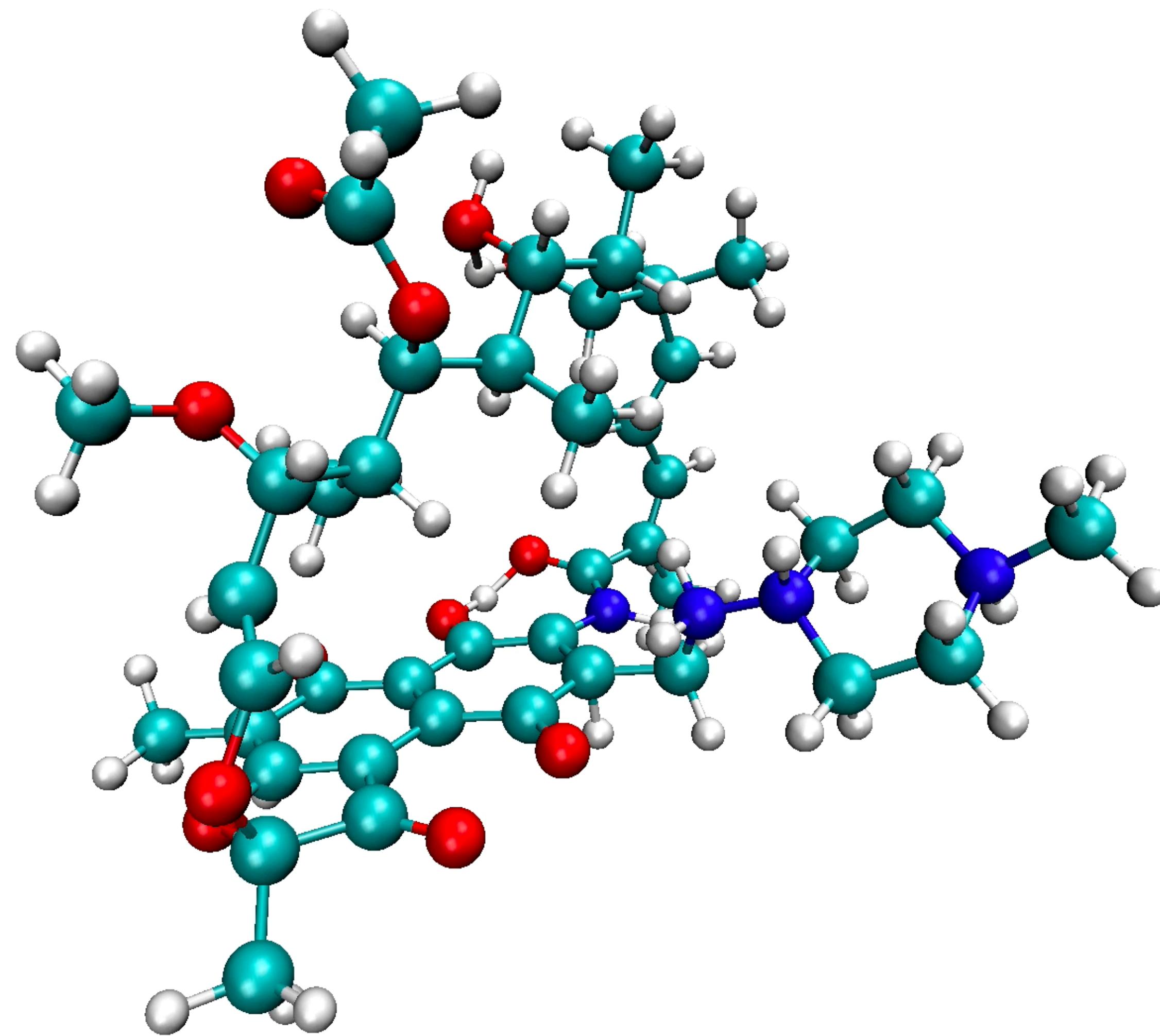
# Potential Energy Surface



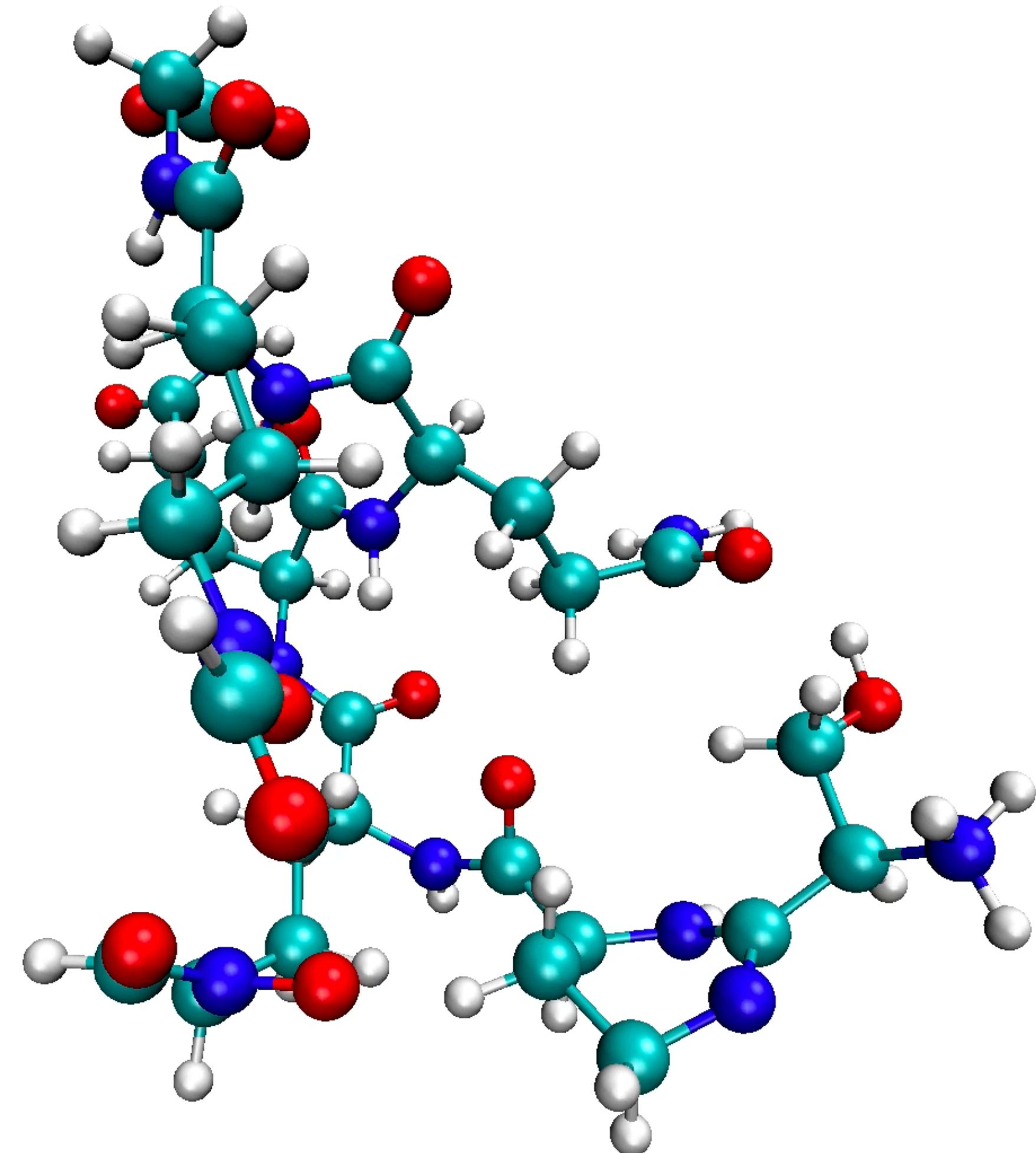
# MD of Incorrect Structure



# Optg of (same) Incorrect Structure

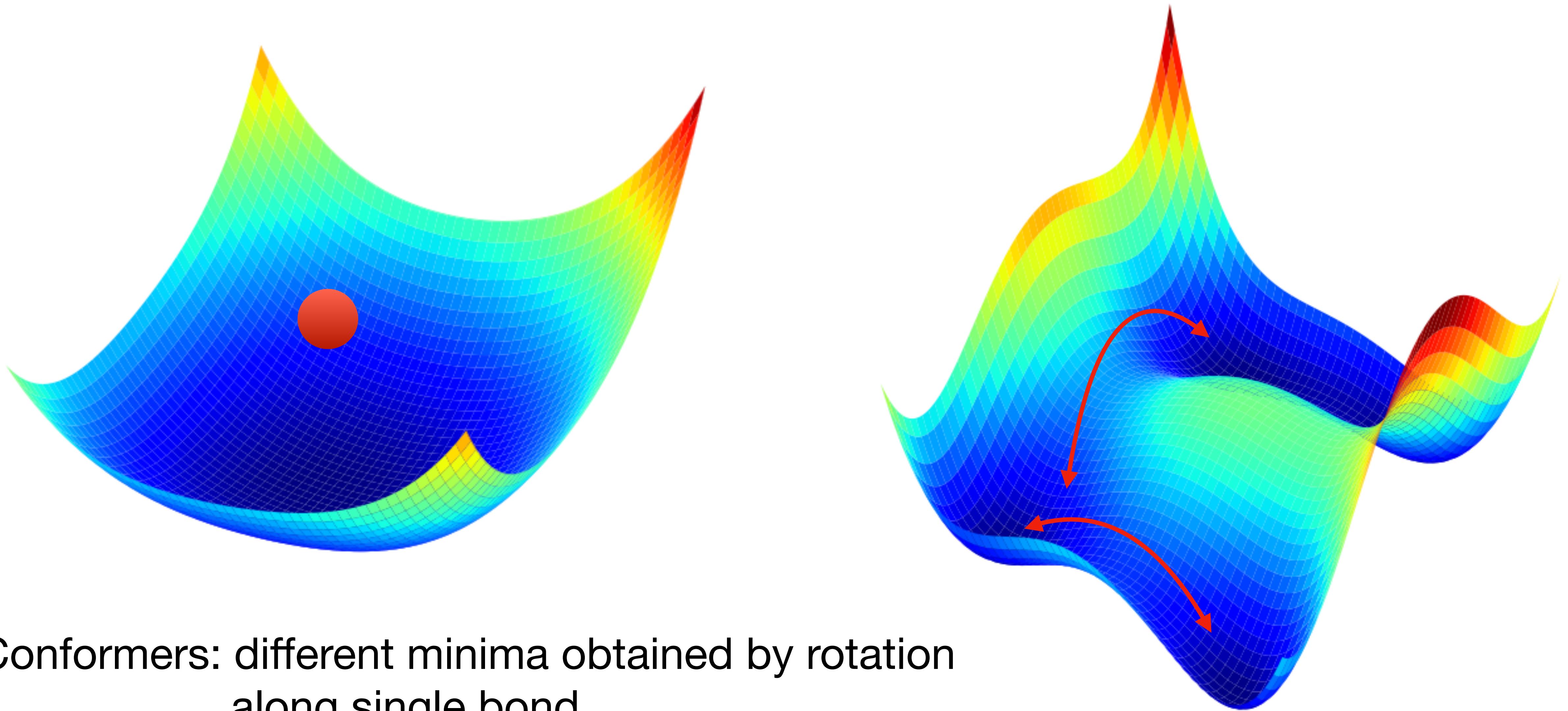


# Radicals Force Their Way



- Two electrons were missing (radical).
- Those atoms panicked: incomplete shells
- They did all they could to refill shells

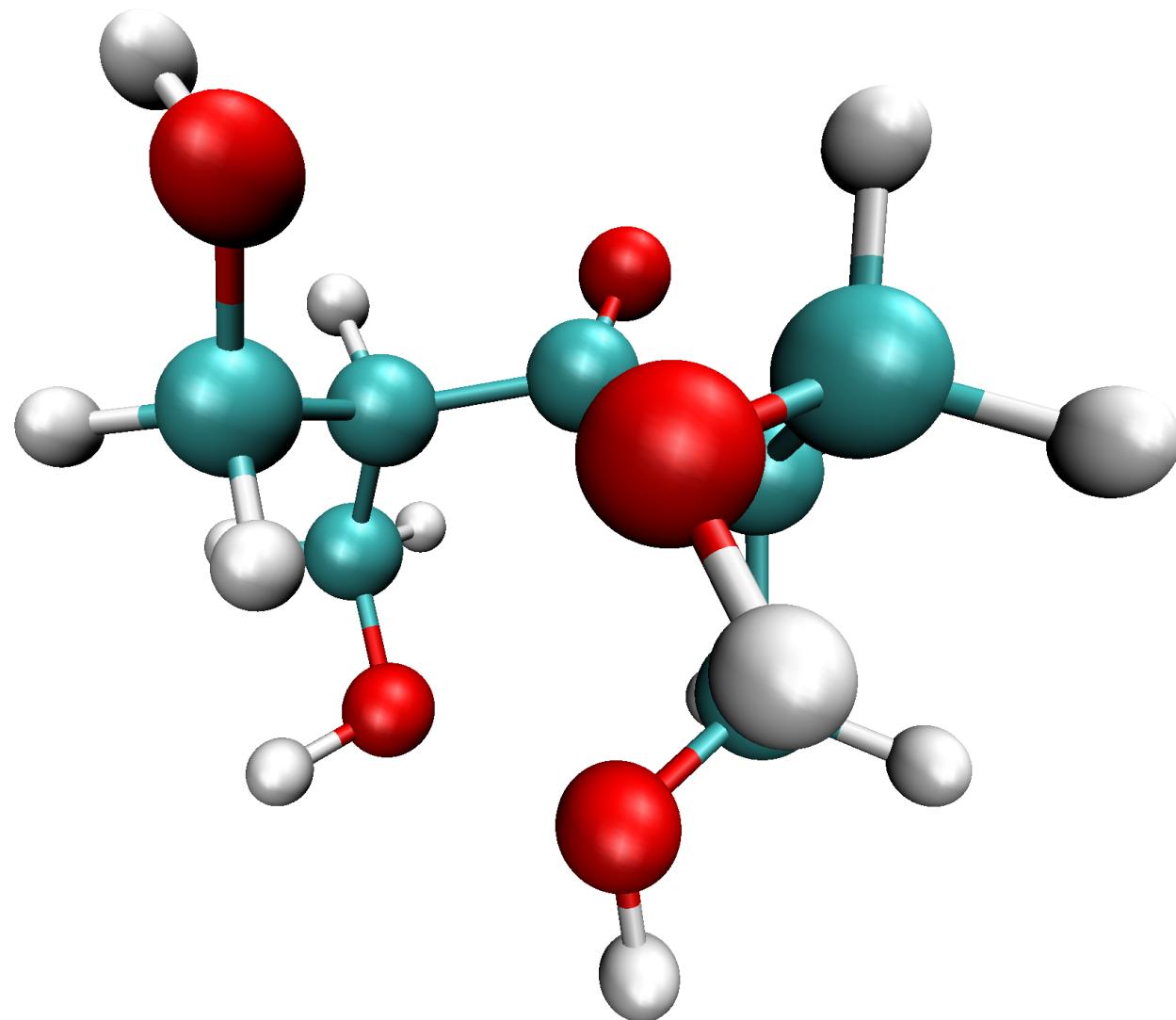
# Molecules as Minima in Surfaces



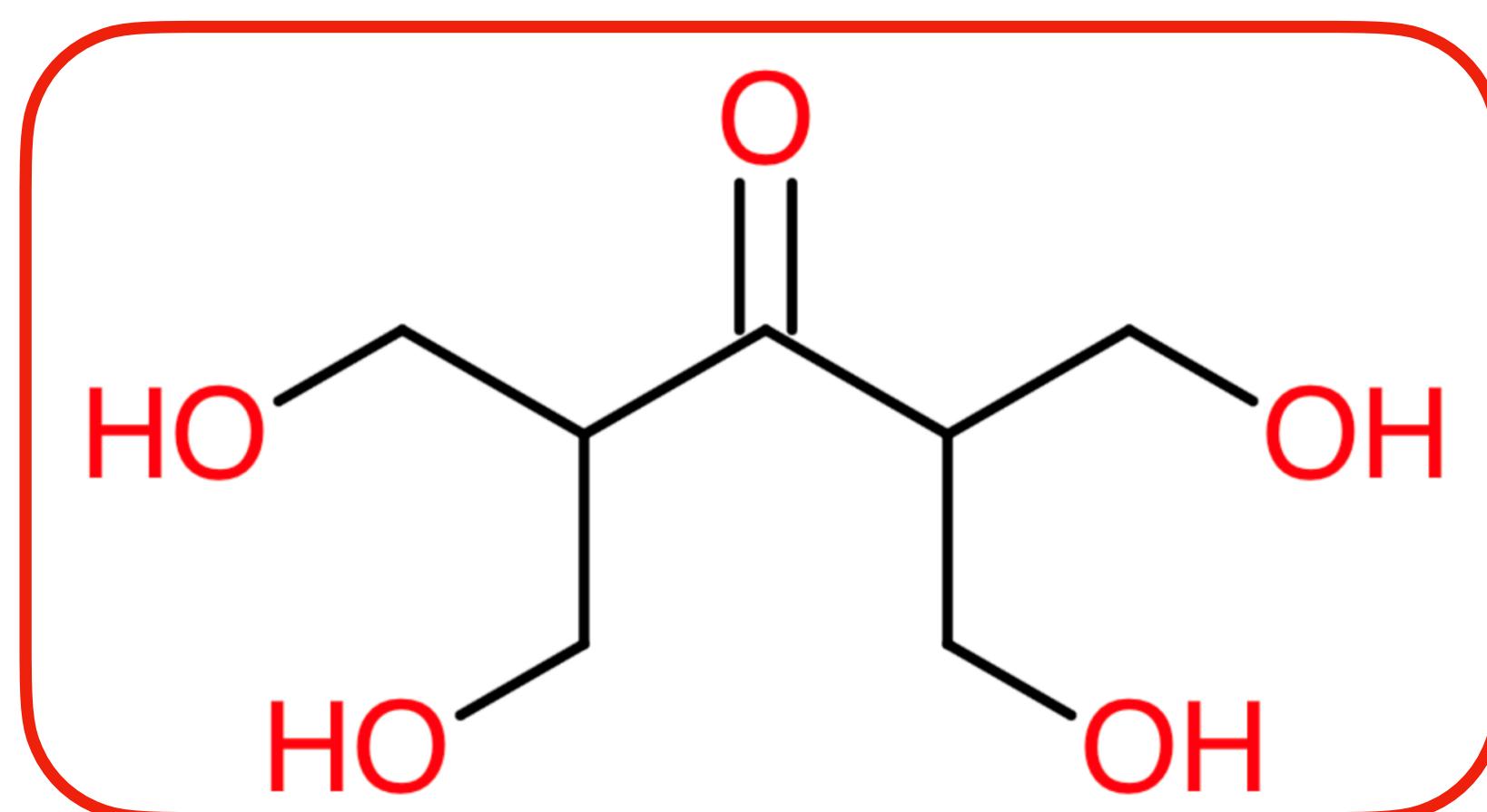
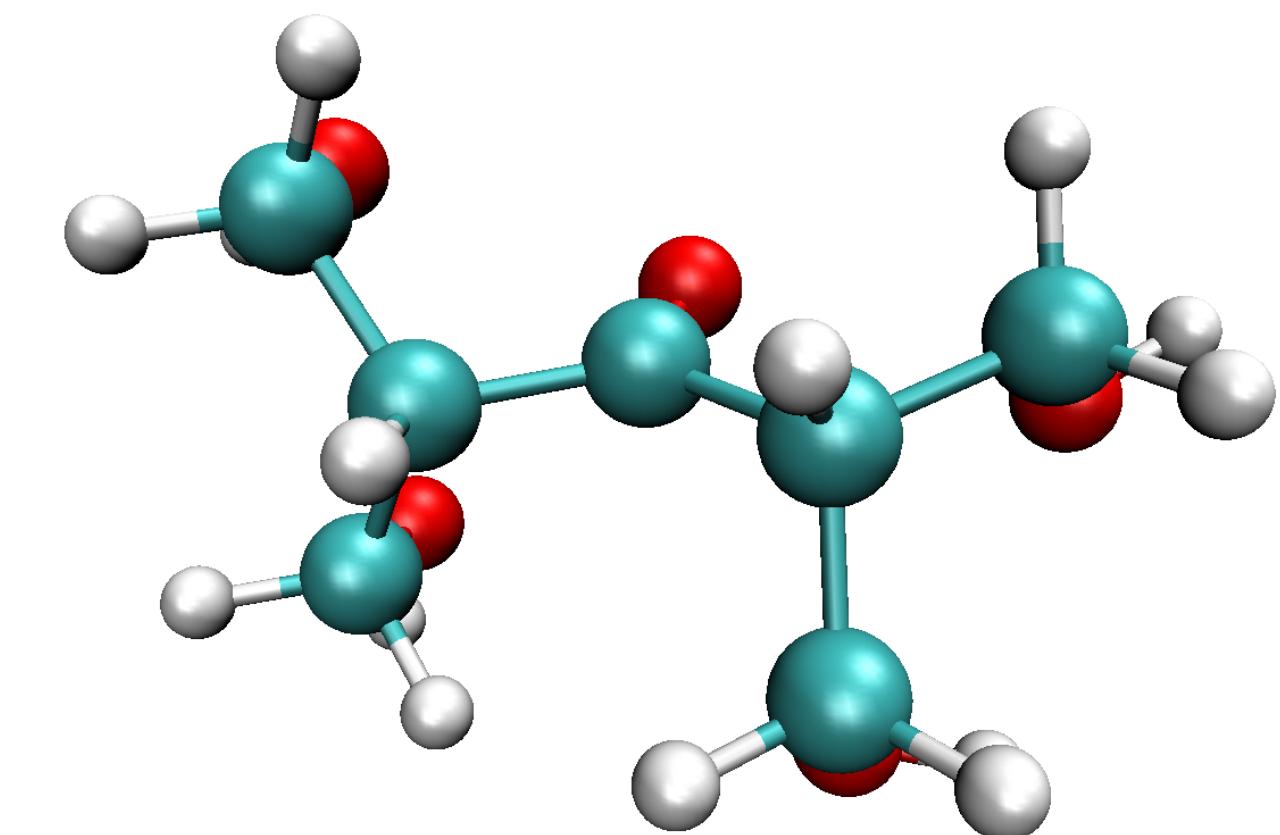
Conformers: different minima obtained by rotation  
along single bond

# All Conformers are equal... are they?

Conformer A

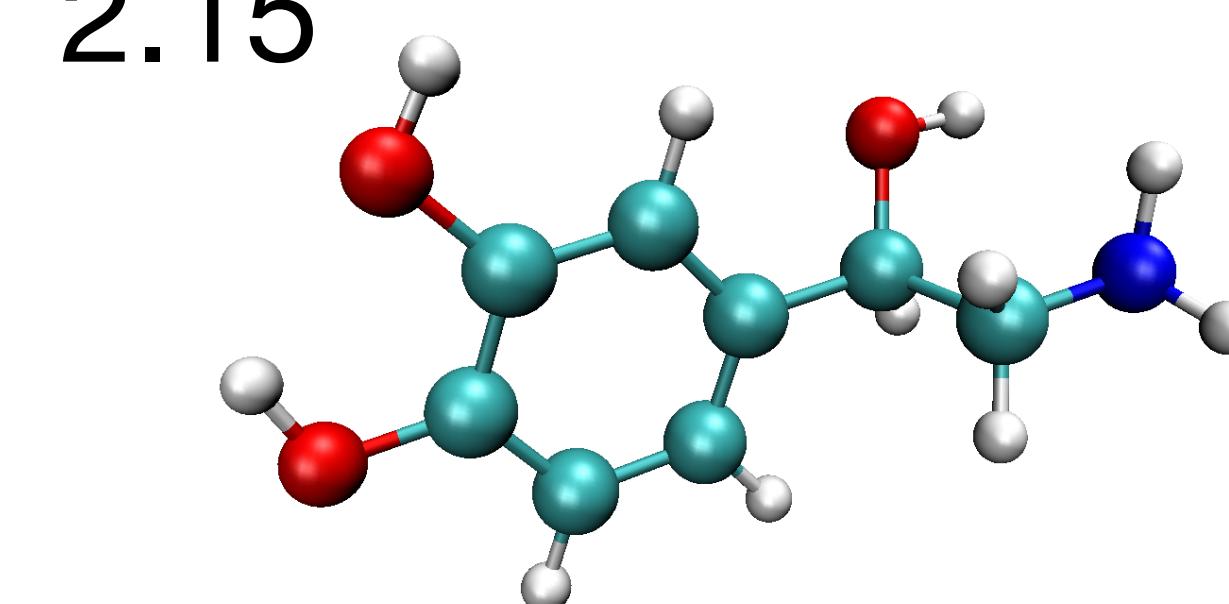
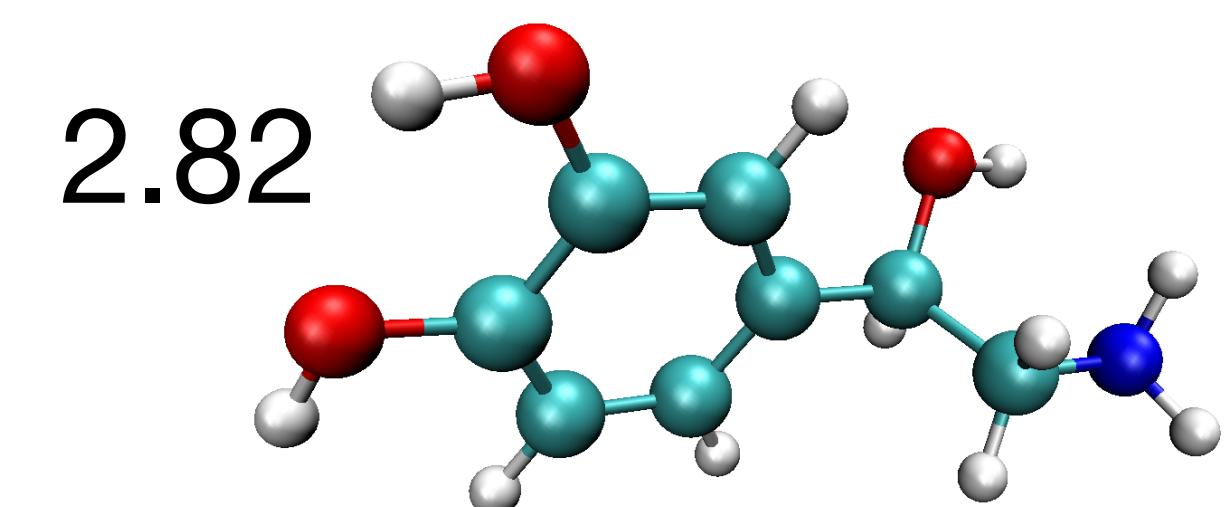
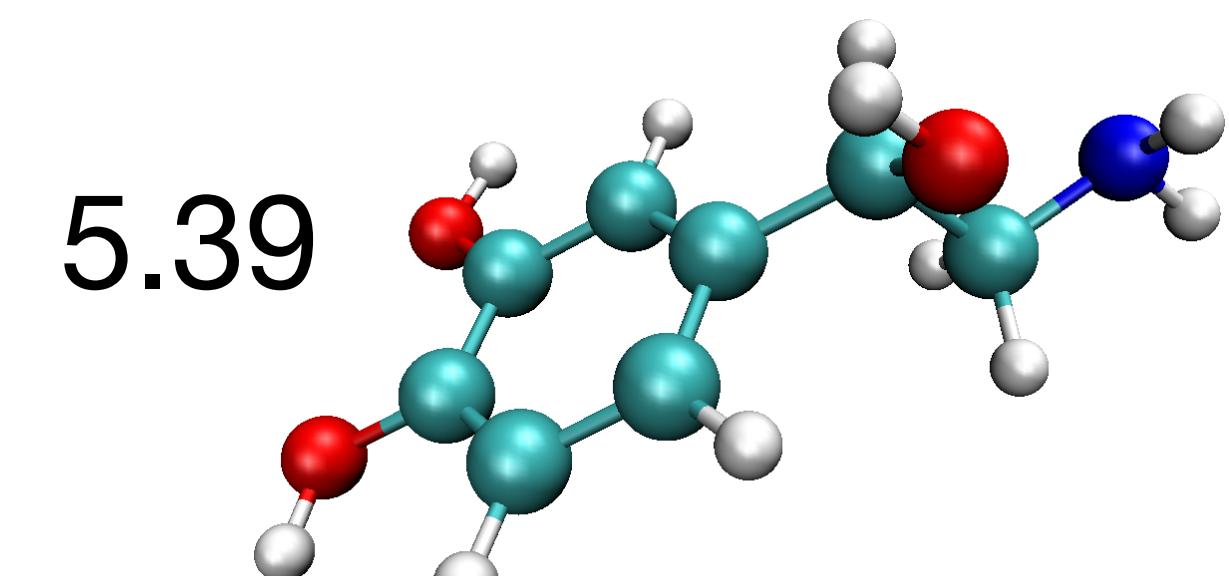
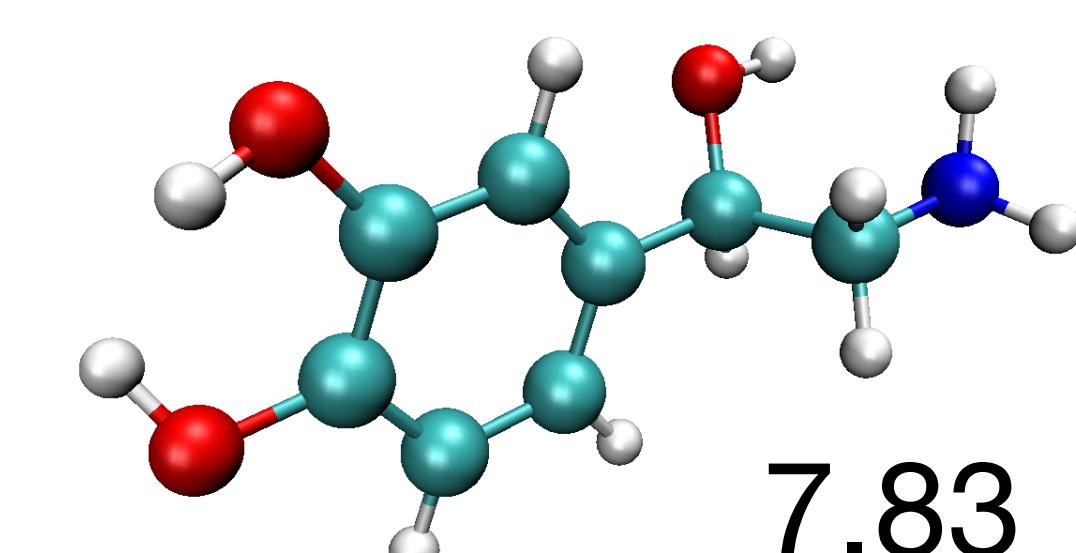
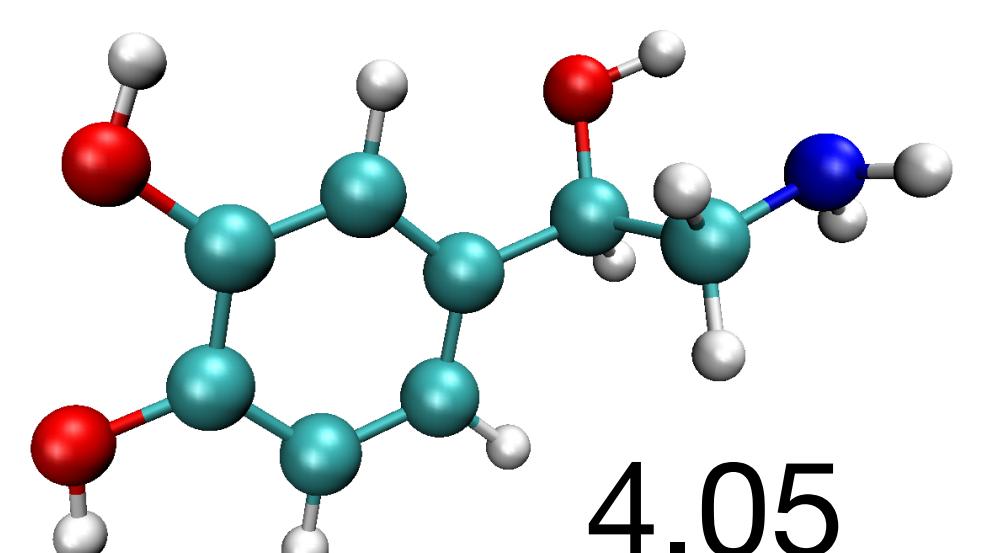
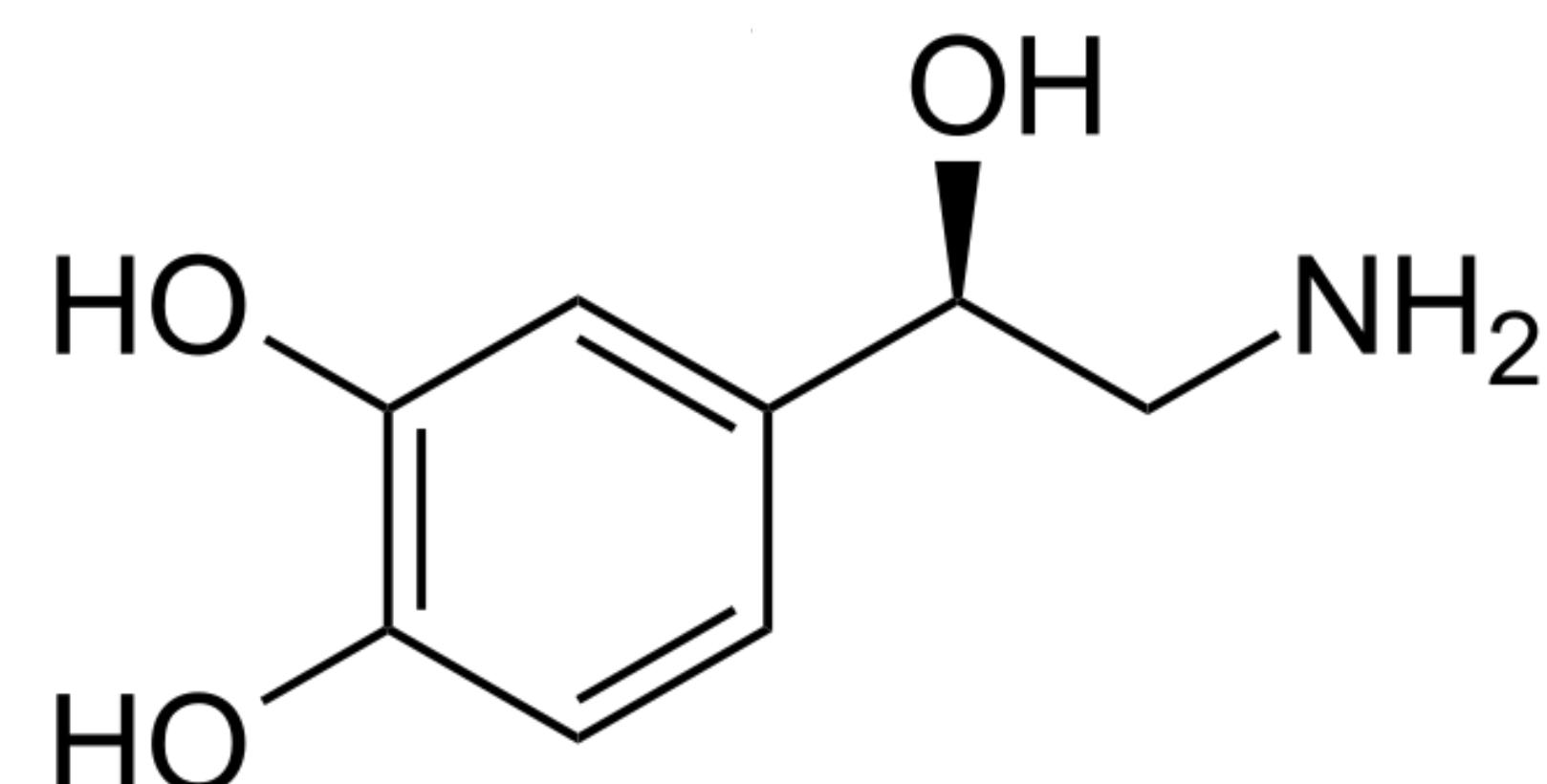
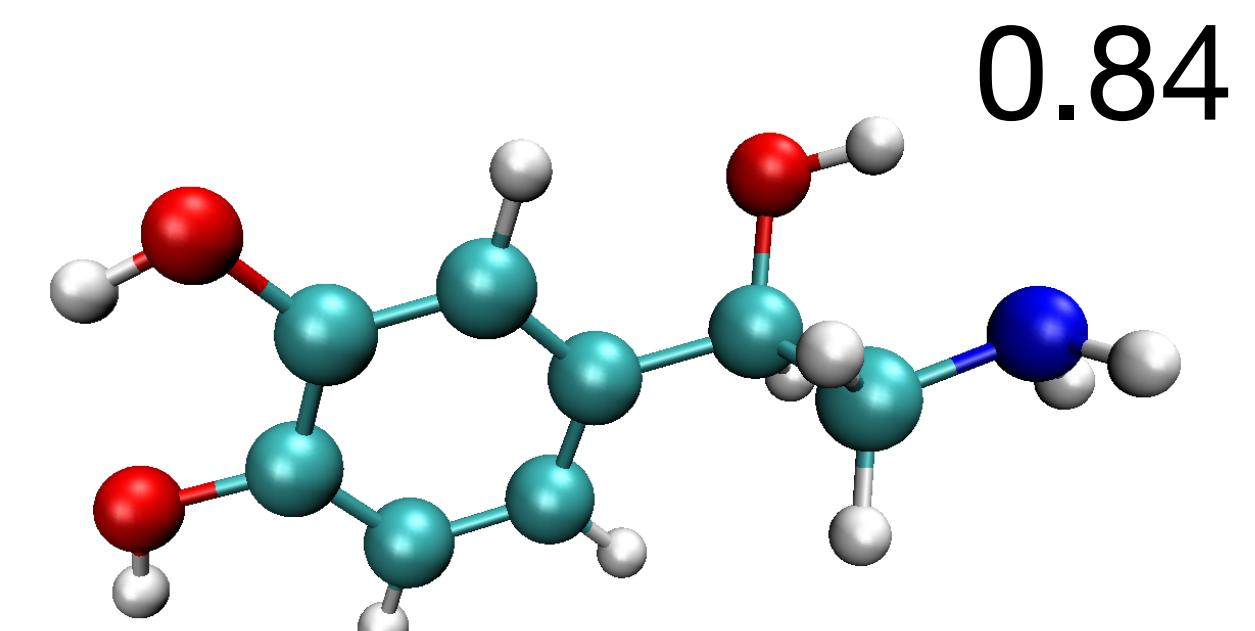
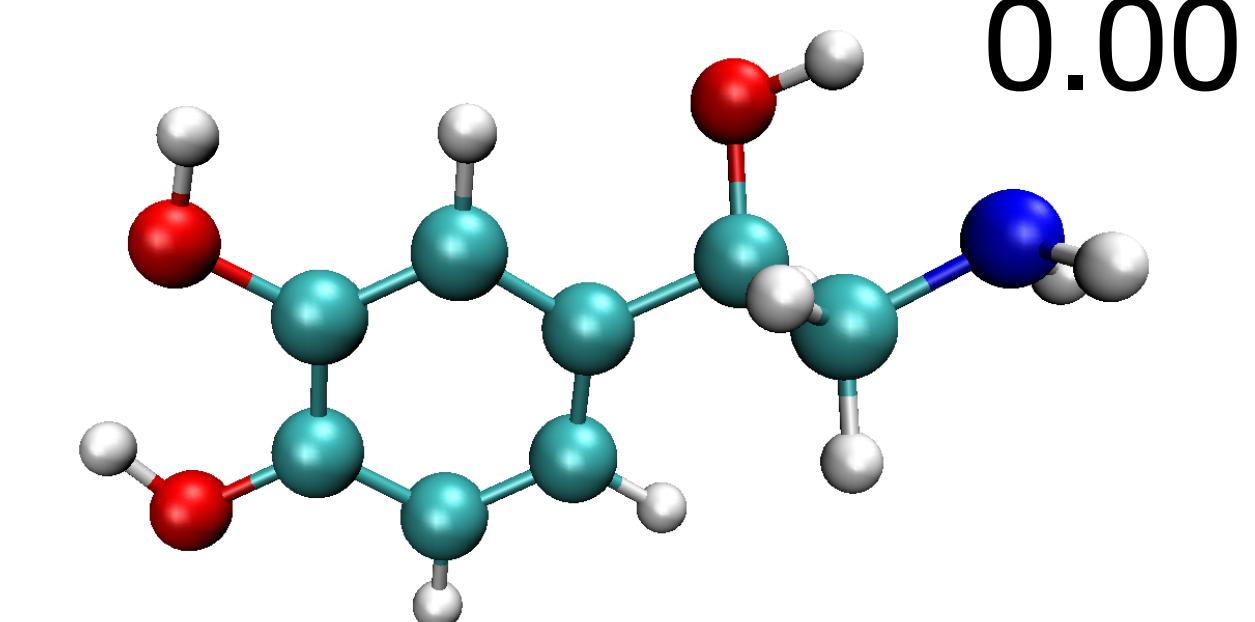
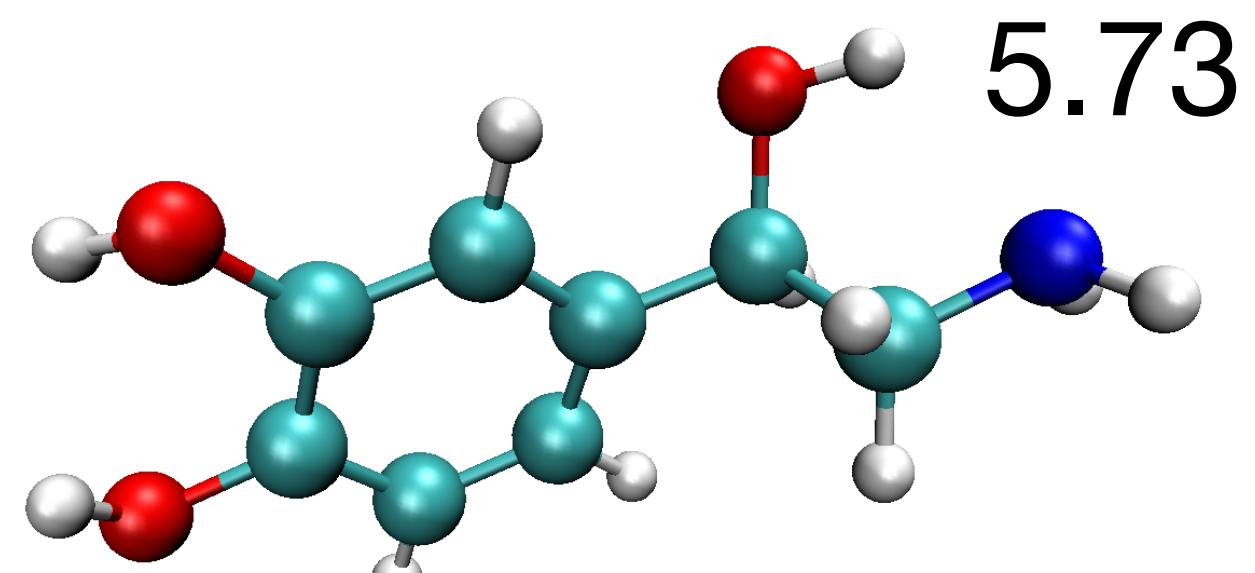


Conformer B



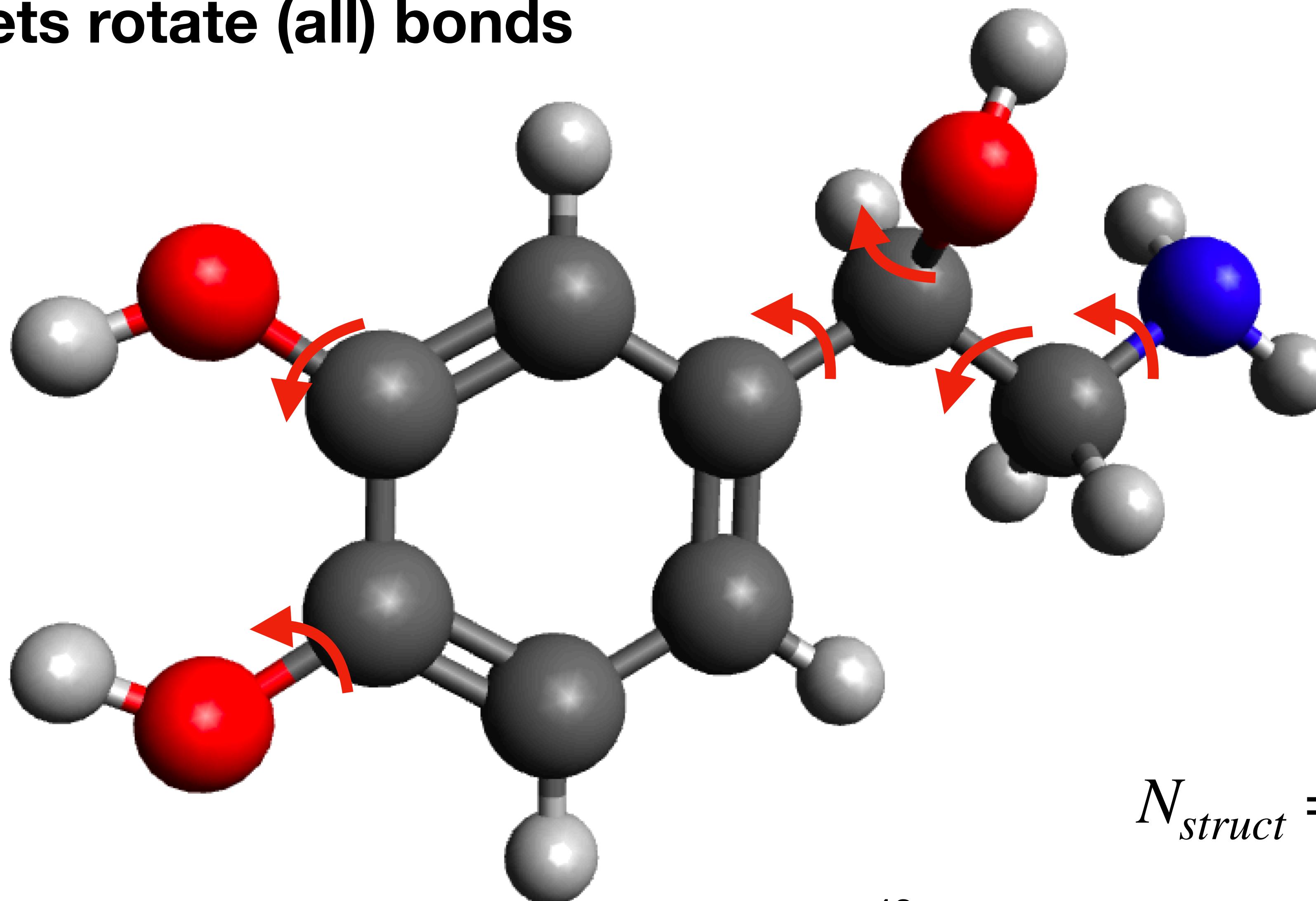
Solvent 1

# All Conformers are equal...



# Simple Algorithms for Finding Conformers

Lets rotate (all) bonds

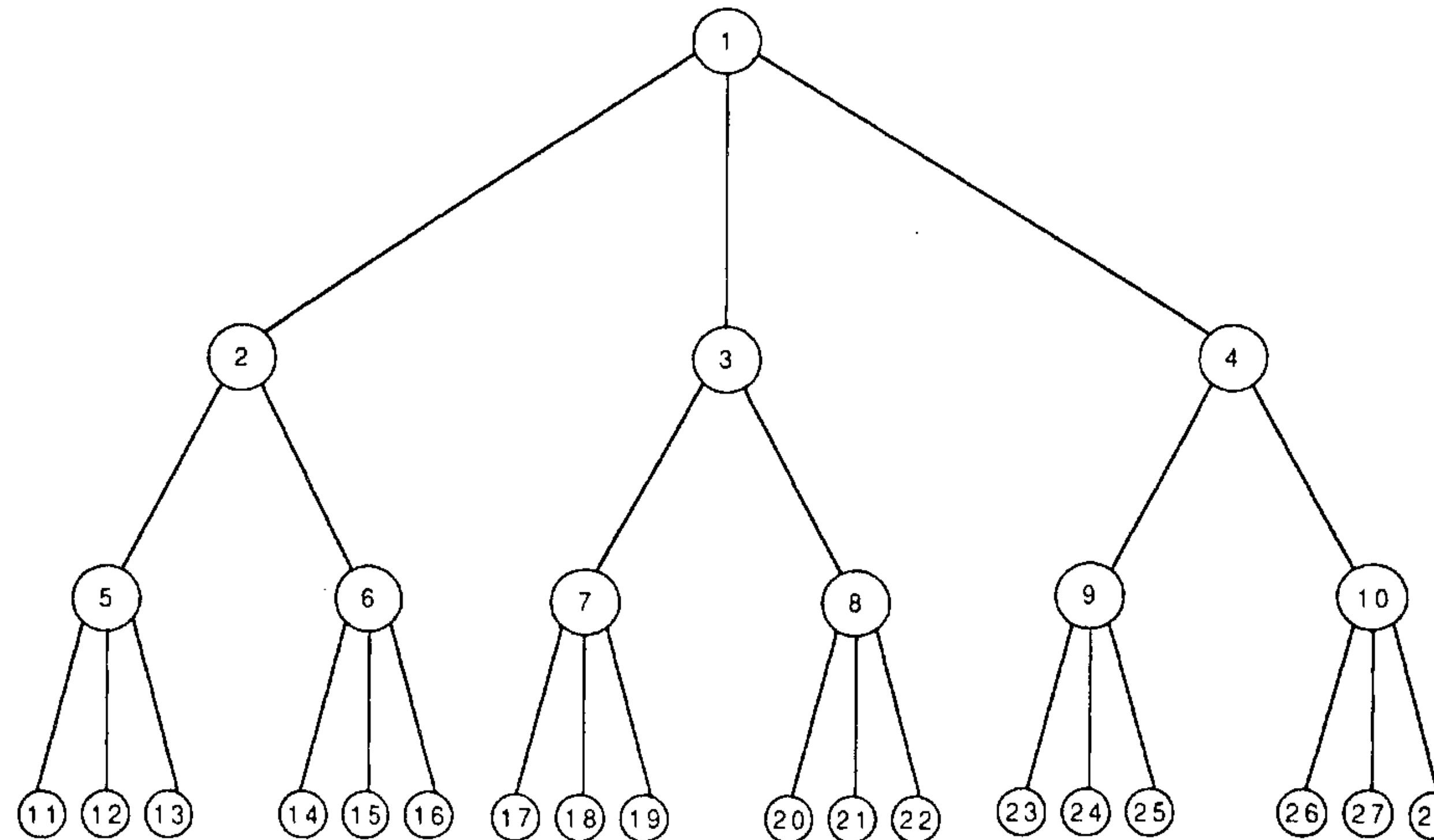


$$N_{struct} = \left( \frac{360}{\theta} \right)^N$$

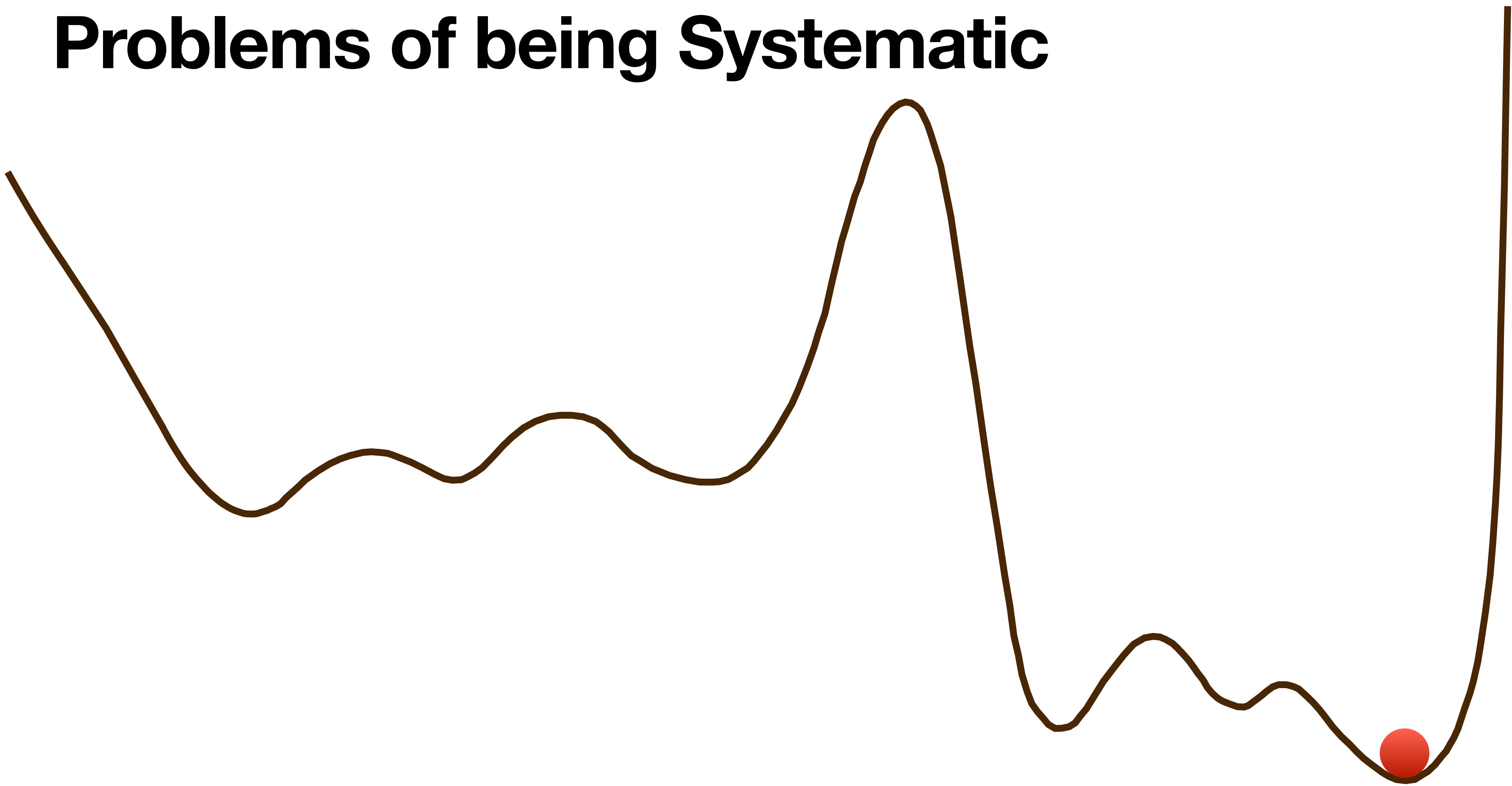
$$N_{struct} = \left( \frac{360}{45} \right)^6 = 262144$$

# Simple Algorithms for Finding Conformers

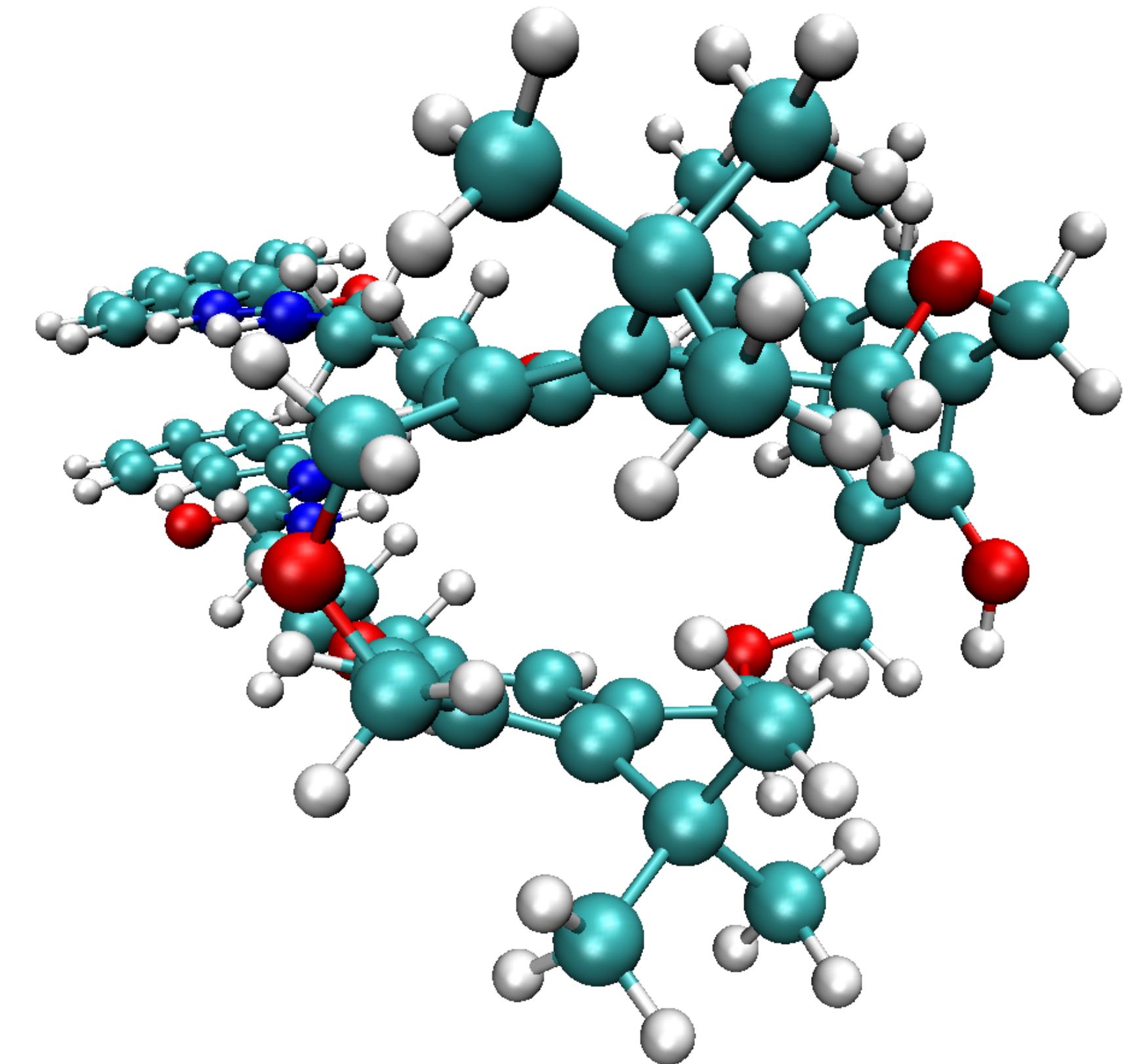
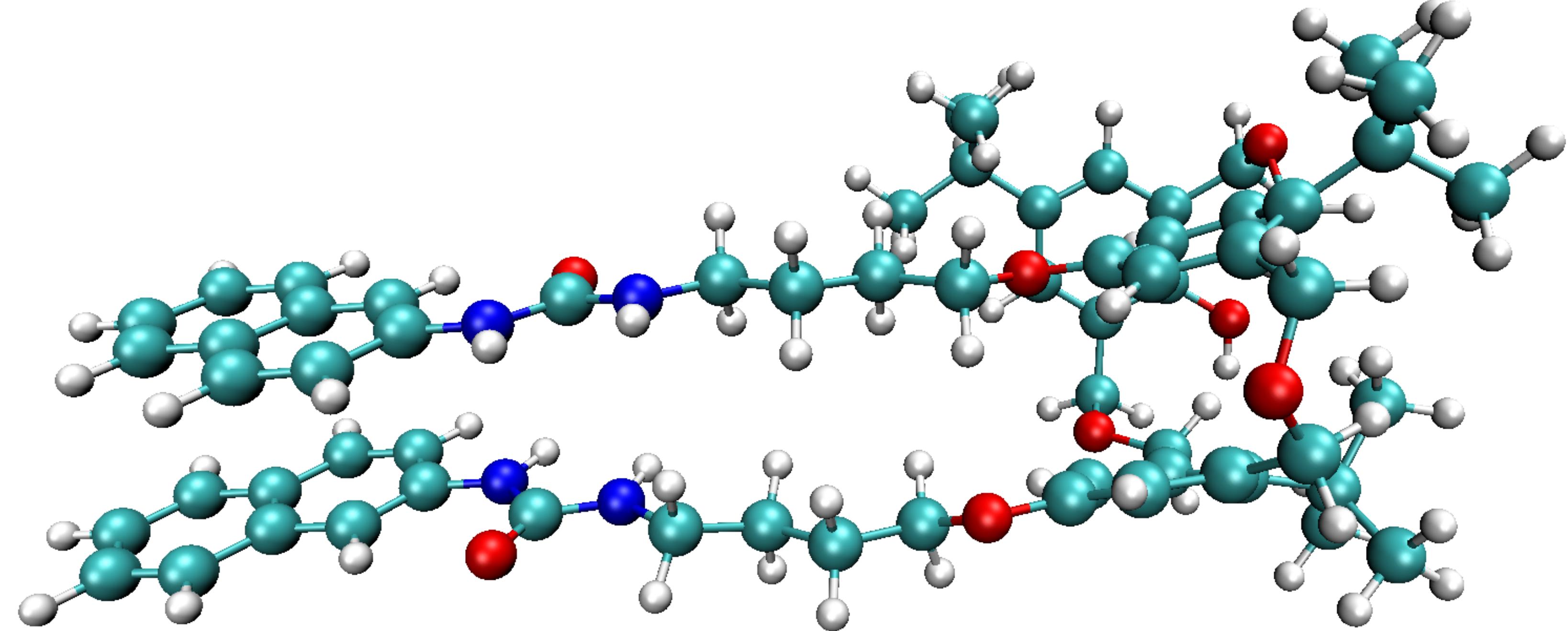
Lets use tree... searches



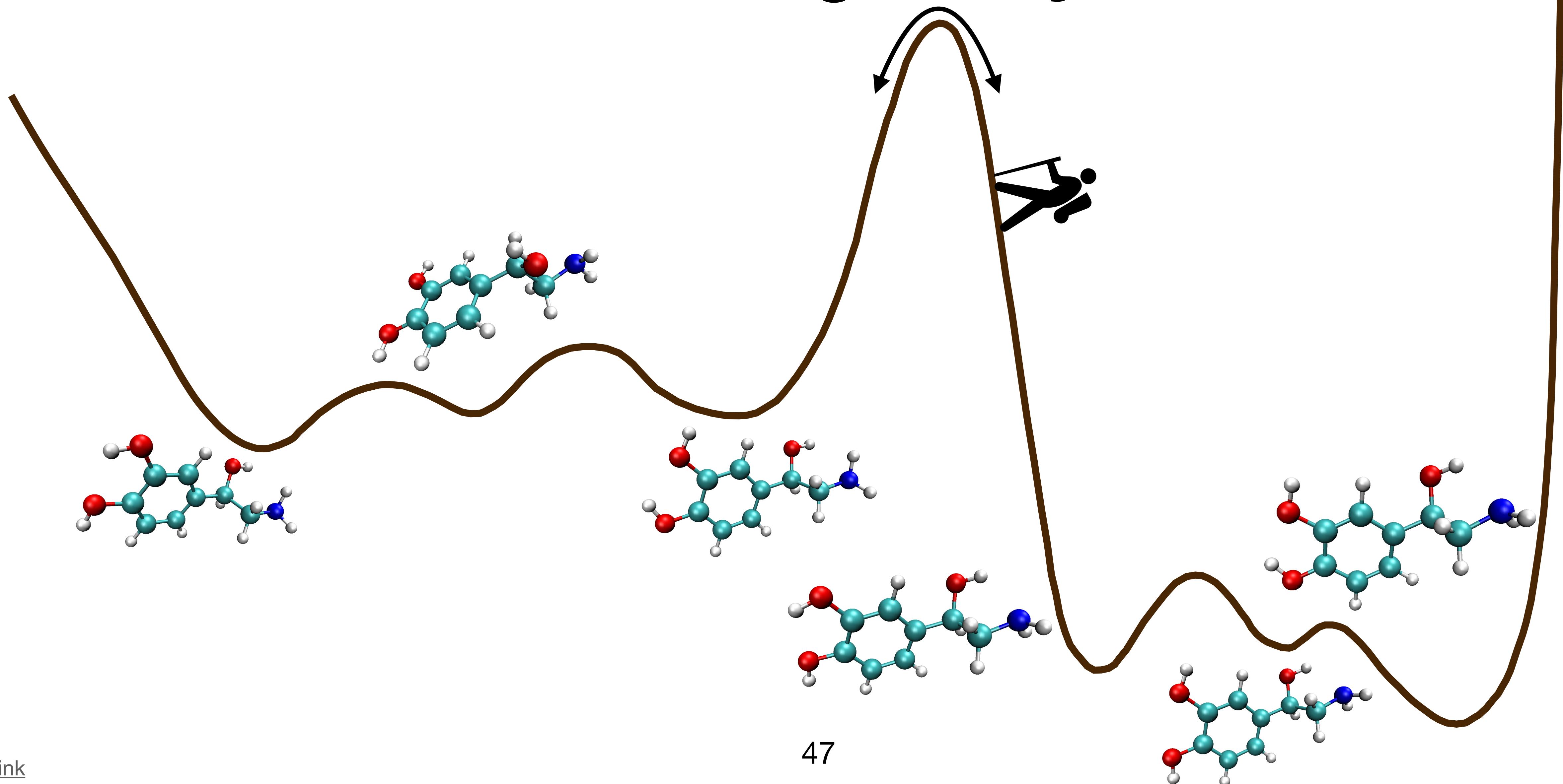
# Problems of being Systematic



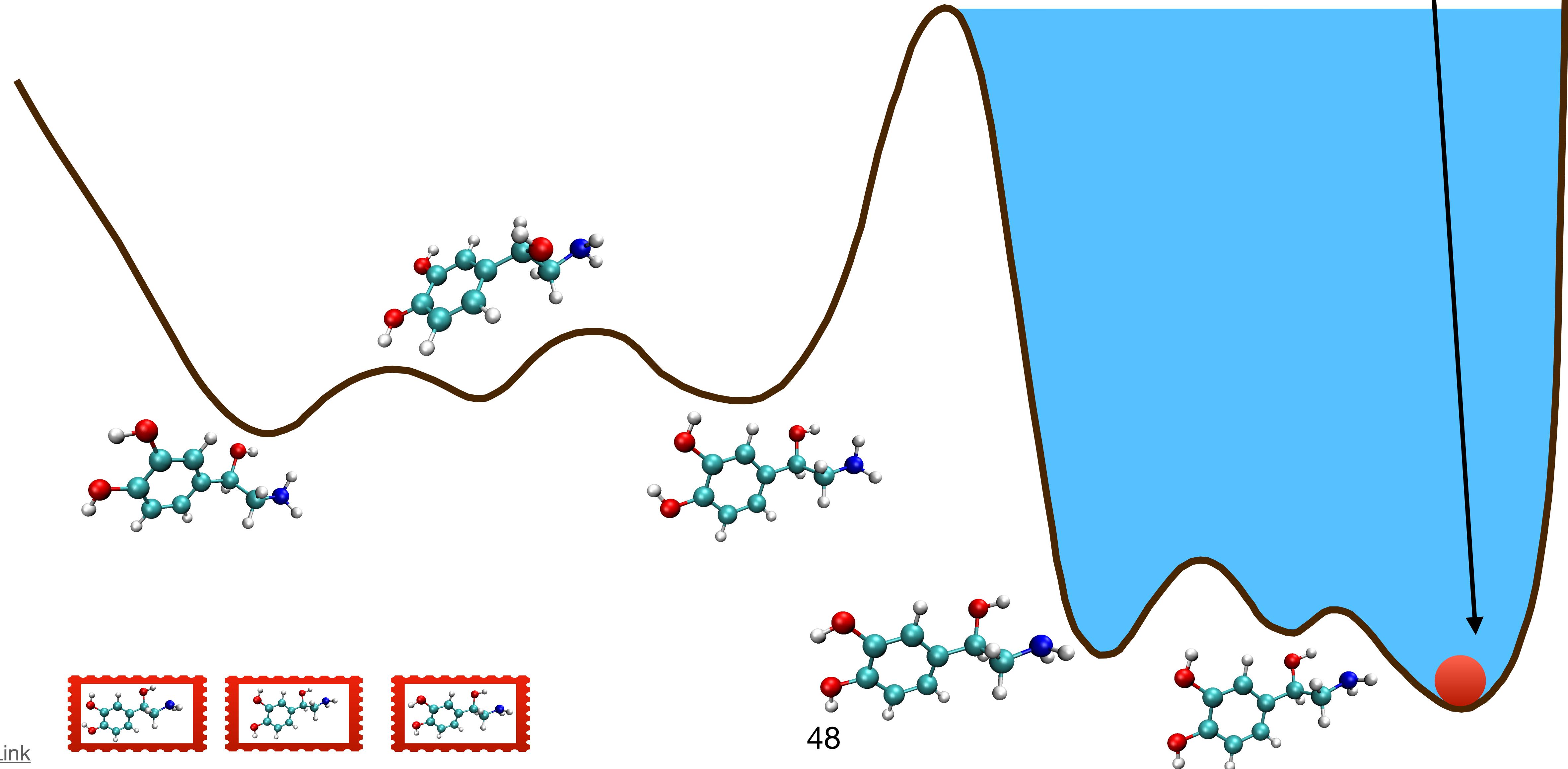
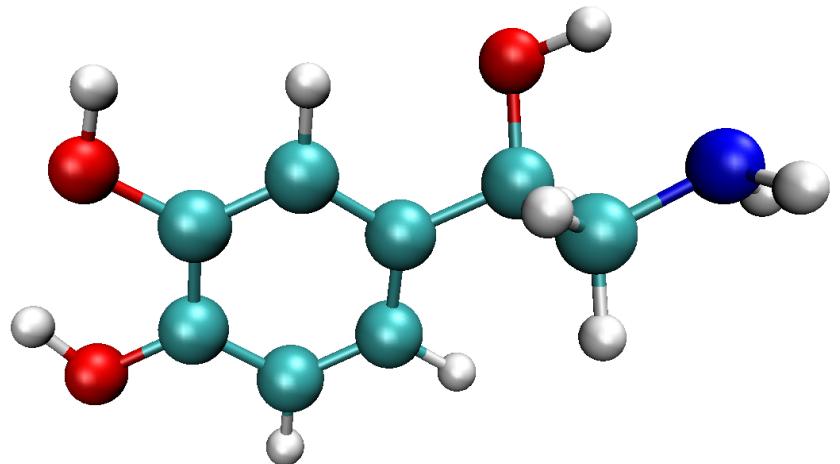
# Problems of being Systematic



# Walk Over PES and Ergodicity



# Speeding Ergodicity



Link

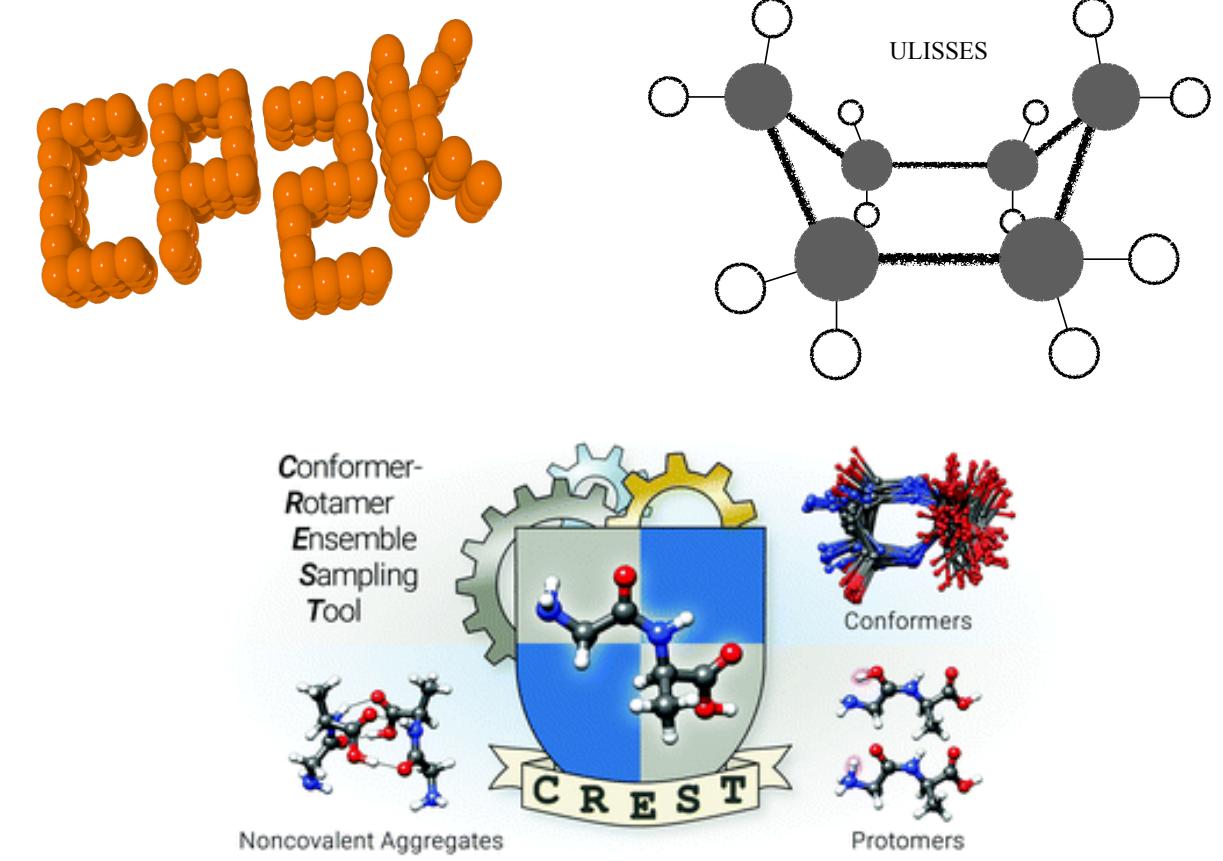
48

# Metadynamics Conformer Search

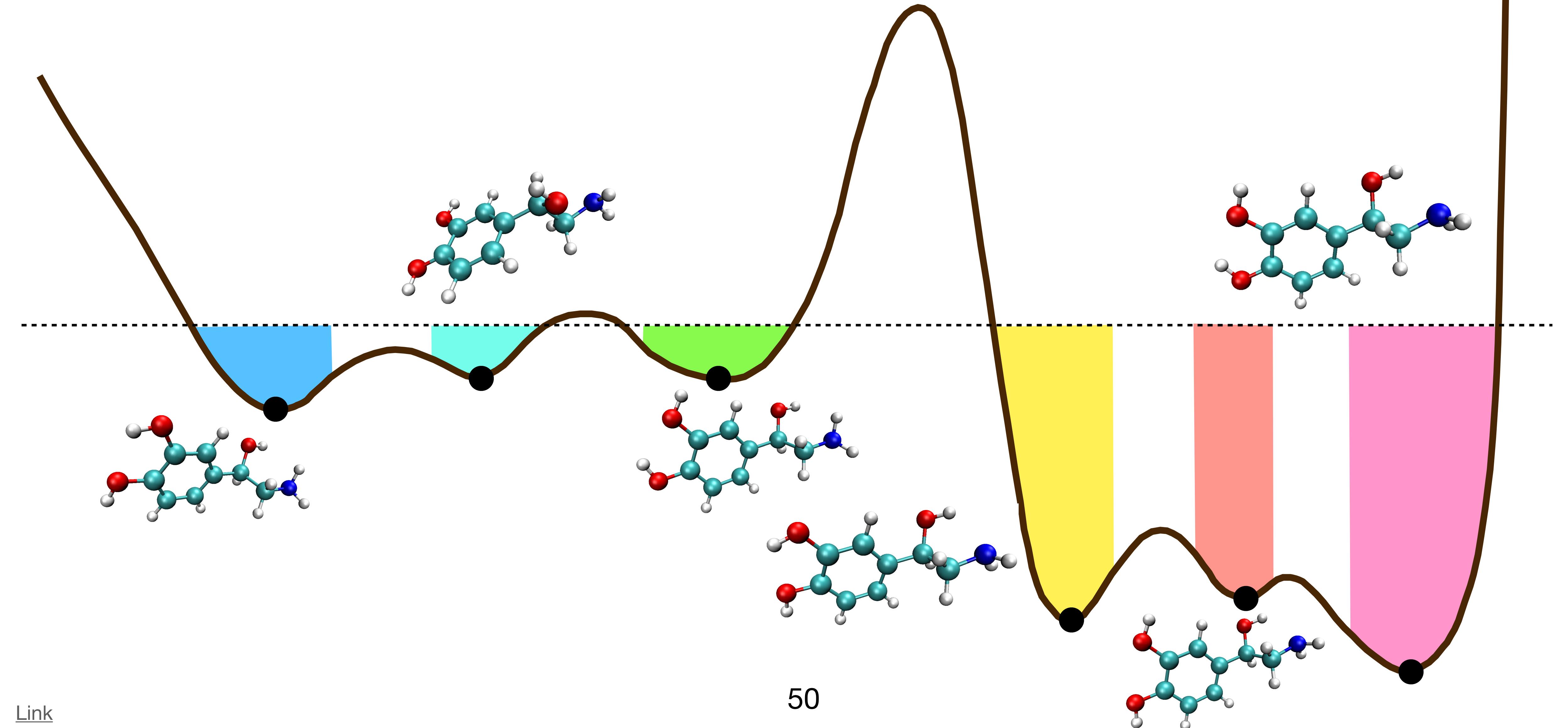
- Start with fast, but reasonable  
Only for small molecules, different methods yield similar minima.
- After running low-level model chemistries, increase the level as much as possible for your resources.

$$FF < SEQM < DFT \leq \Psi$$

- In between runs, optimise all structures.
- Use previous ensemble as starting point for next run.



# How Important are the conformers?

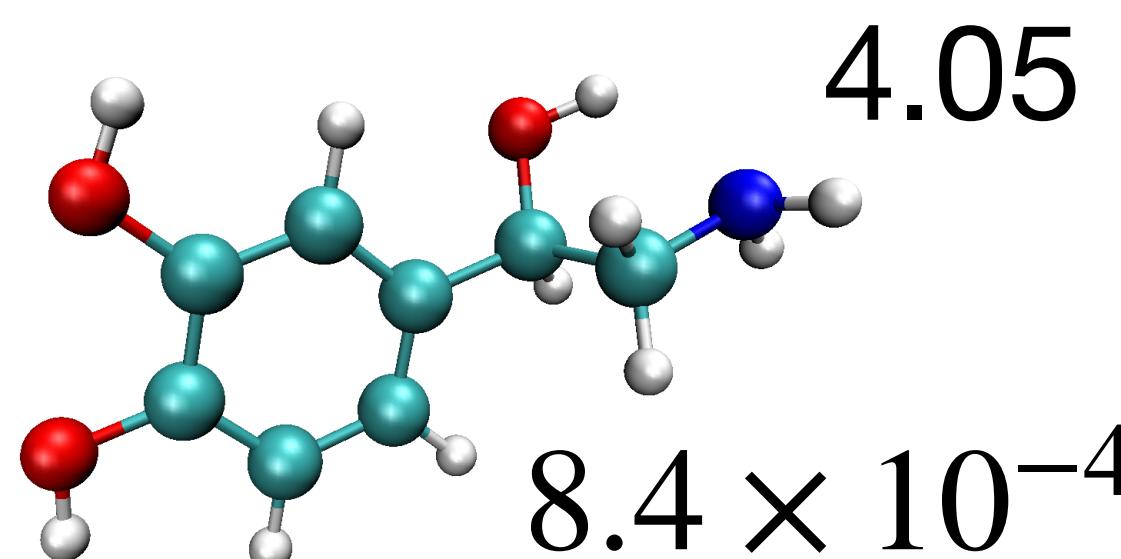
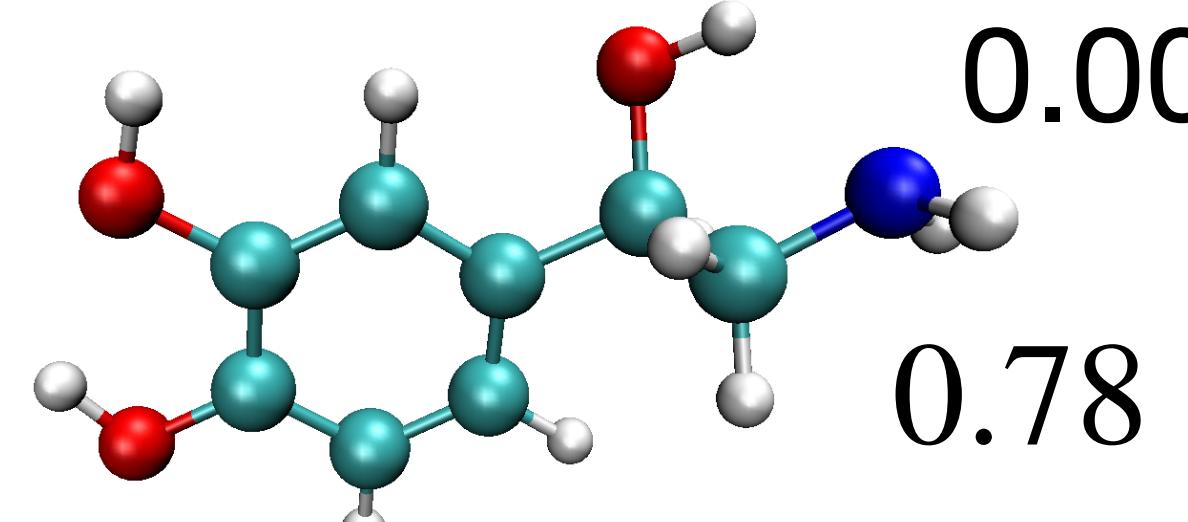
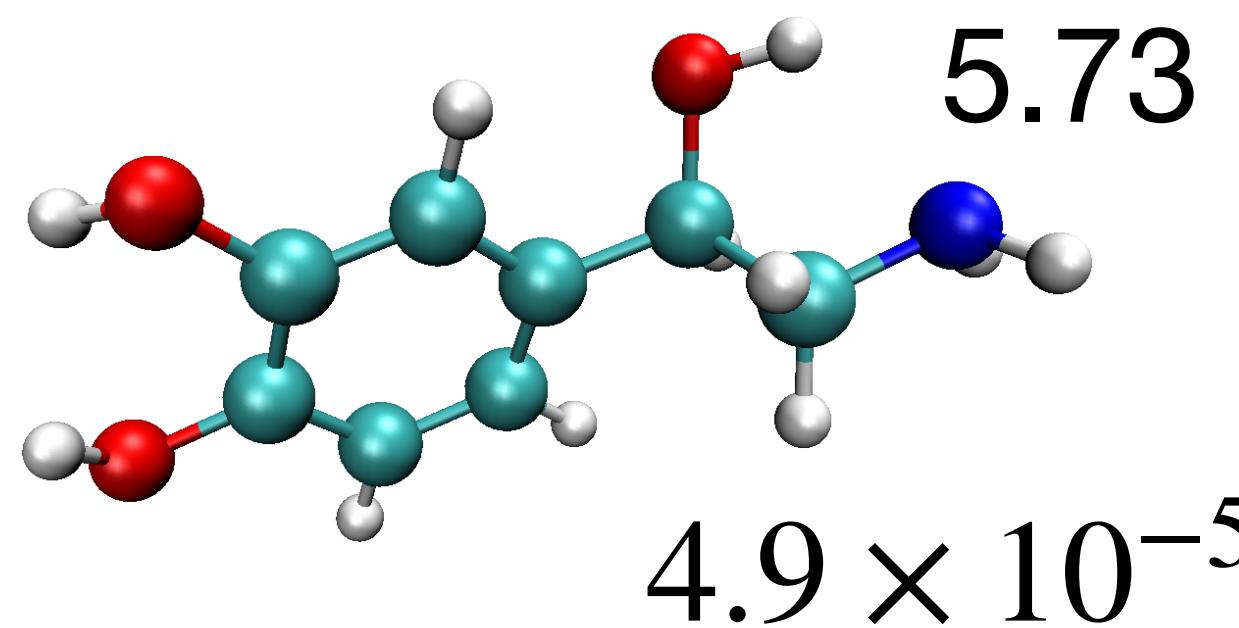


50

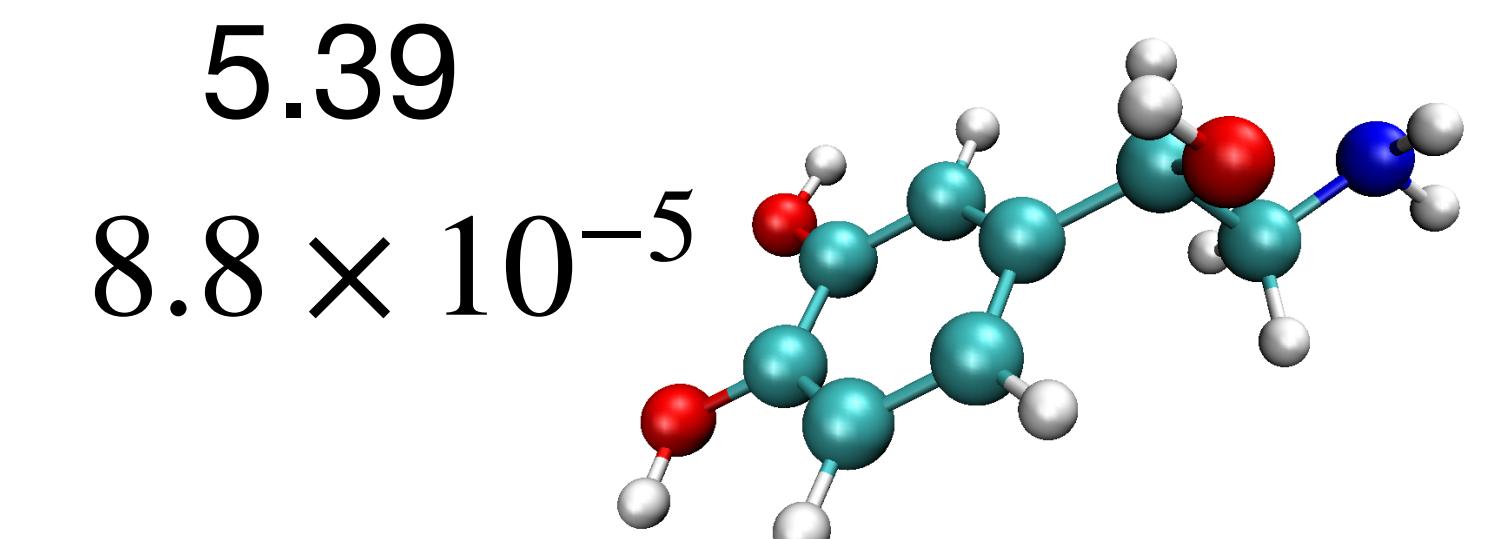
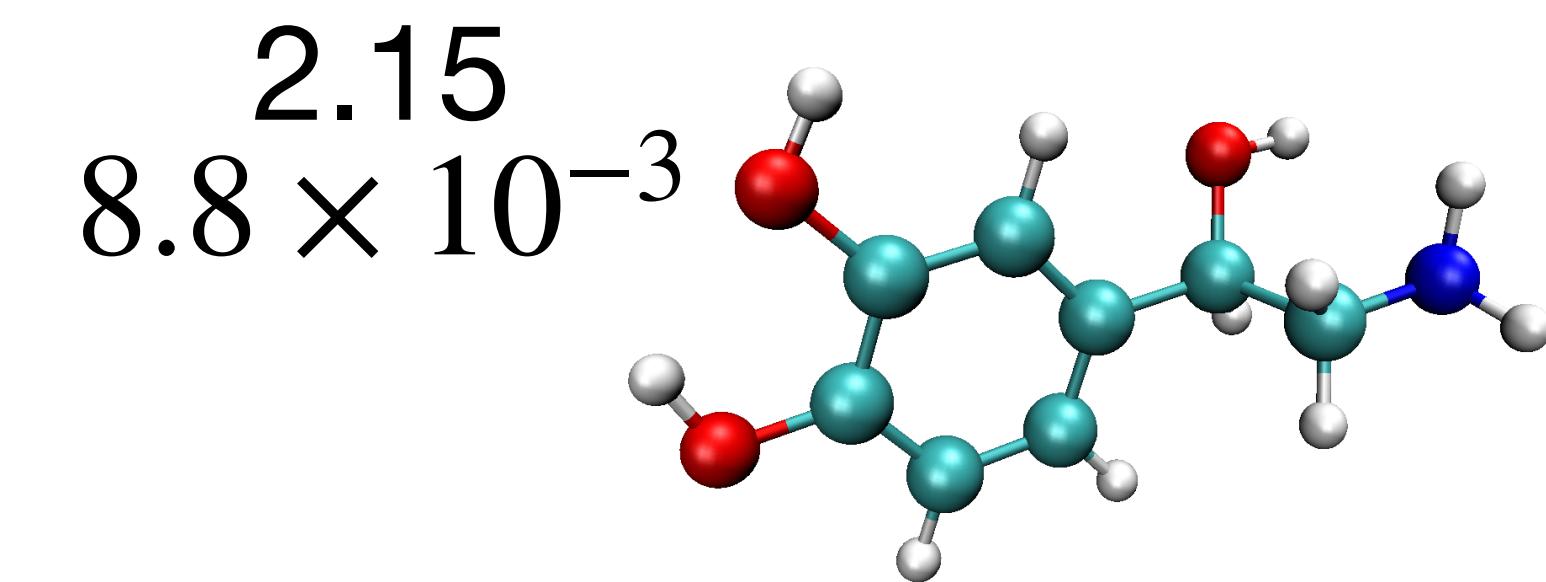
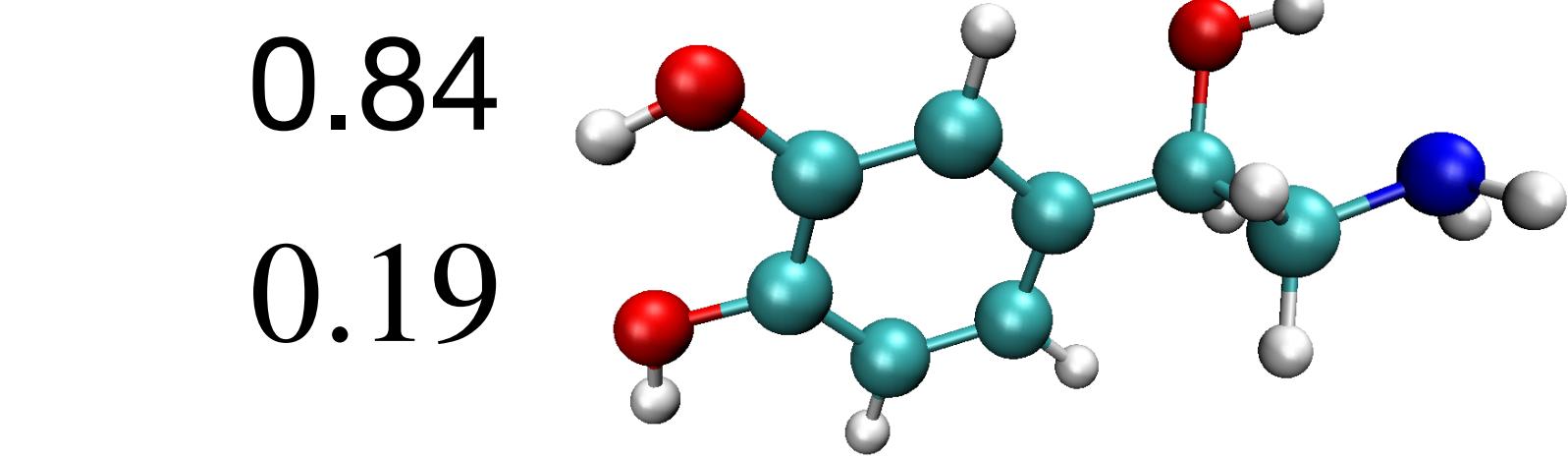
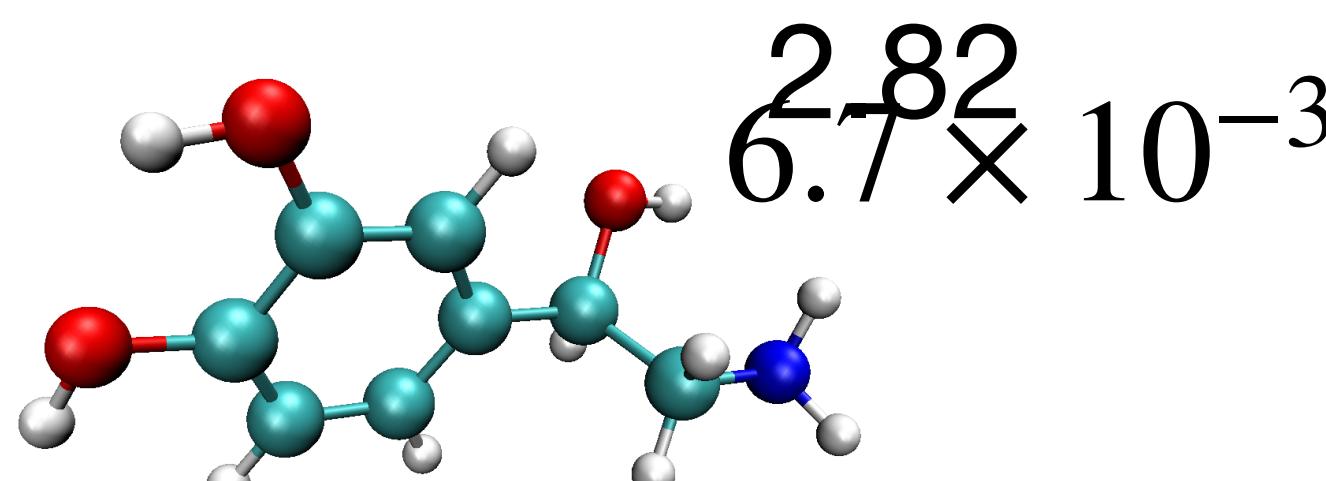


Ludwig Boltzmann

# Which Conformers are relevant?



$$P_i = \frac{\sigma \cdot \rho \chi n}{1 + \left( \frac{\Delta E_i}{k_B T} \right)^{1.4 \times 10^{-6}}}$$



This is a semi-classical approach, valid when quantum effects are not relevant for statistics.



# Which Conformers are relevant?

Ludwig Boltzmann

$$P_i = \frac{g_i \exp\left(-\frac{\Delta E_i}{k_B T}\right)}{\sum g_j \exp\left(-\frac{\Delta E_j}{k_B T}\right)}$$

$$P_i = \frac{g_i \exp\left(-\frac{\Delta G_i}{k_B T}\right)}{\sum g_j \exp\left(-\frac{\Delta G_j}{k_B T}\right)}$$

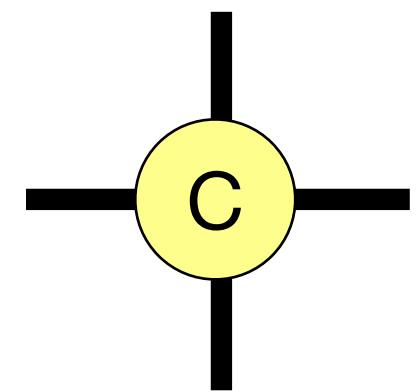
## **VALIDITY:**

- Gas phase
- All structures similar  
(Similar vibrations)

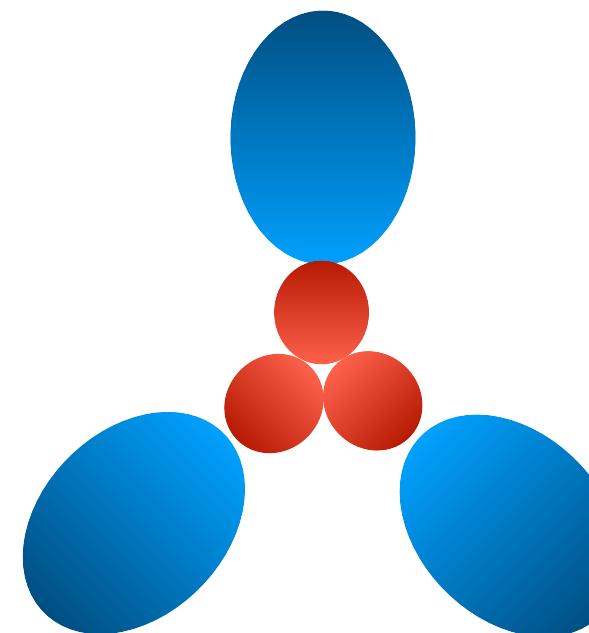
## **VALIDITY:**

- Any phase
- Entropy is considered

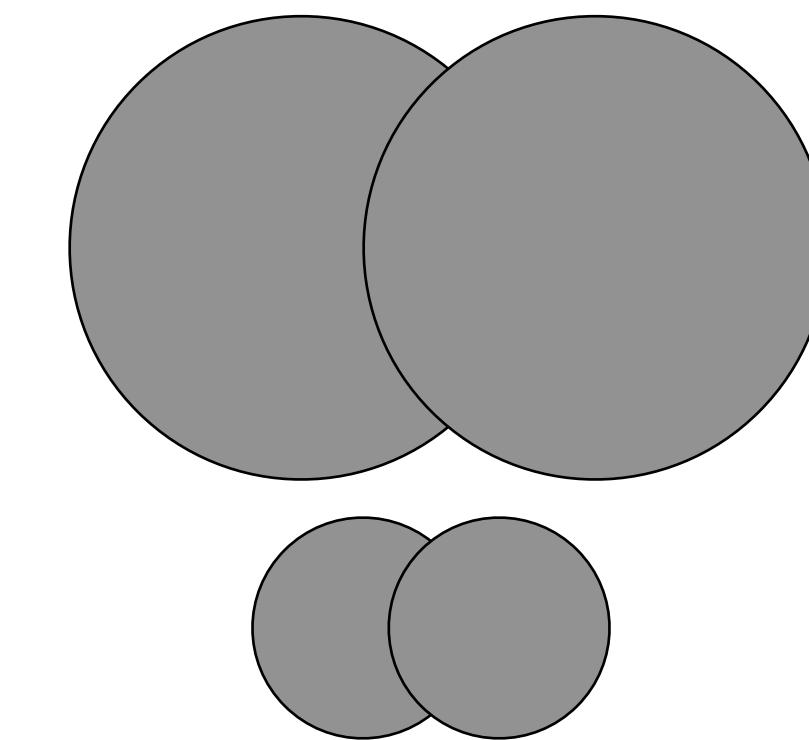
# Summary



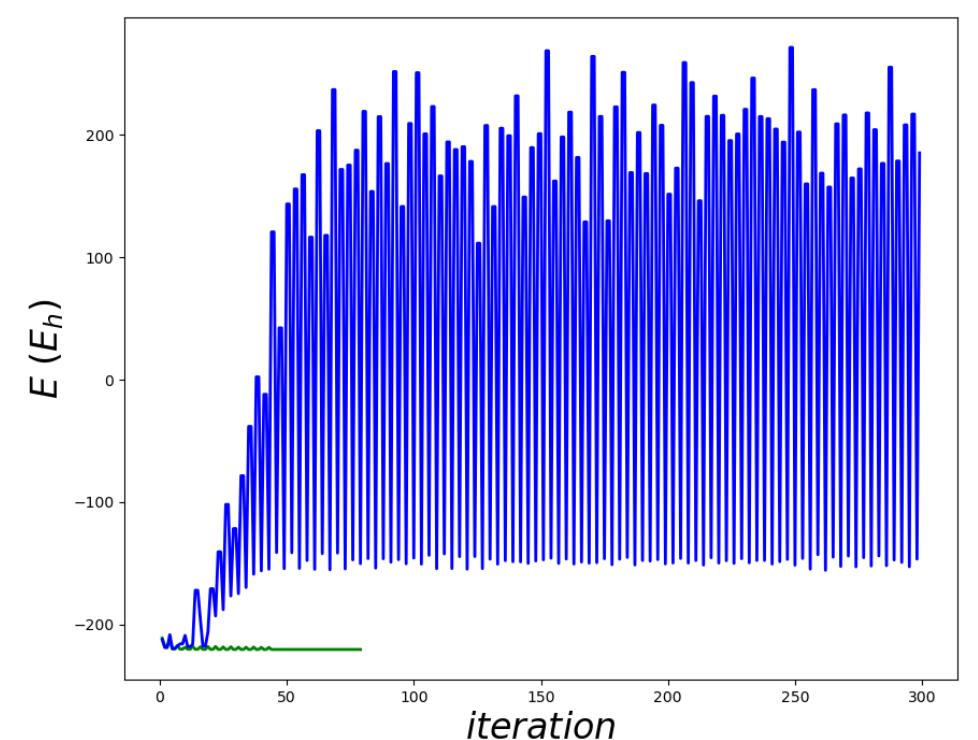
Octet Rule



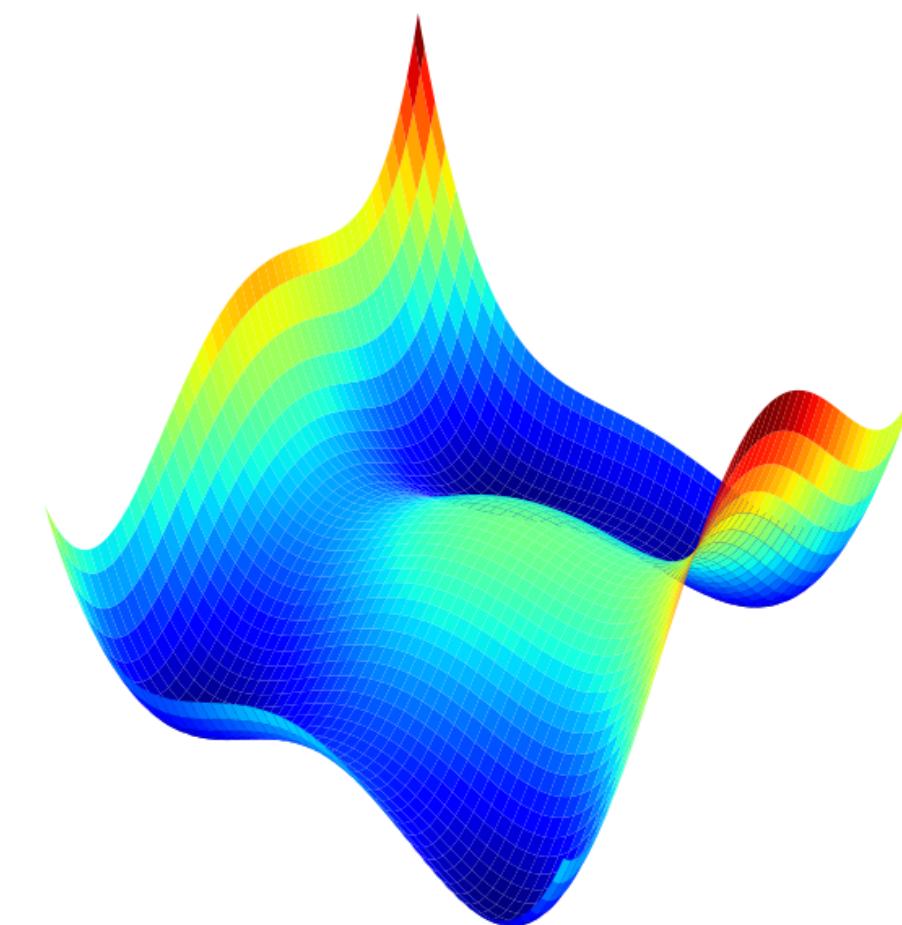
Hybridization



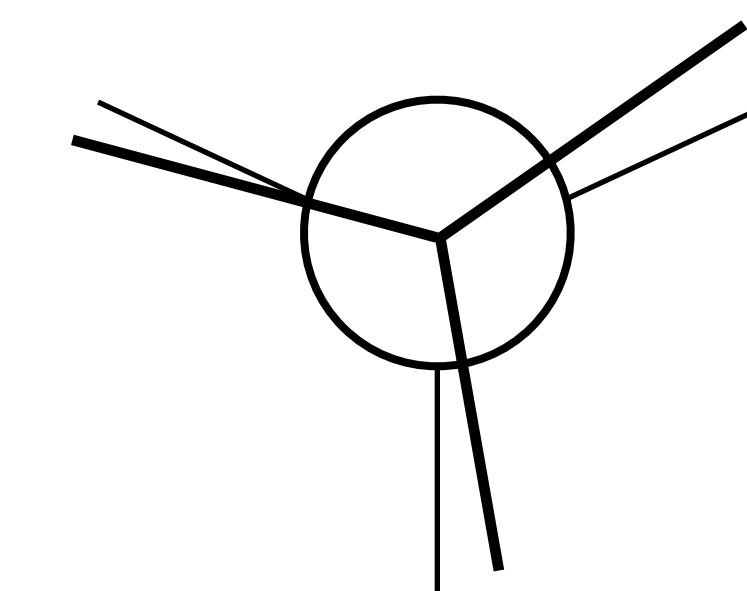
Hard-Soft



Convergence Analysis



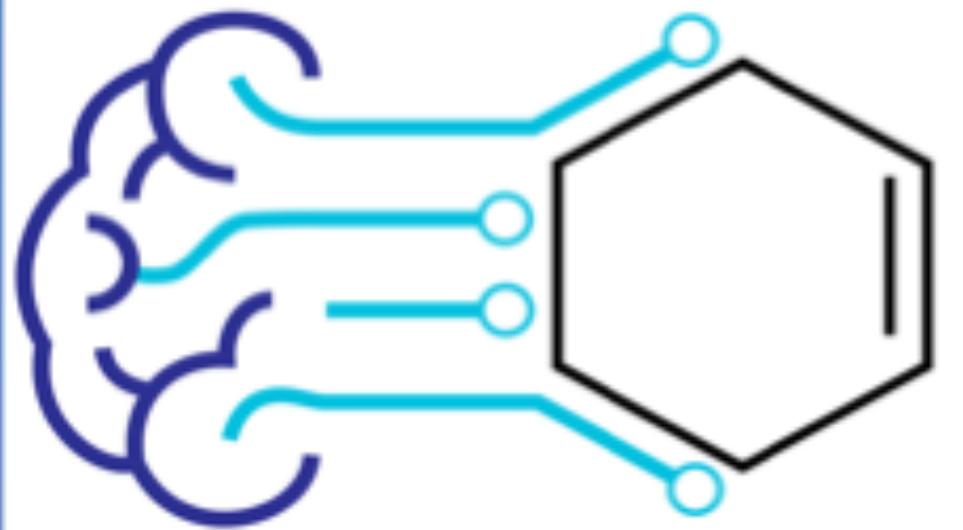
Optg, MD



Conformers

# Acknowledgments

- Grzegorz Popowicz
- Michael Sattler
- Igor Tetko
- Pavel Karpov
- BMWi ZIM. KK 5197901TS0
- BMBF, SUPREME, 031L0268



Advanced machine learning for Innovative Drug Discovery

- Khumbu AI
- Sattler Group
- All Participants



**NOT SURE IF WE DID A GREAT JOB...**



*The End*

**...OR NO ONE PAID ATTENTION**

# Conformational Entropy and Ergodicity

$$S_{conf} = -R \sum P_i \log P_i$$

56

$$\Delta S \geq 0$$

# Conformational Entropy and Ergodicity

