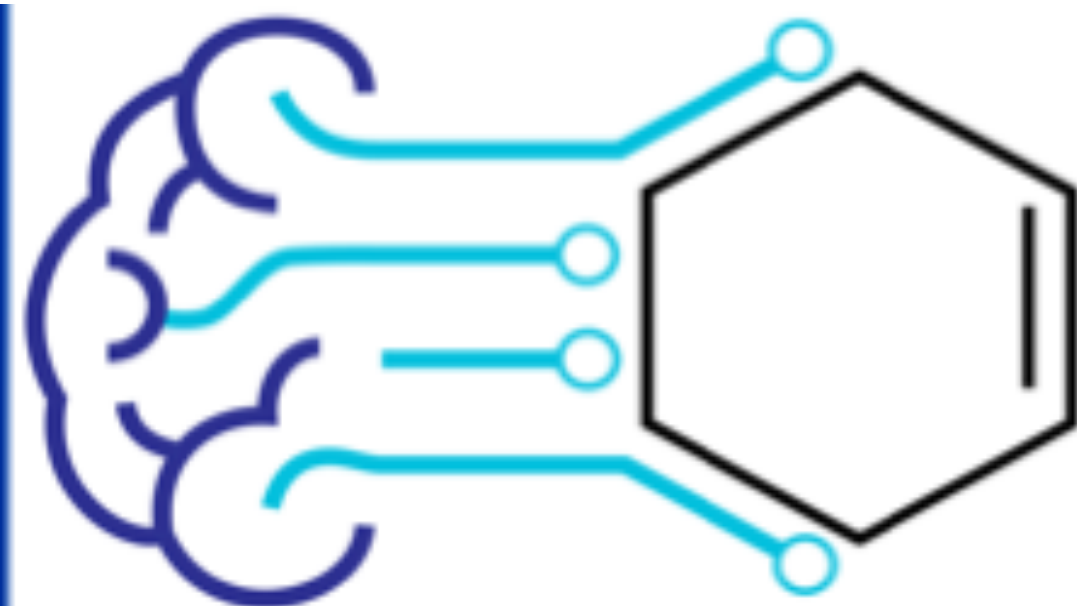




Advanced machine learning for  
Innovative Drug Discovery (AIDD)

# 3D-Structure Refinement

Simple Principles for Meaningful Structures

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

HELMHOLTZ  
MUNICH

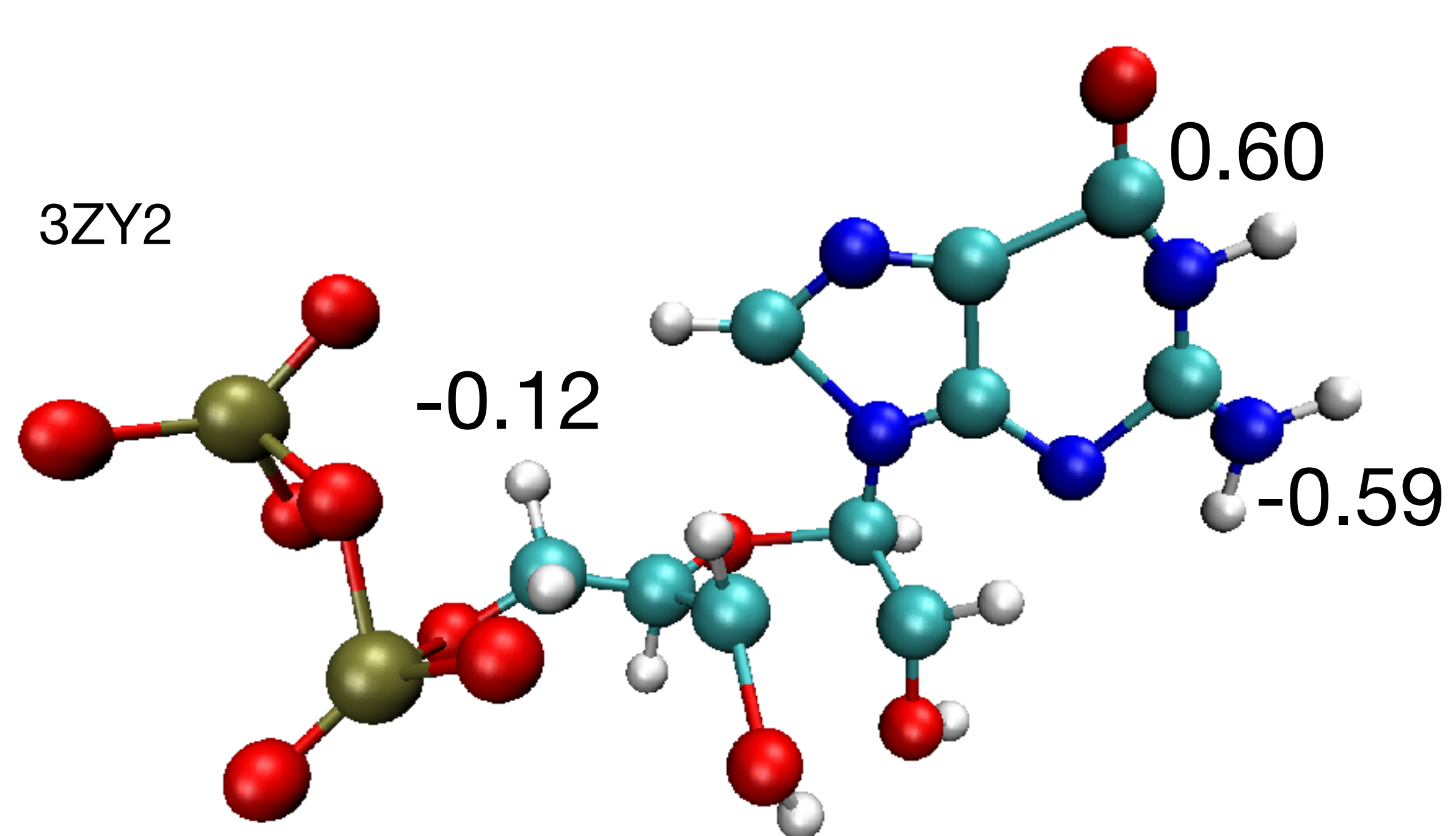


Institute of  
Structural Biology

Bayerisches NMR Zentrum

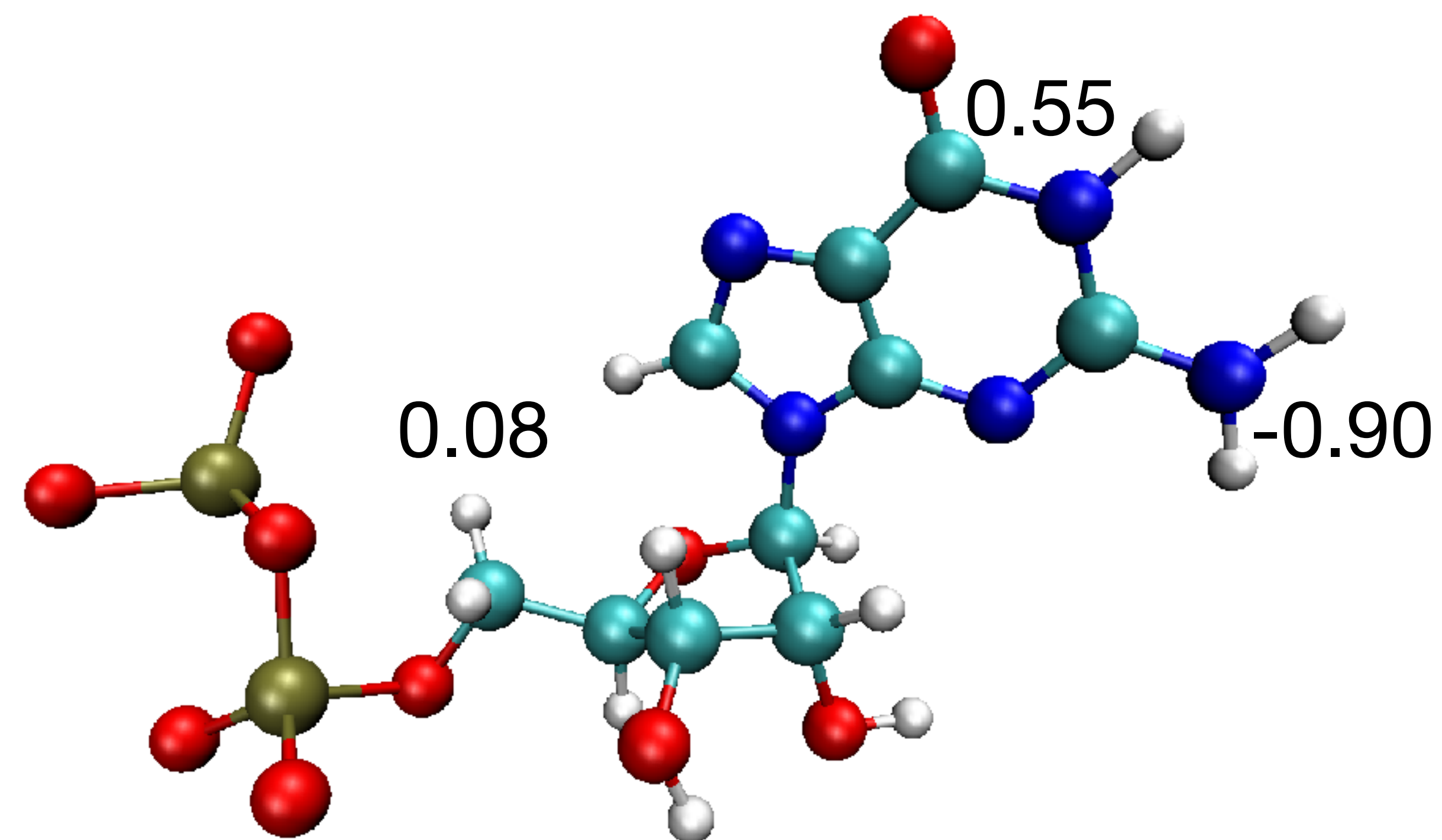


# Why do we need good structures?



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

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

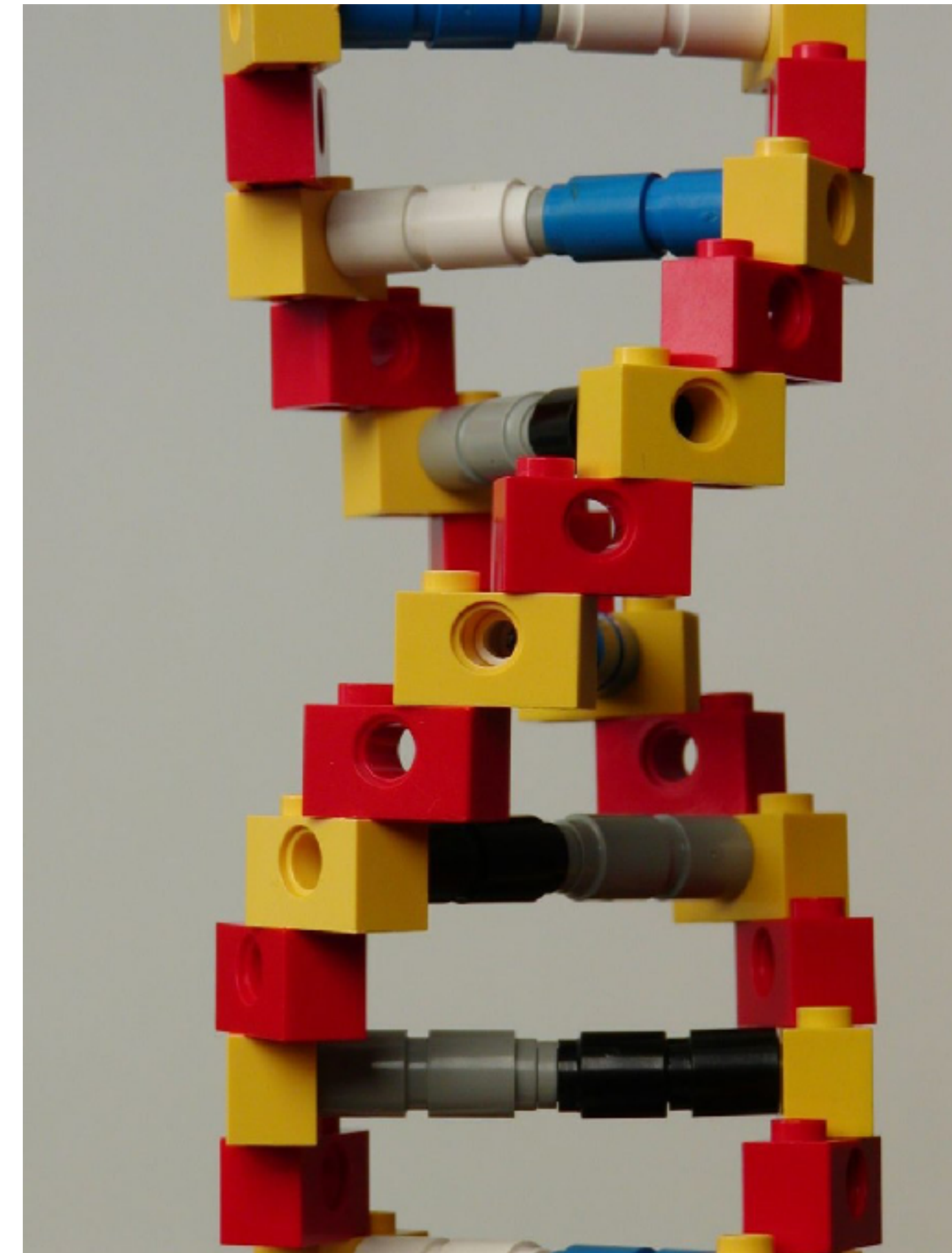


$$\Delta_{solv}G = -513.39 \text{ kcal/mol}$$

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

# Part 1

## Assembling Molecules



[Link](#)

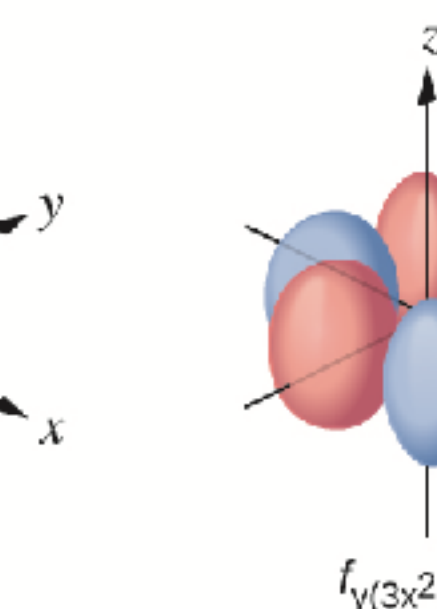
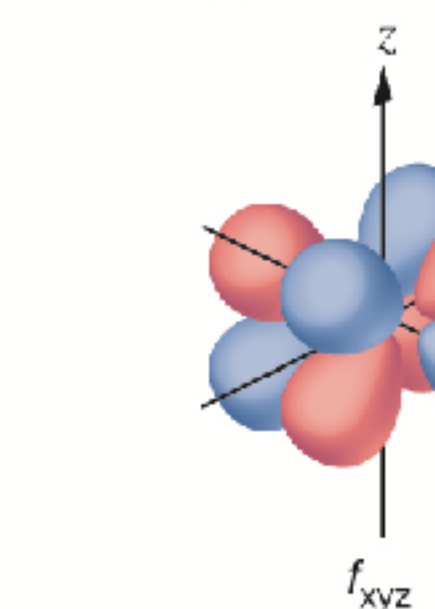
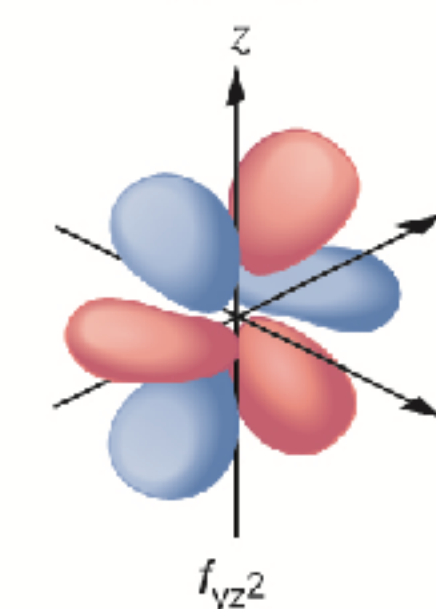
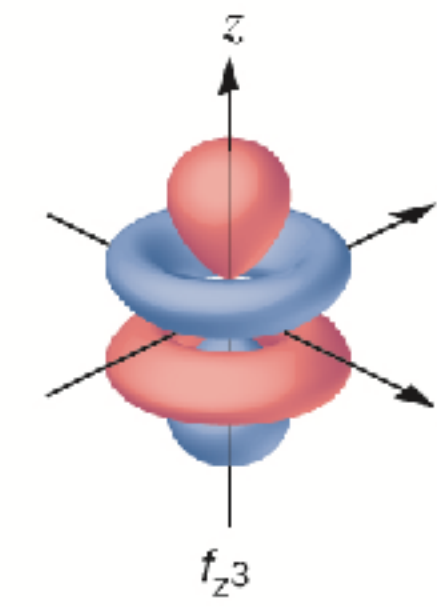
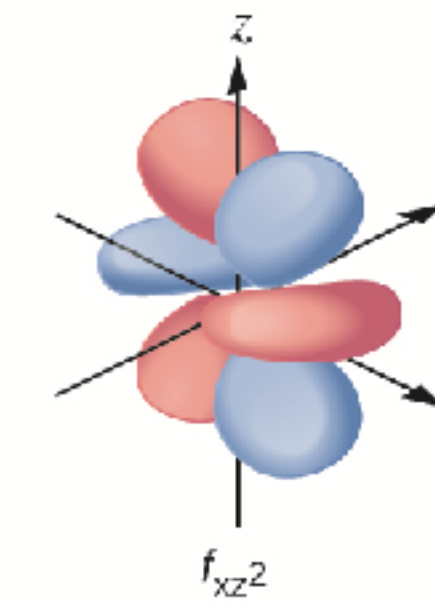
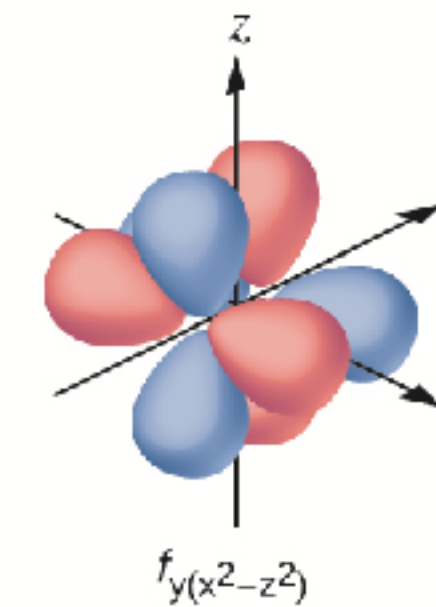
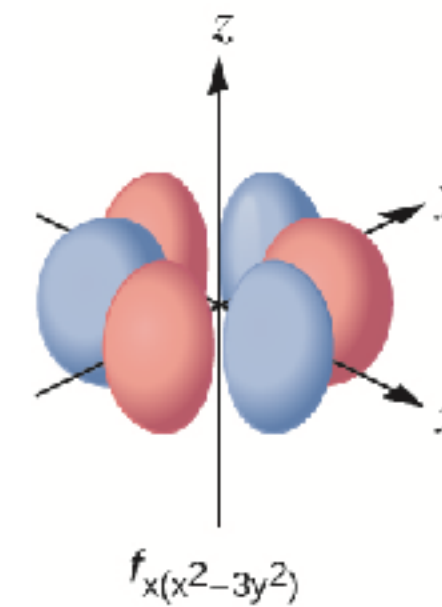
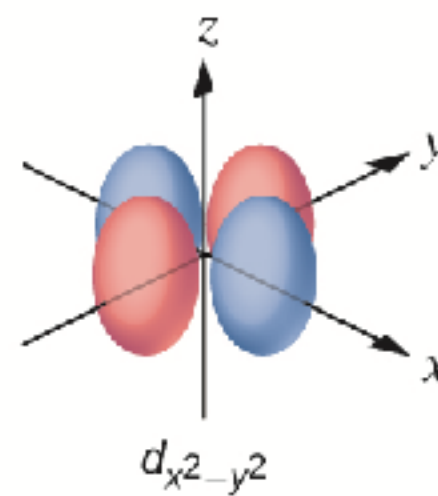
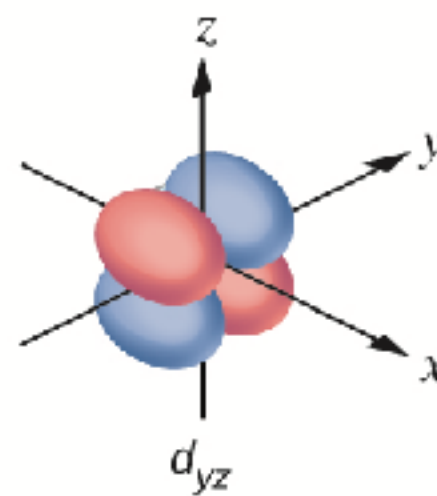
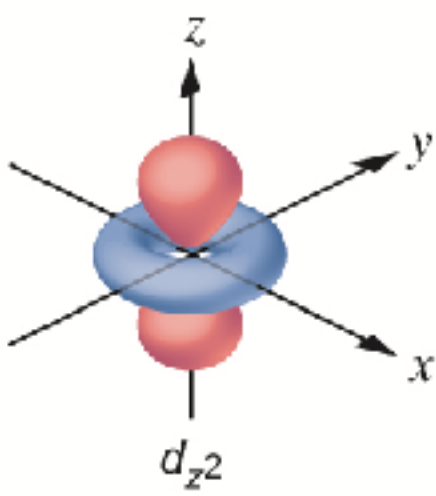
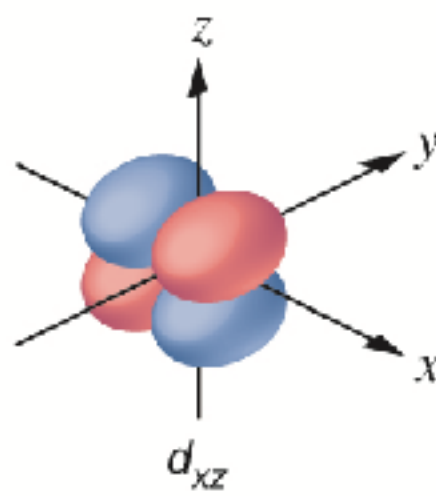
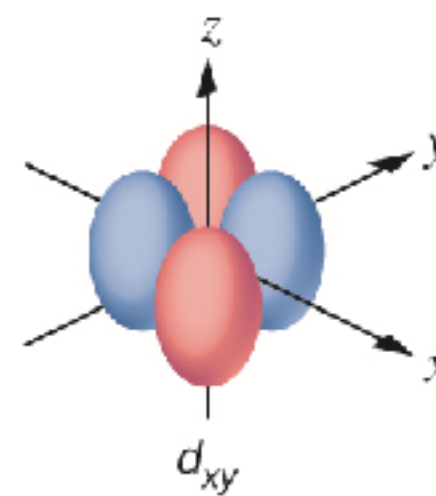
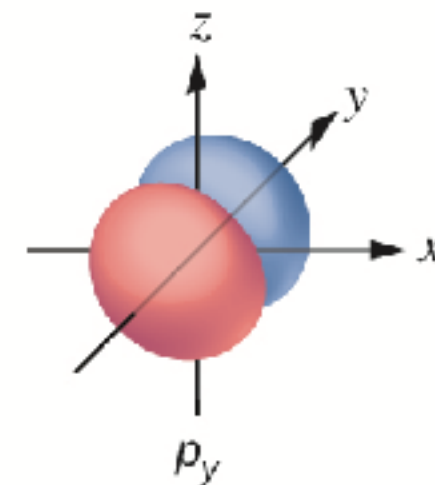
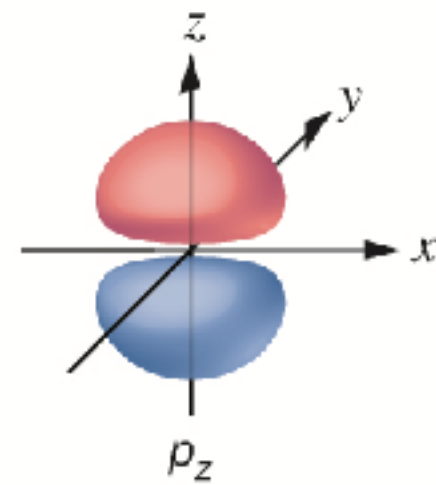
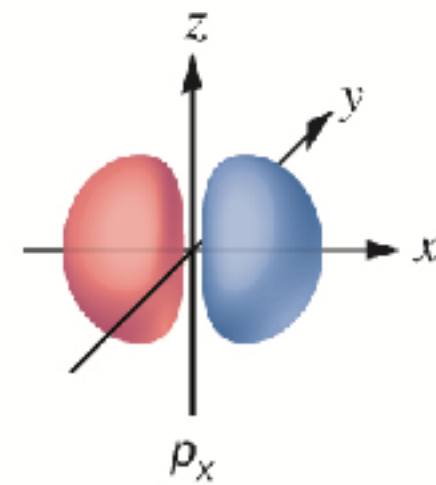
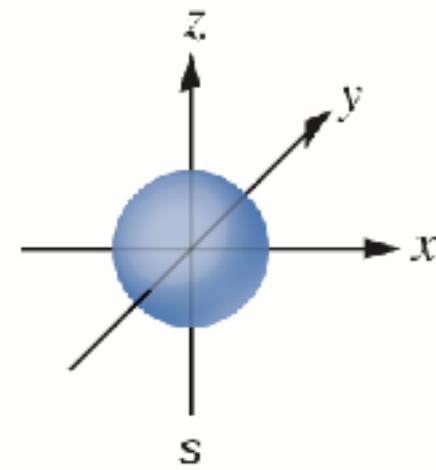


# The Schrödinger Equation



Erwin Schrödinger

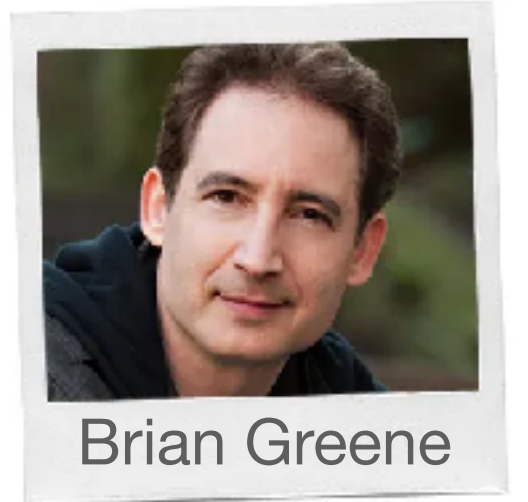
$$(\hat{T} + \hat{V})\Psi = E\Psi$$





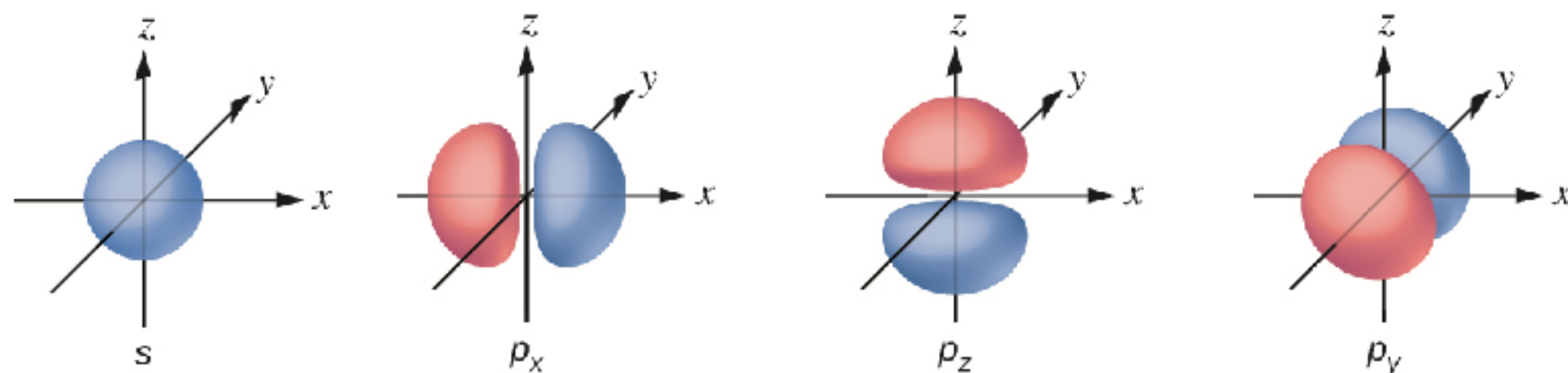
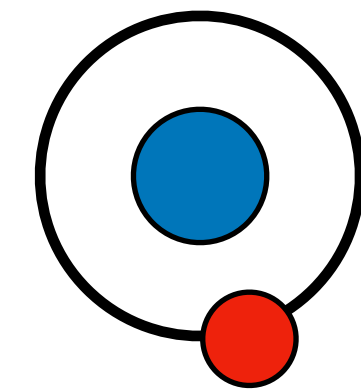
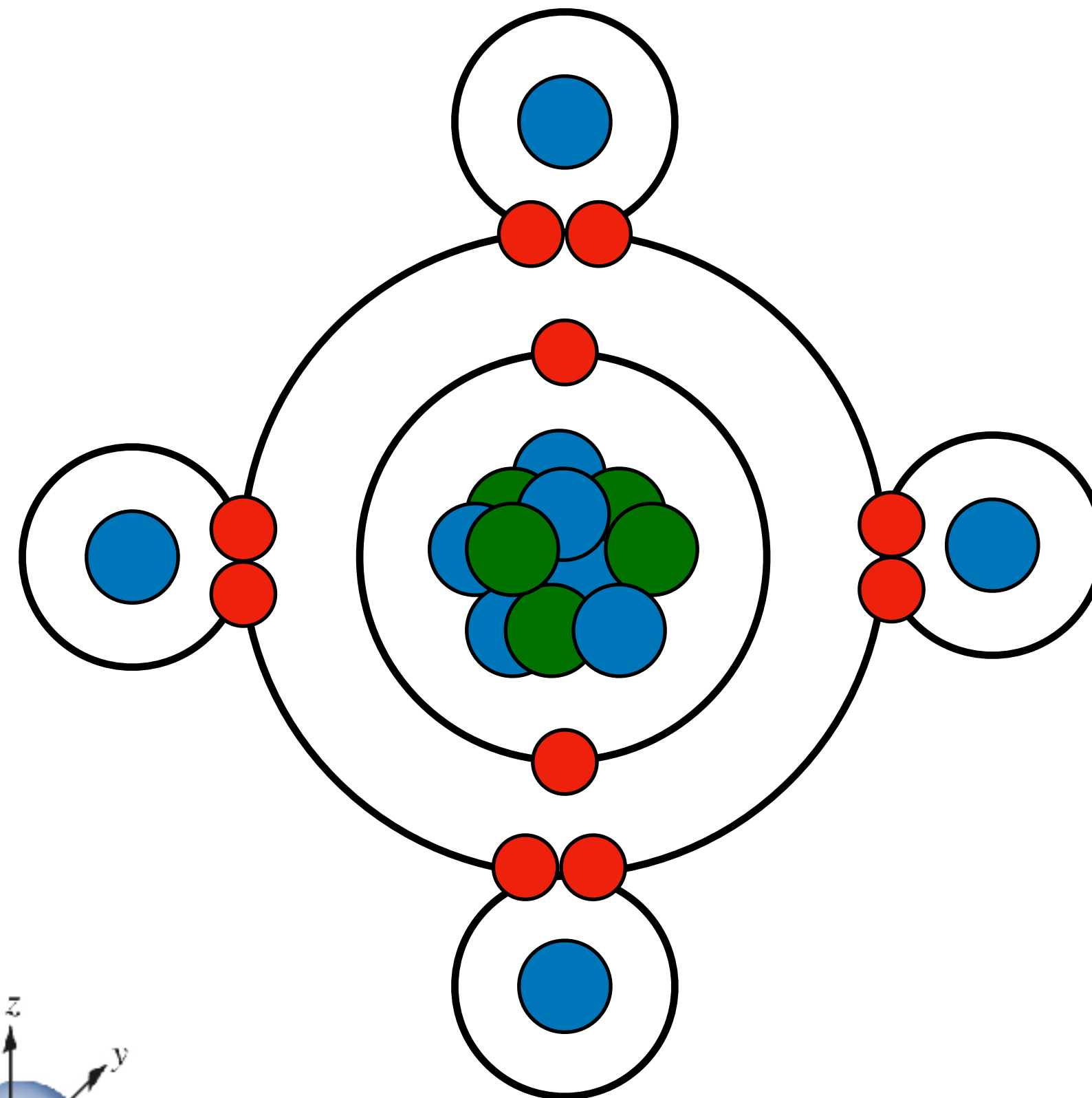
# Base Principle

Atoms have OCD (obsessive compulsive disorder)



## Sharing is Caring

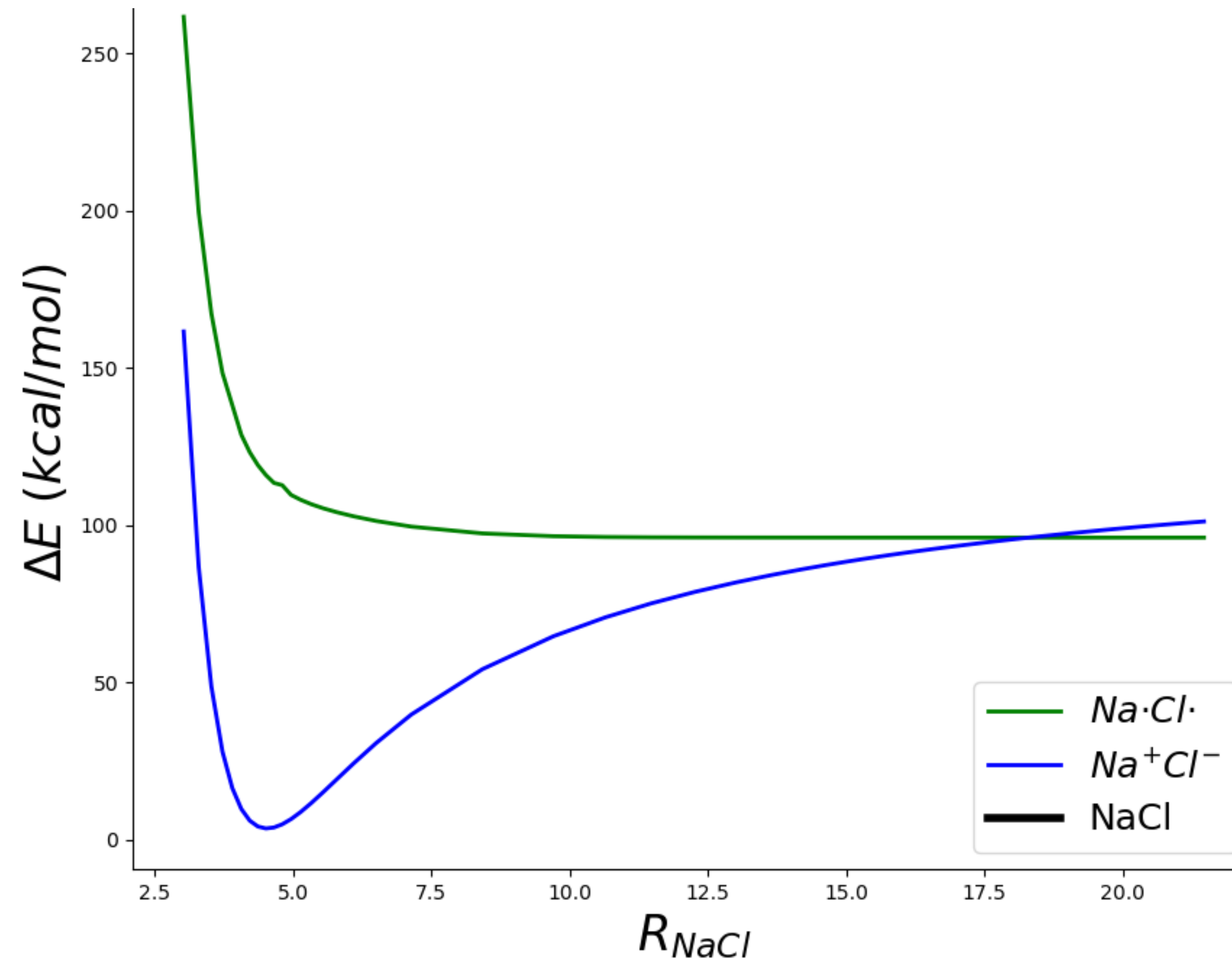
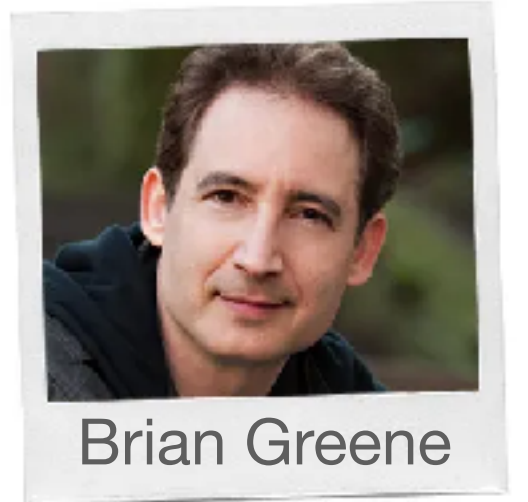
rather share electrons than having incomplete shells



- Proton
- Neutron
- Electron

# Base Principle

Atoms have OCD (obsessive compulsive disorder)

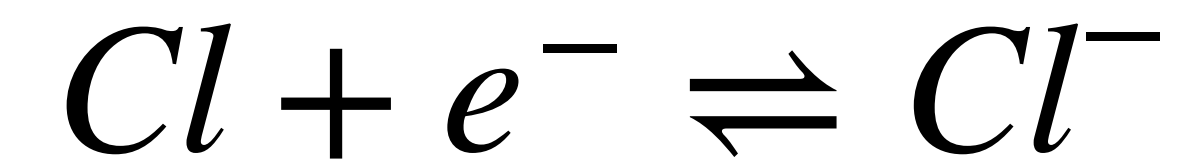


They will do anything to avoid unpaired electrons

- “Philanthropist atoms”:

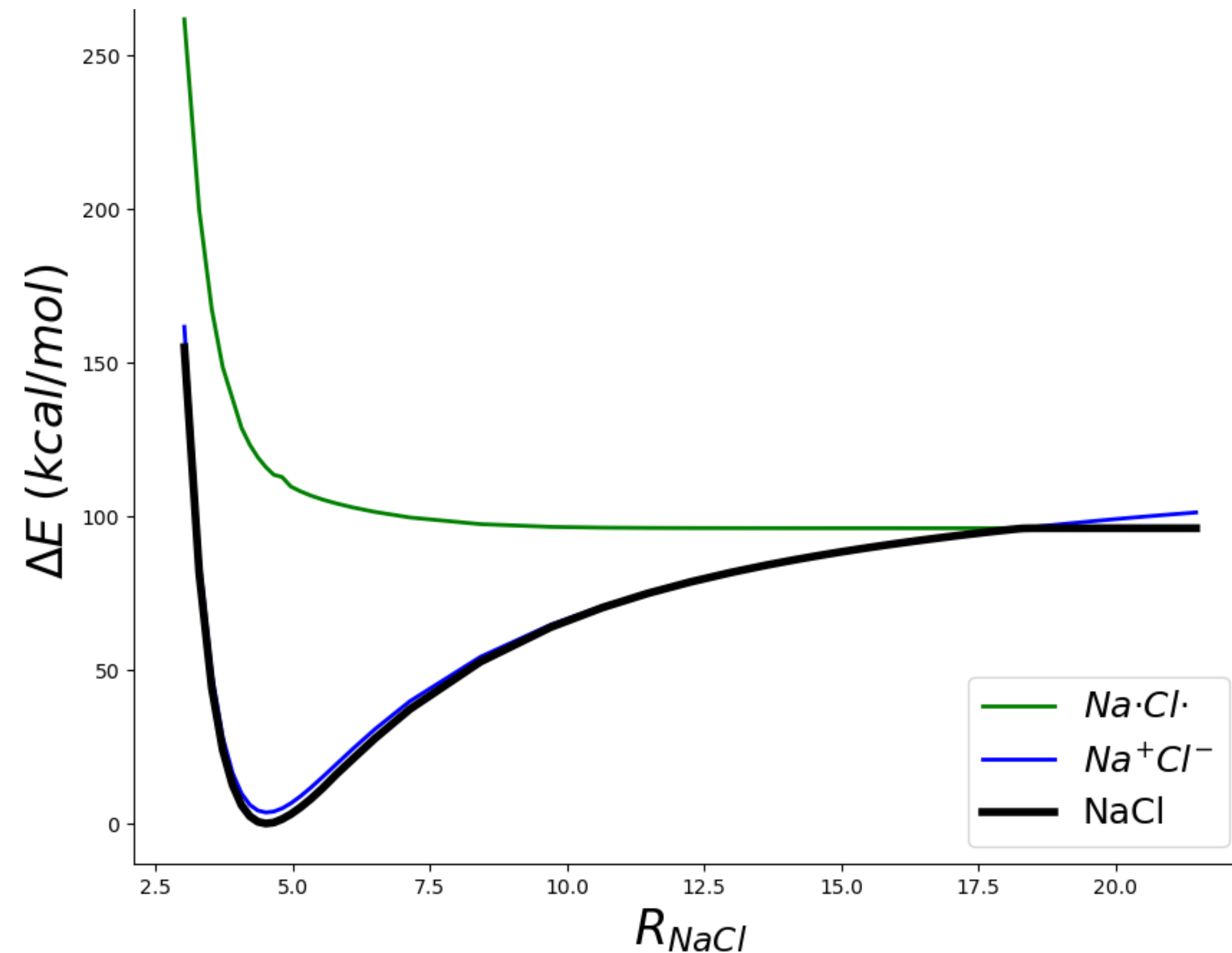
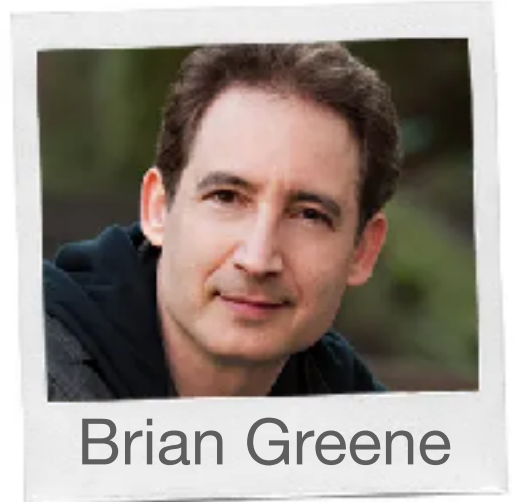


- “Looter atoms”:



# Base Principle

Atoms have OCD (obsessive compulsive disorder)

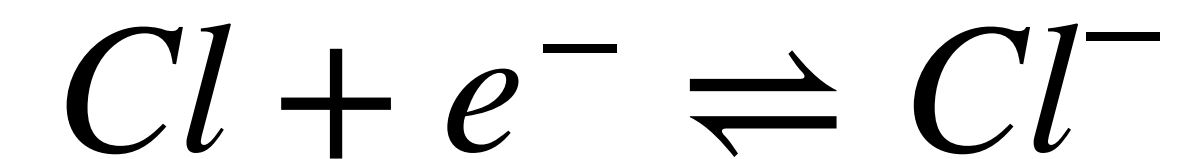


They will do anything to avoid unpaired electrons

- “Philanthropist atoms”:



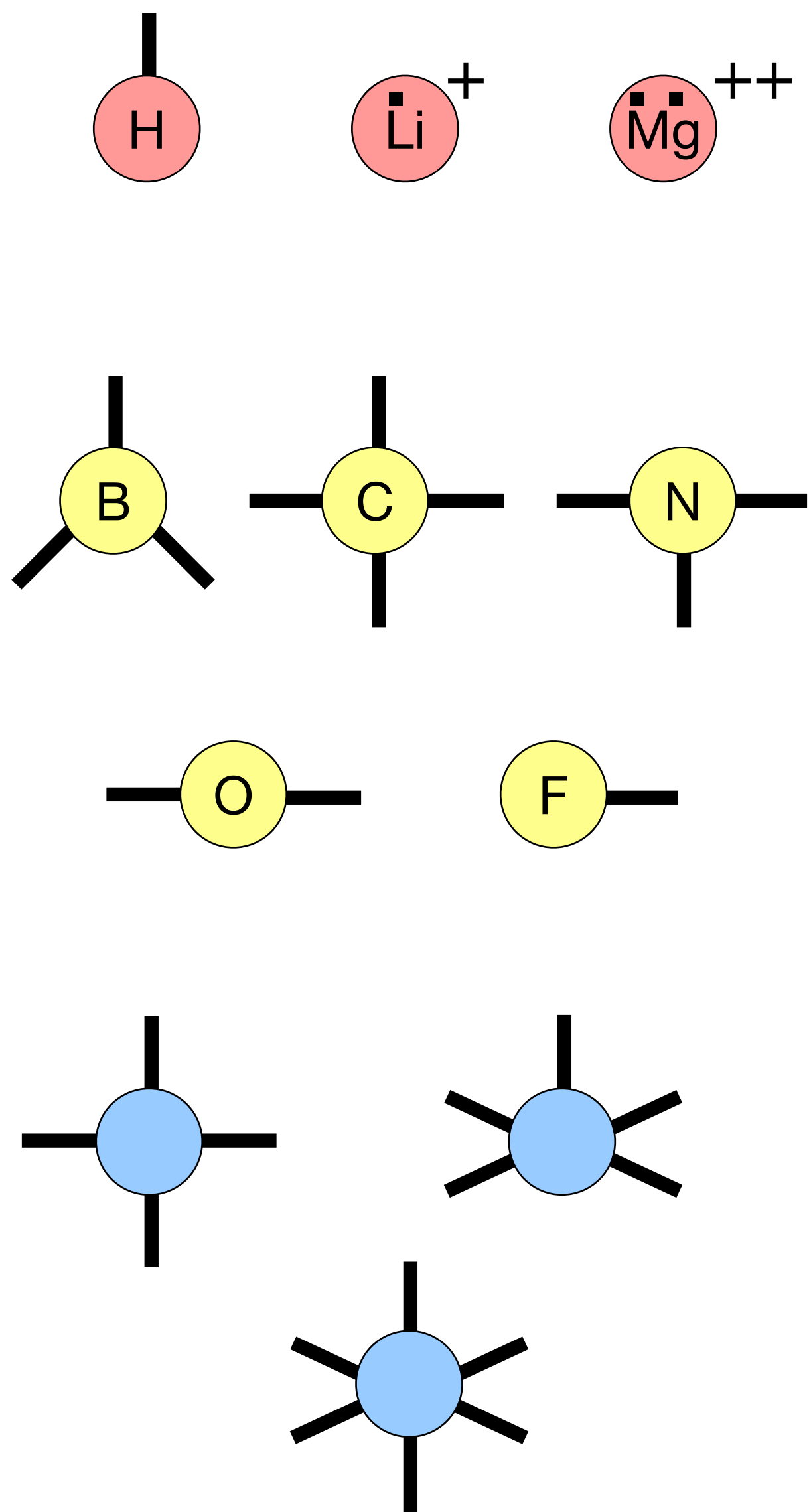
- “Looter atoms”:



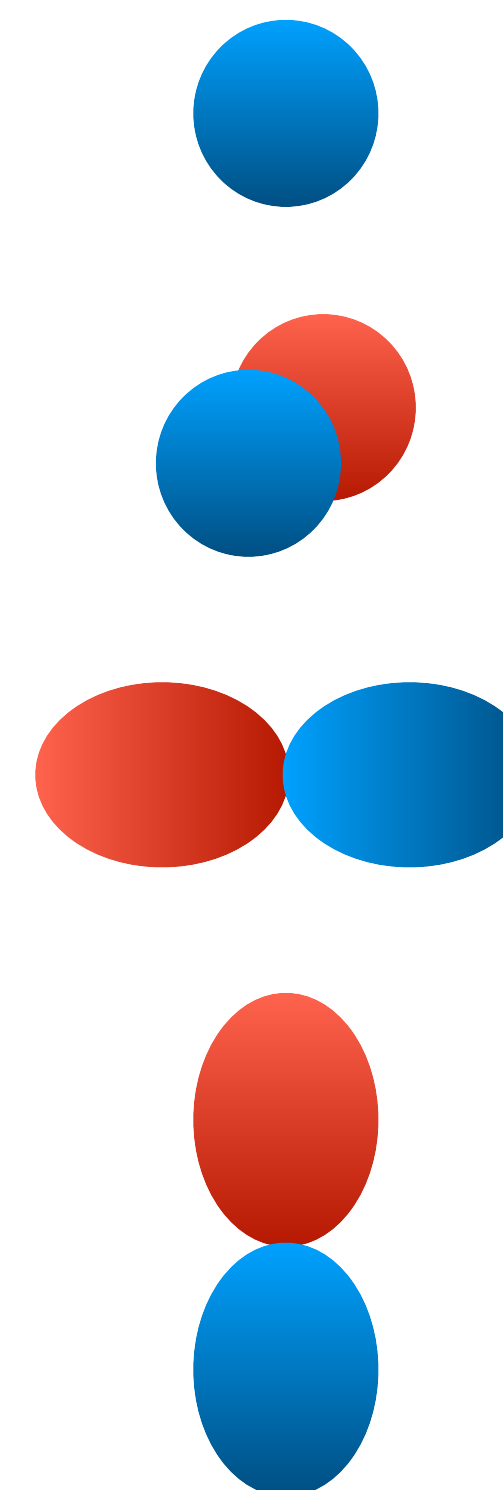
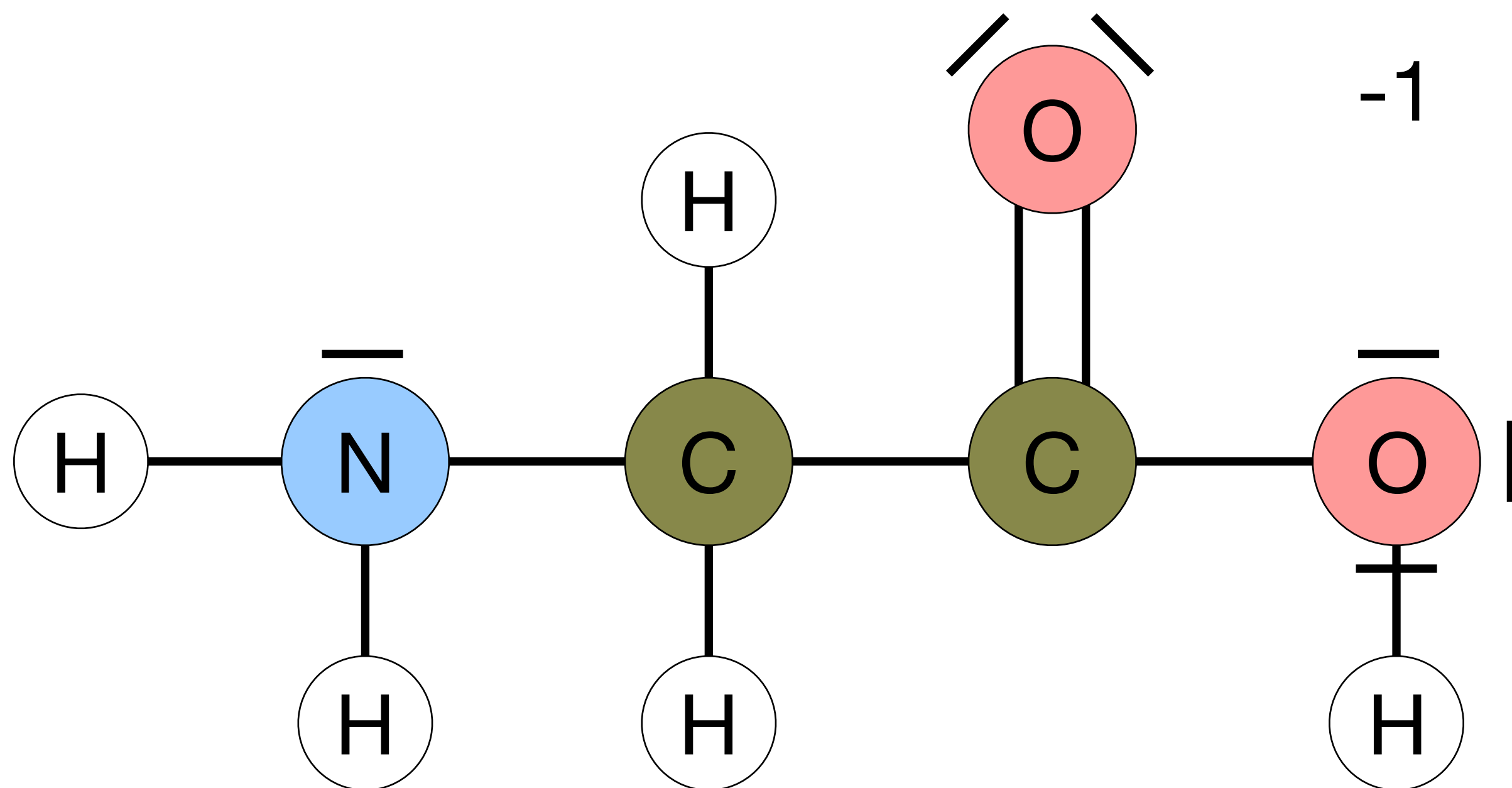


# Octet And 18-Electron Rules

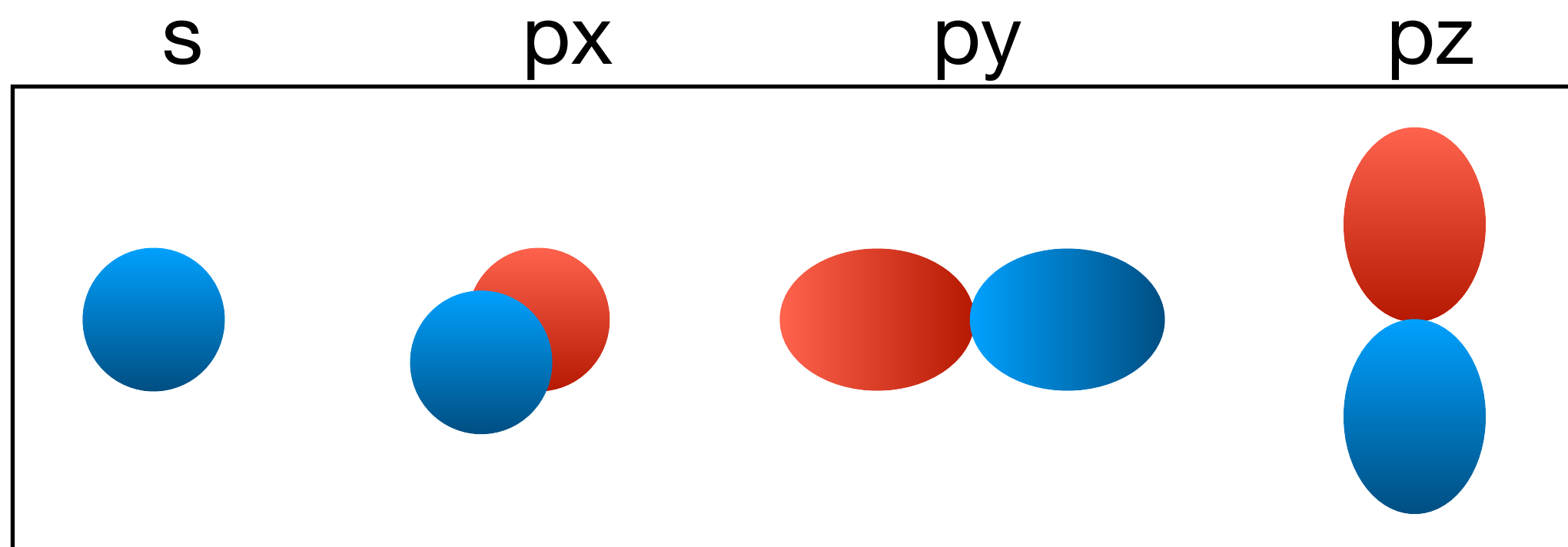
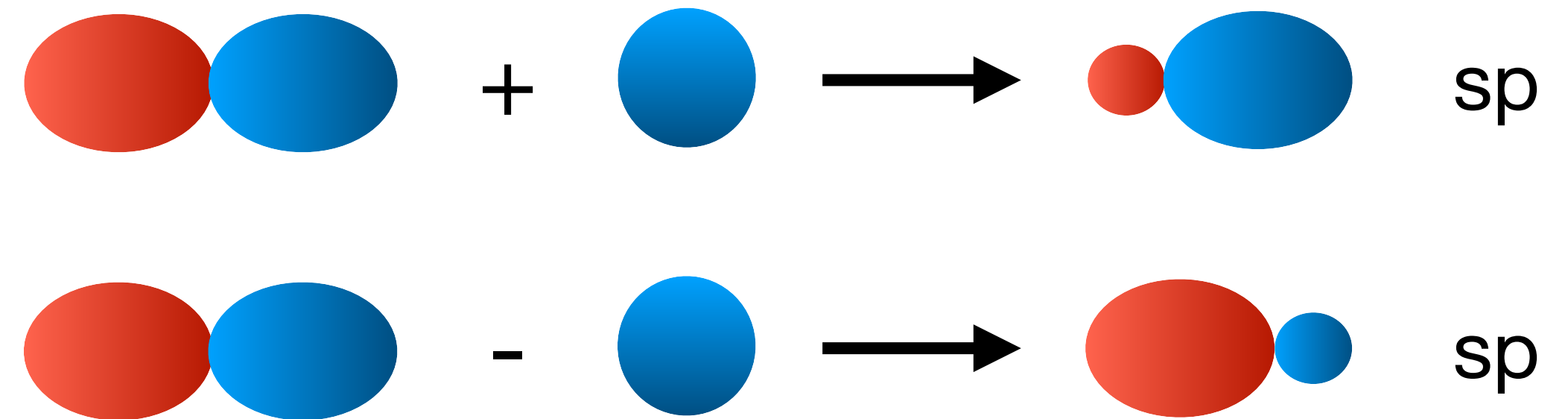
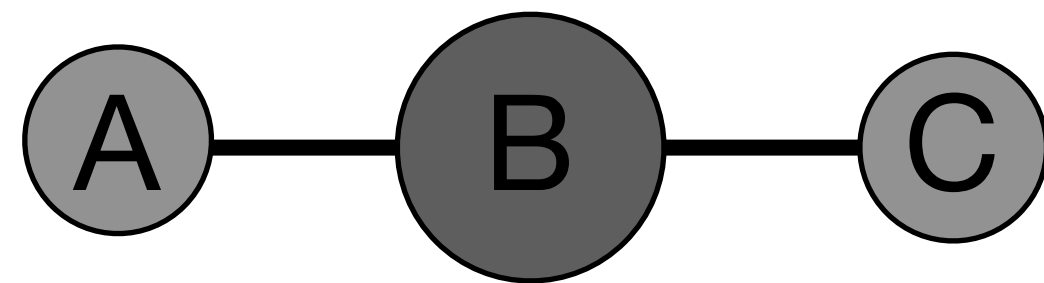
| Group<br>Period | → 1      | 2        | 3           | 4         | 5         | 6         | 7         | 8         | 9         | 10        | 11        | 12        | 13        | 14        | 15        | 16        | 17        | 18        |
|-----------------|----------|----------|-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1               | 1<br>H   |          |             |           |           |           |           |           |           |           |           |           |           |           |           |           |           | 2<br>He   |
| 2               | 3<br>Li  | 4<br>Be  |             |           |           |           |           |           |           |           |           |           | 5<br>B    | 6<br>C    | 7<br>N    | 8<br>O    | 9<br>F    | 10<br>Ne  |
| 3               | 11<br>Na | 12<br>Mg |             |           |           |           |           |           |           |           |           |           | 13<br>Al  | 14<br>Si  | 15<br>P   | 16<br>S   | 17<br>Cl  | 18<br>Ar  |
| 4               | 19<br>K  | 20<br>Ca | 21<br>Sc    | 22<br>Ti  | 23<br>V   | 24<br>Cr  | 25<br>Mn  | 26<br>Fe  | 27<br>Co  | 28<br>Ni  | 29<br>Cu  | 30<br>Zn  | 31<br>Ga  | 32<br>Ge  | 33<br>As  | 34<br>Se  | 35<br>Br  | 36<br>Kr  |
| 5               | 37<br>Rb | 38<br>Sr | 39<br>Y     | 40<br>Zr  | 41<br>Nb  | 42<br>Mo  | 43<br>Tc  | 44<br>Ru  | 45<br>Rh  | 46<br>Pd  | 47<br>Ag  | 48<br>Cd  | 49<br>In  | 50<br>Sn  | 51<br>Sb  | 52<br>Te  | 53<br>I   | 54<br>Xe  |
| 6               | 55<br>Cs | 56<br>Ba | * 71<br>Lu  | 72<br>Hf  | 73<br>Ta  | 74<br>W   | 75<br>Re  | 76<br>Os  | 77<br>Ir  | 78<br>Pt  | 79<br>Au  | 80<br>Hg  | 81<br>Tl  | 82<br>Pb  | 83<br>Bi  | 84<br>Po  | 85<br>At  | 86<br>Rn  |
| 7               | 87<br>Fr | 88<br>Ra | * 103<br>Lr | 104<br>Rf | 105<br>Db | 106<br>Sg | 107<br>Bh | 108<br>Hs | 109<br>Mt | 110<br>Ds | 111<br>Rg | 112<br>Cn | 113<br>Nh | 114<br>Fl | 115<br>Mc | 116<br>Lv | 117<br>Ts | 118<br>Og |
|                 |          |          | * 57<br>La  | 58<br>Ce  | 59<br>Pr  | 60<br>Nd  | 61<br>Pm  | 62<br>Sm  | 63<br>Eu  | 64<br>Gd  | 65<br>Tb  | 66<br>Dy  | 67<br>Ho  | 68<br>Er  | 69<br>Tm  | 70<br>Yb  |           |           |
|                 |          |          | * 89<br>Ac  | 90<br>Th  | 91<br>Pa  | 92<br>U   | 93<br>Np  | 94<br>Pu  | 95<br>Am  | 96<br>Cm  | 97<br>Bk  | 98<br>Cf  | 99<br>Es  | 100<br>Fm | 101<br>Md | 102<br>No |           |           |



# Lewis Structures

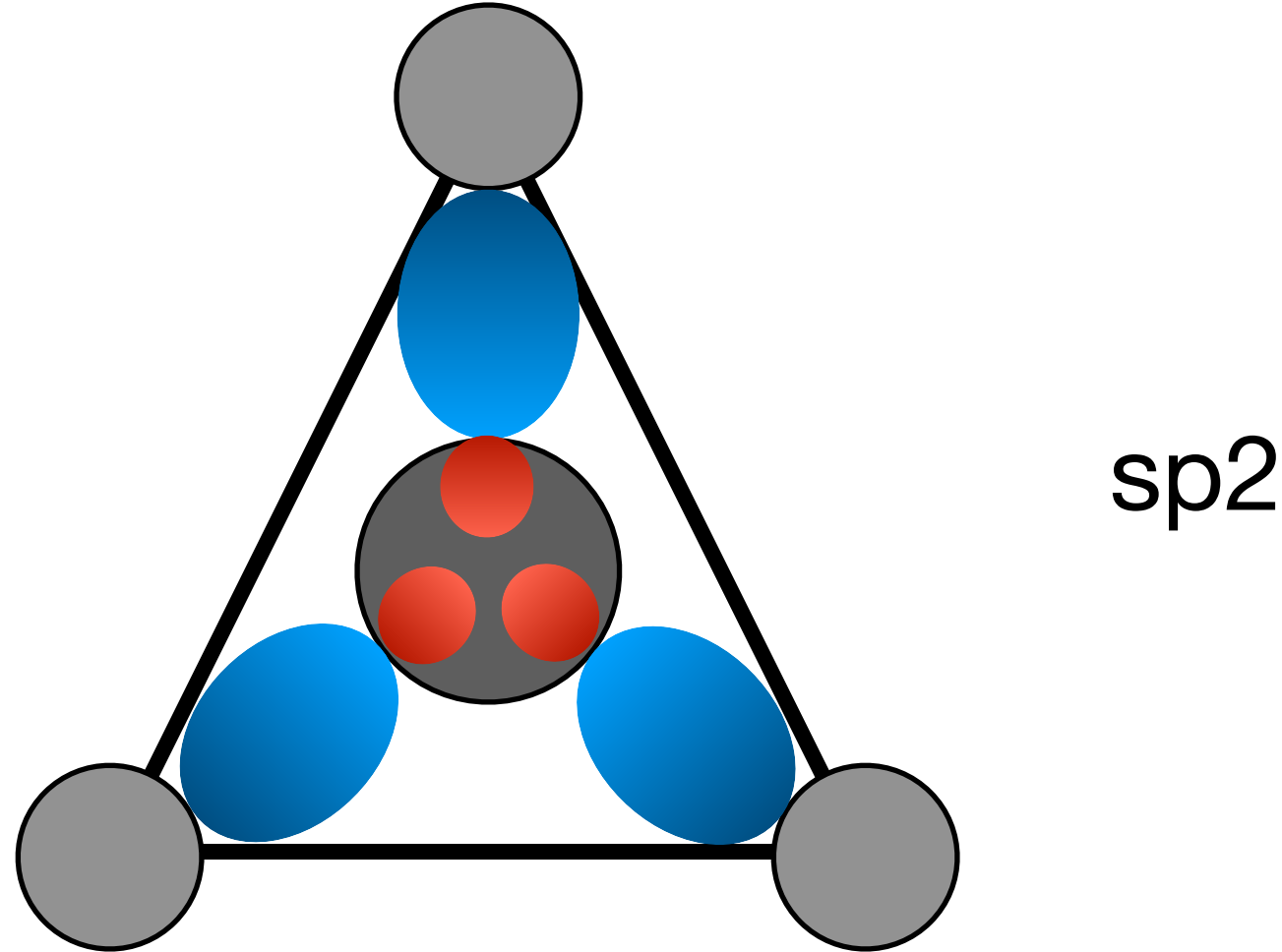


# Orbital Hybridization

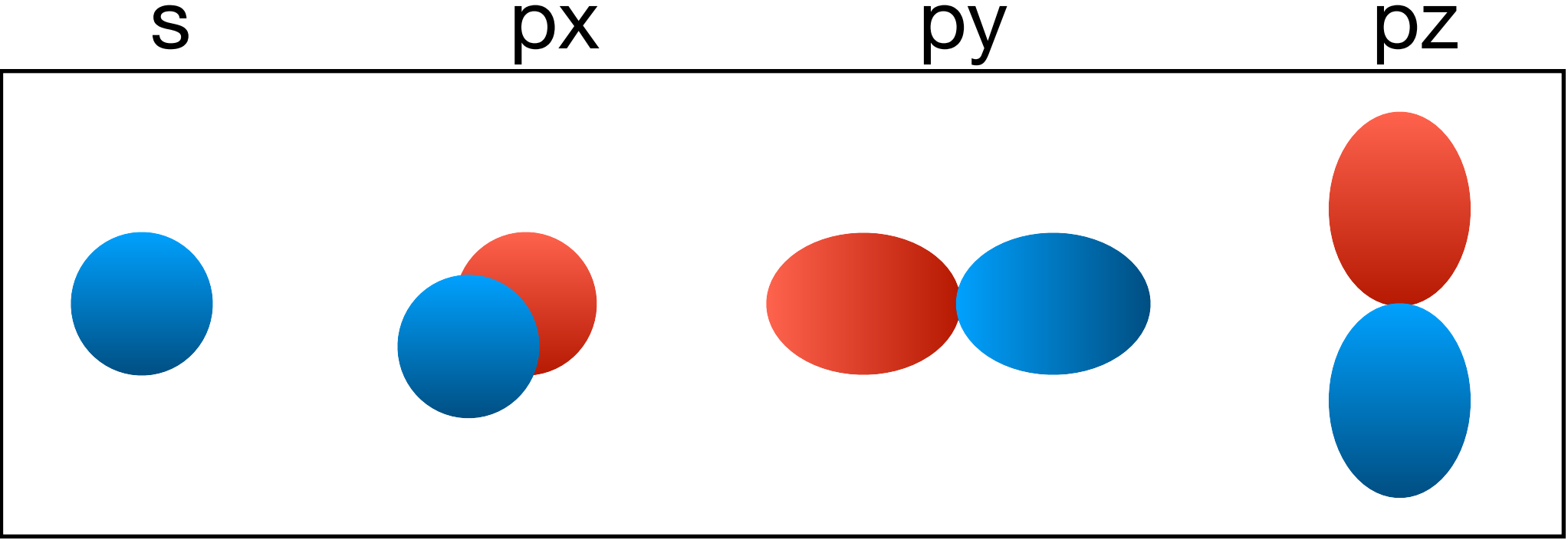
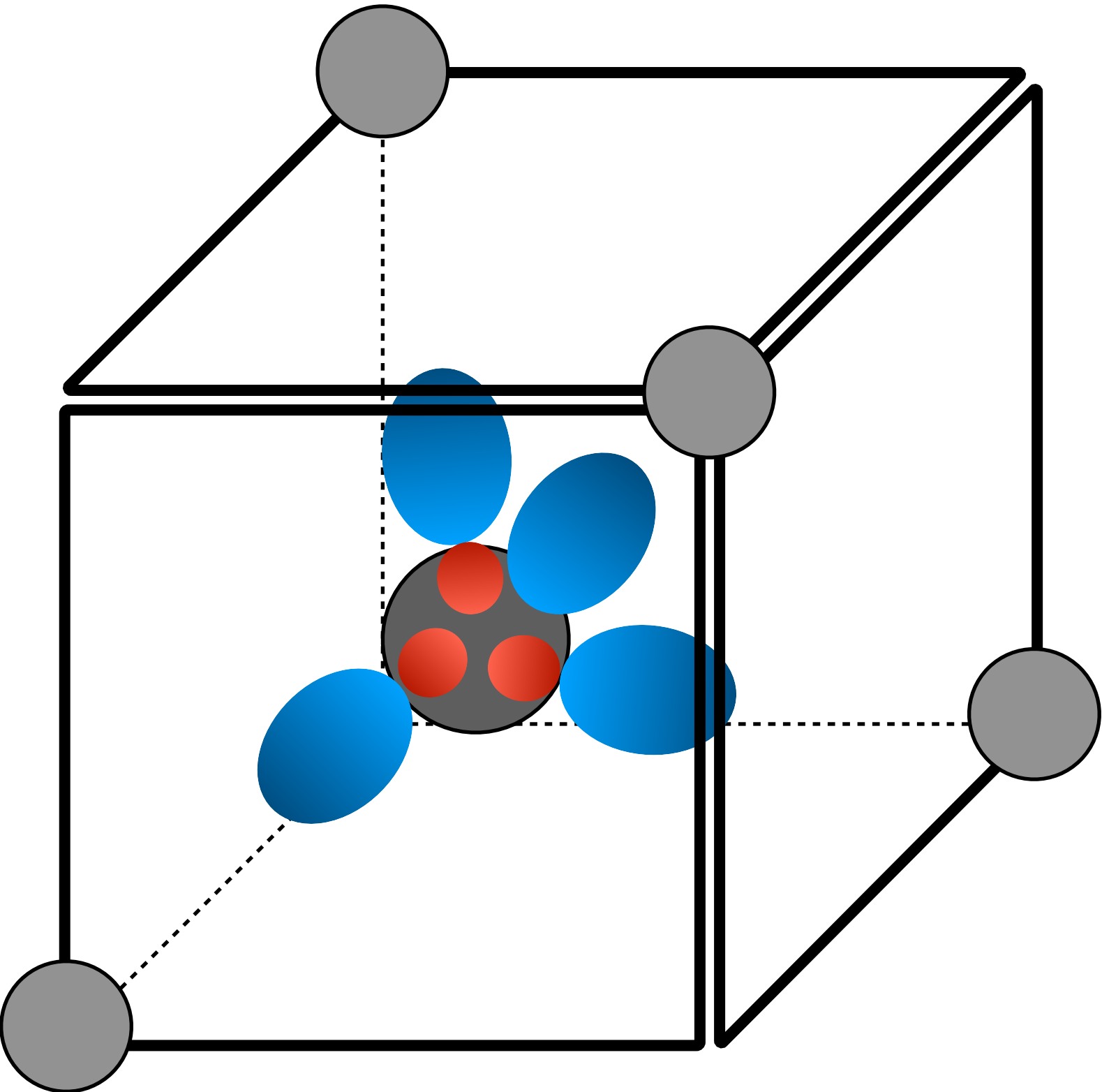




# Orbital Hybridization



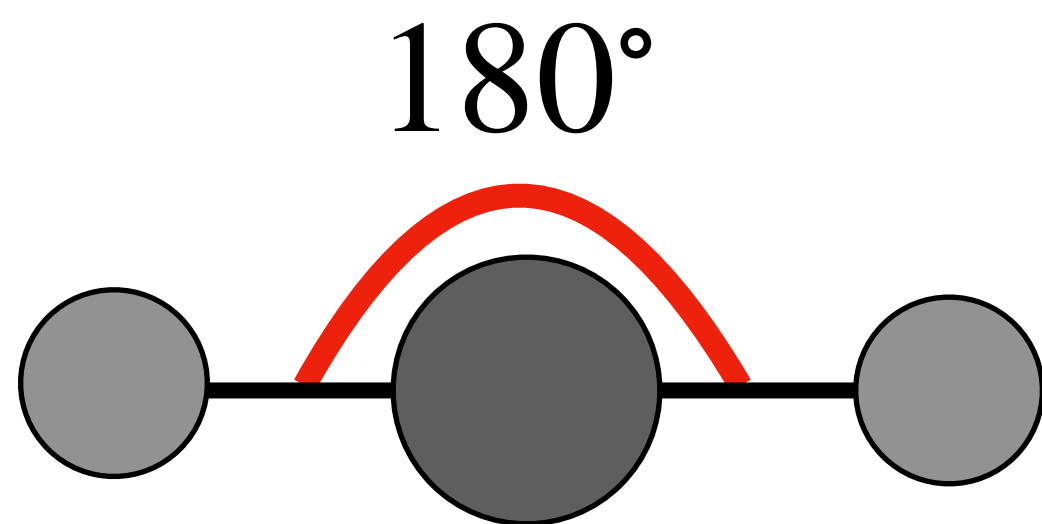
$sp^3$



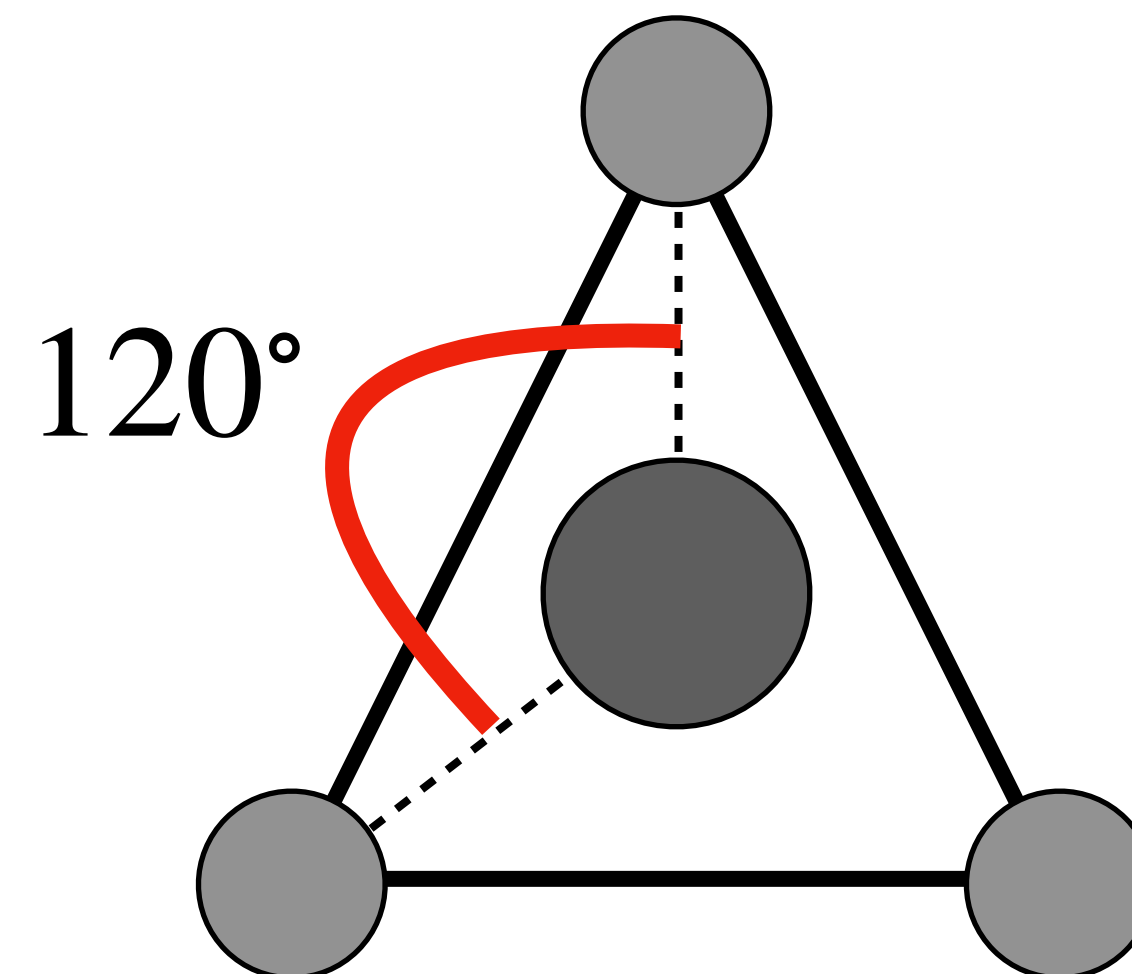


# Orbital Hybridization

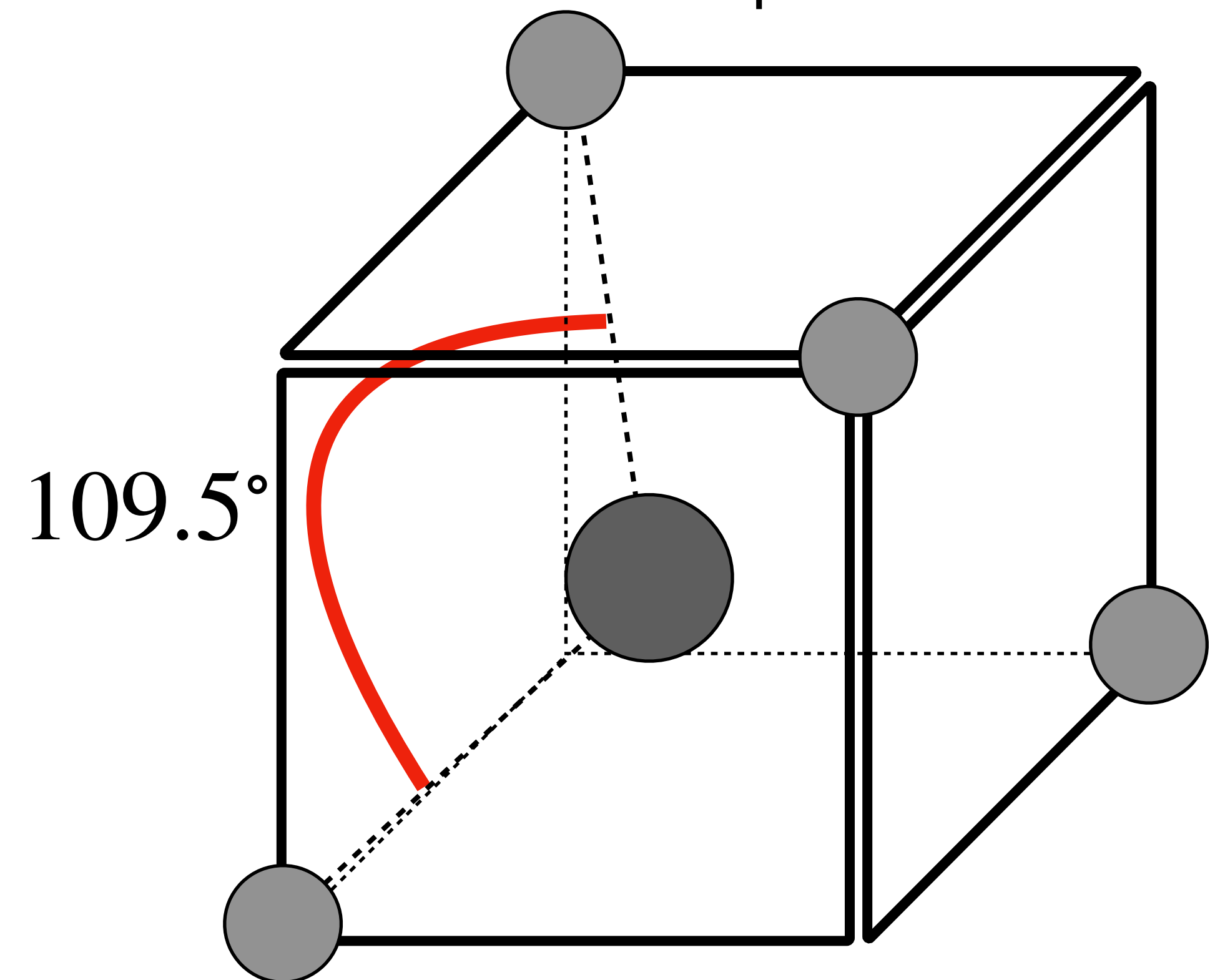
sp



sp<sup>2</sup>

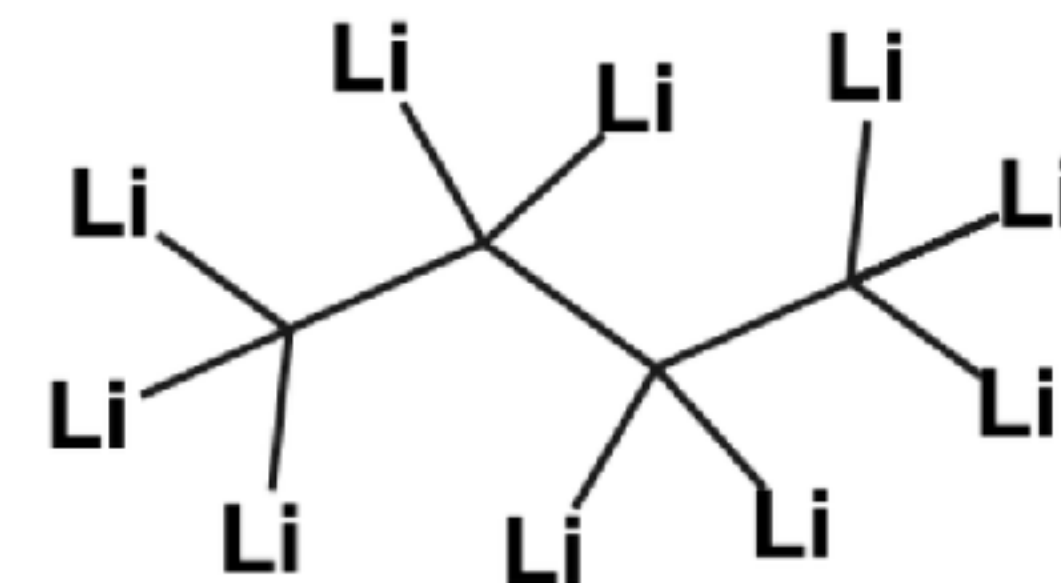
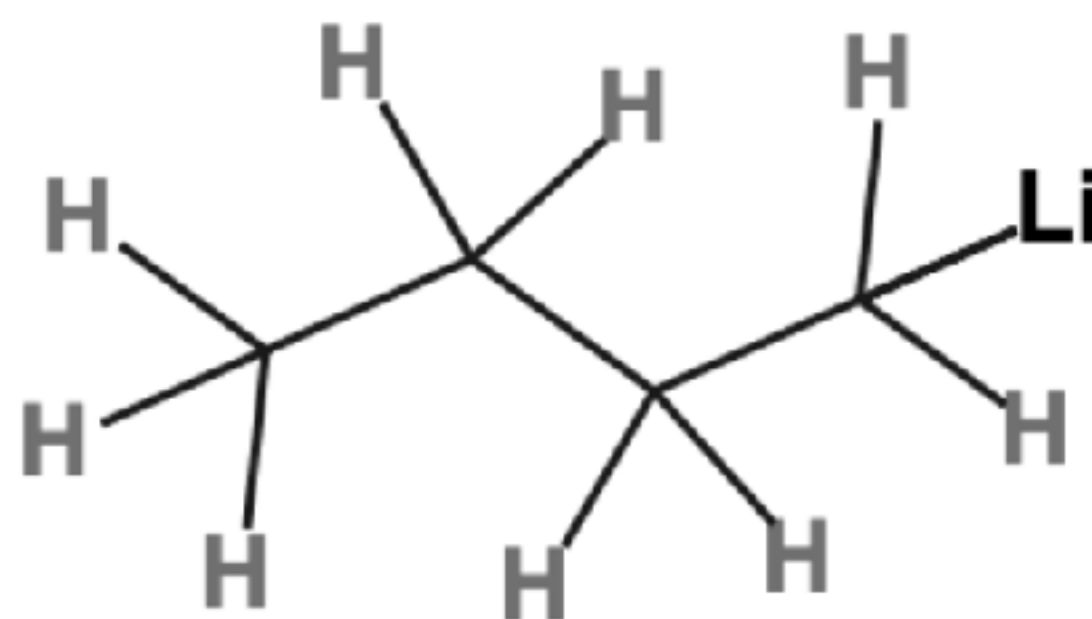
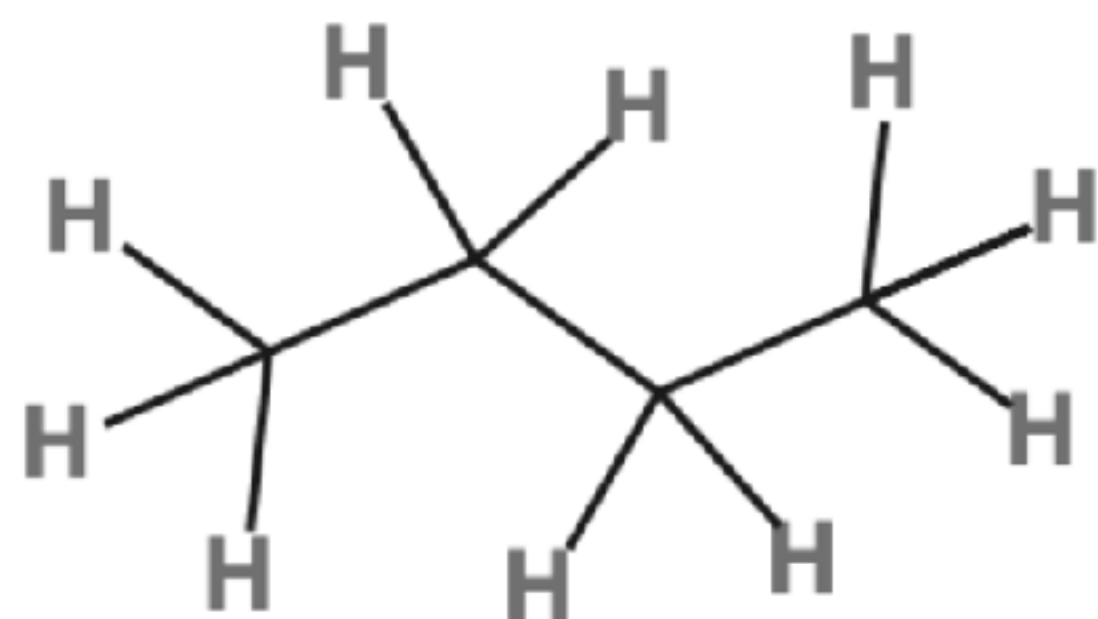


sp<sup>3</sup>



# To Make It Complicated:

Atoms are picky with whom they hang out



[Link](#)

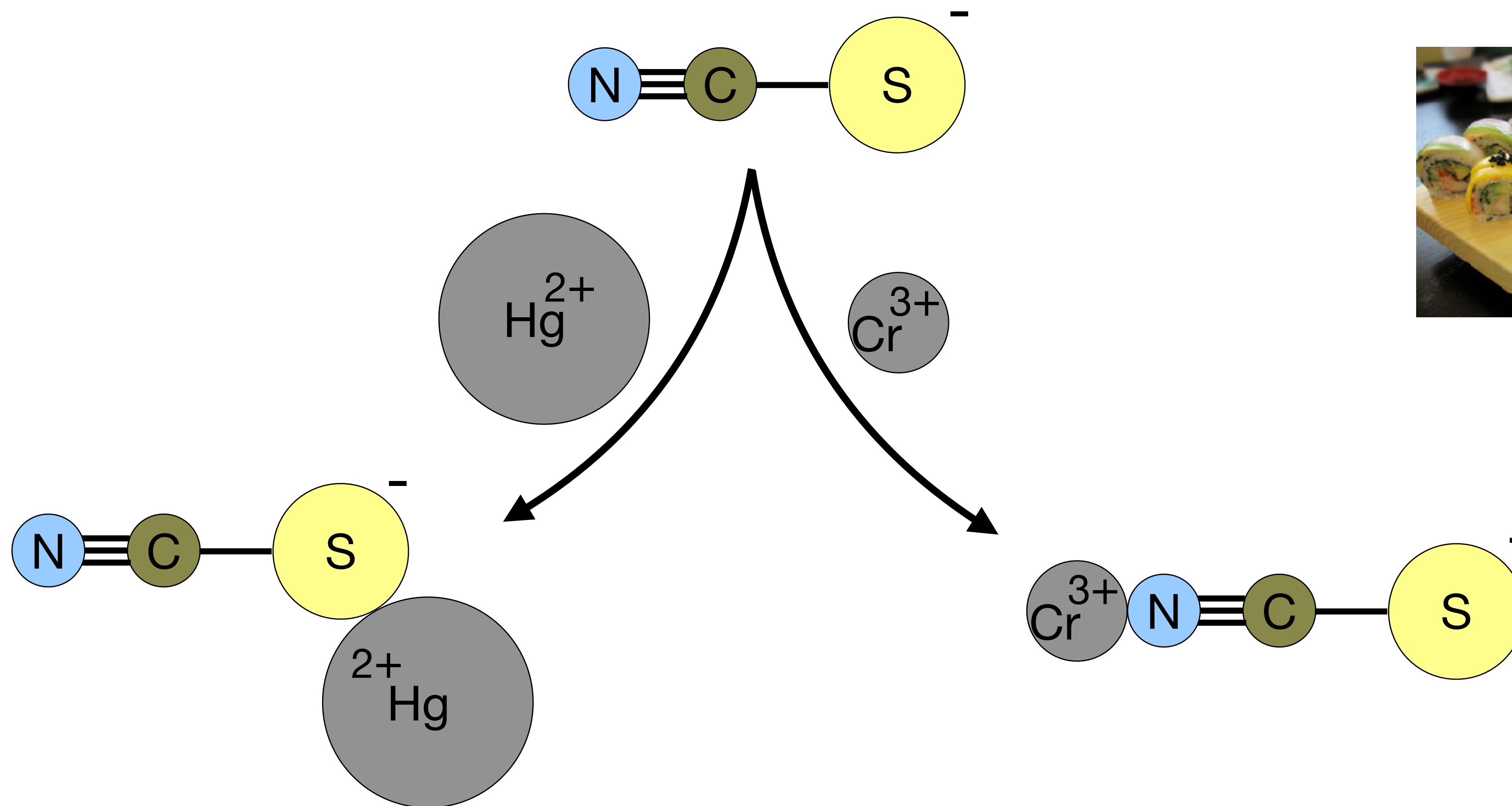


[Link](#)



# Hard-Soft Acid-Base

Atoms are picky with whom they hang out



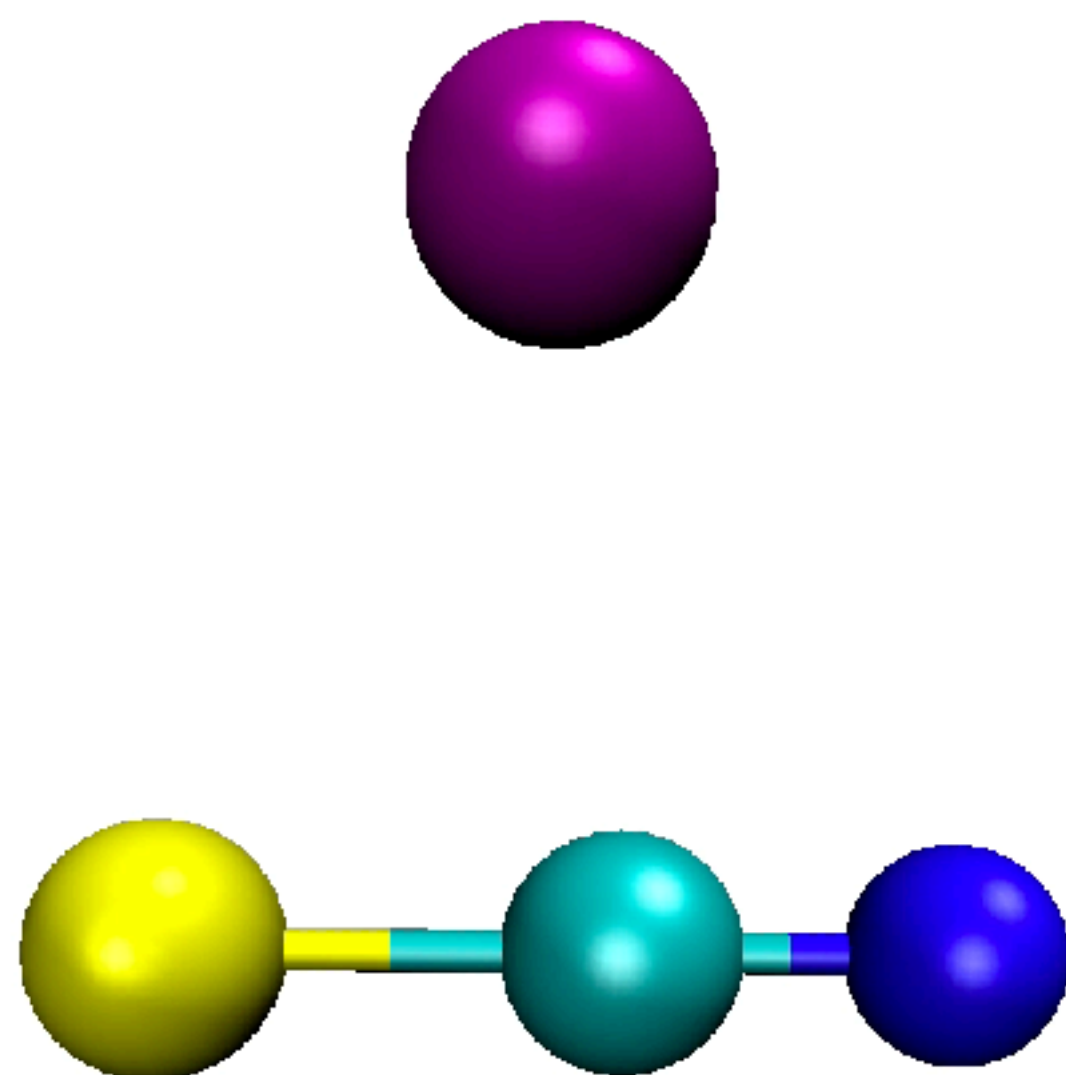
[Link](#)



# Hard-Soft Acid-Base

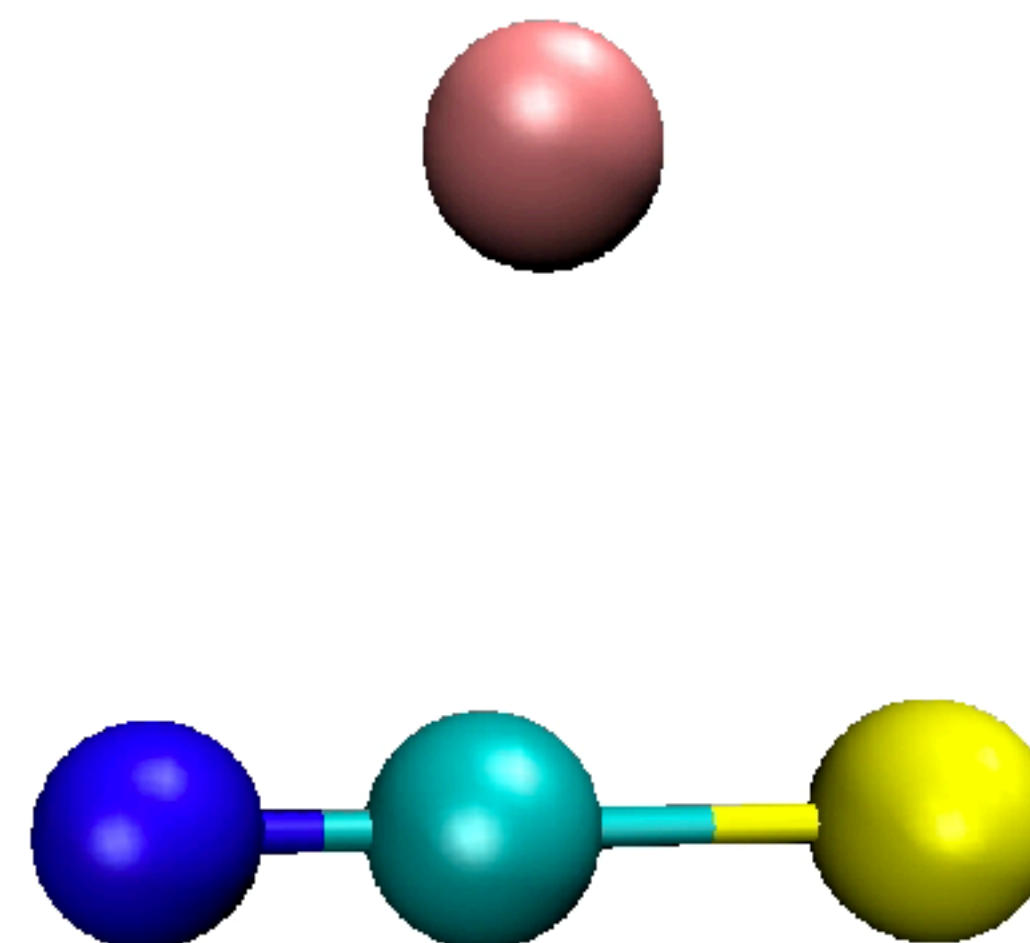
Atoms are picky with whom they hang out

---



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

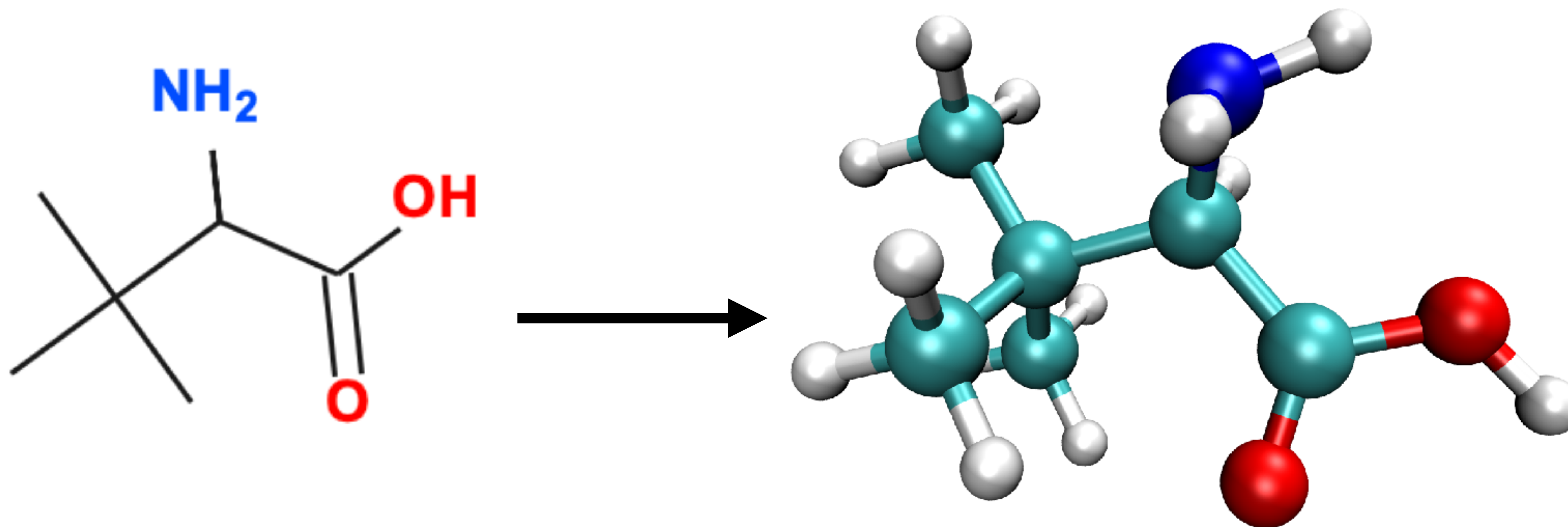
$$(s_S | s_{Hg}) = 0.22$$



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

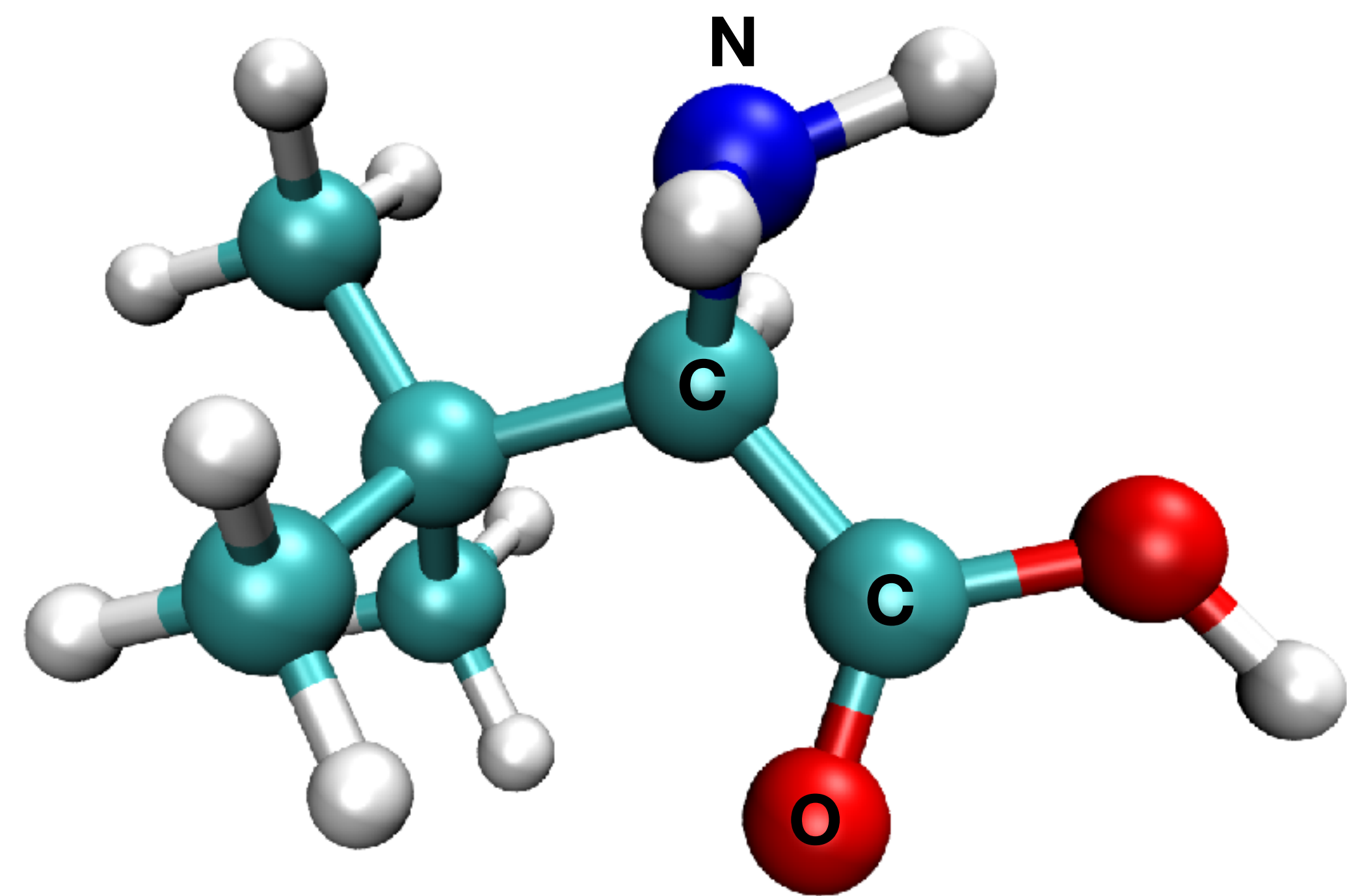
$$(s_S | s_{Cr}) = -$$

# From 2D to 3D



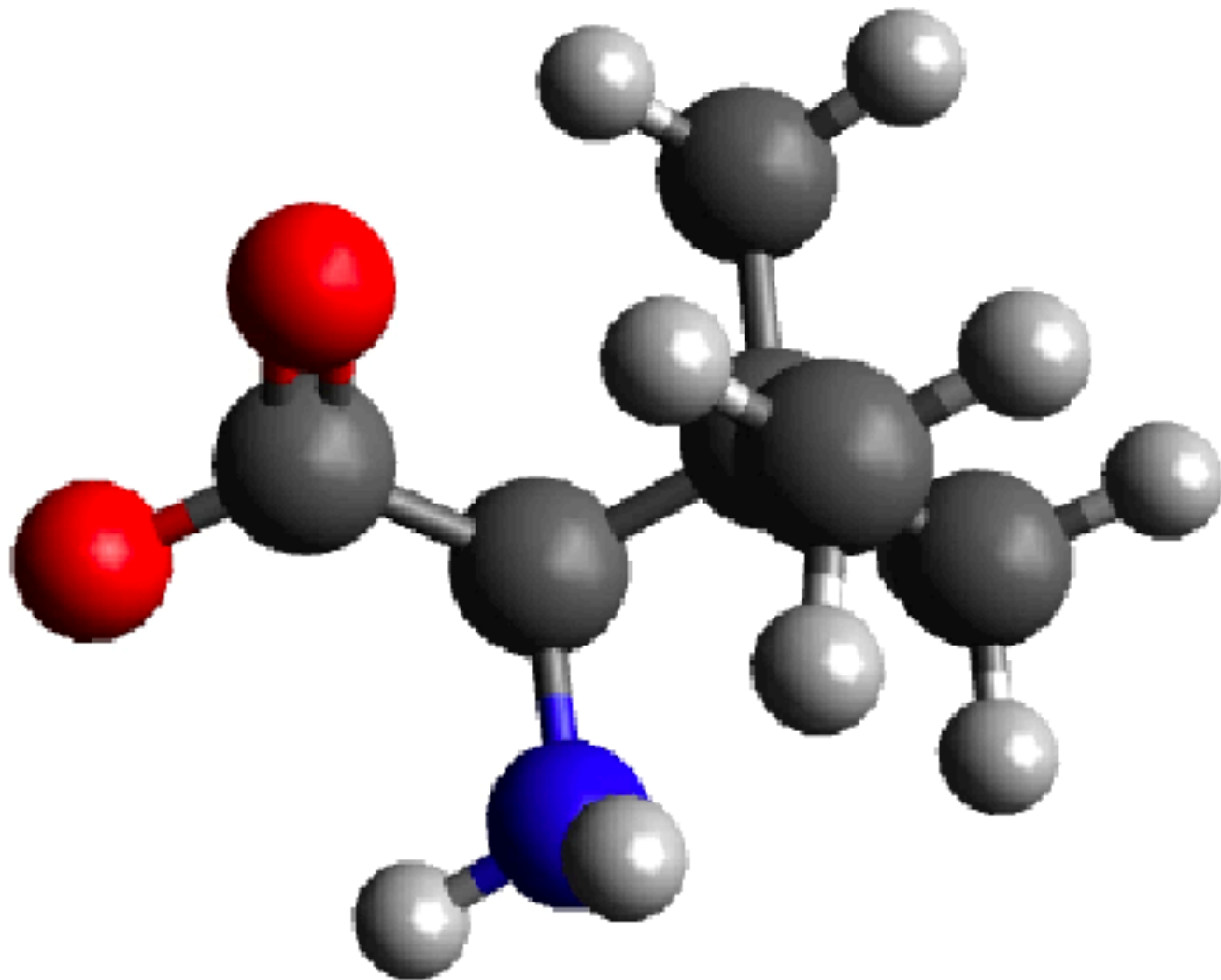


# Z-Matrix



| 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



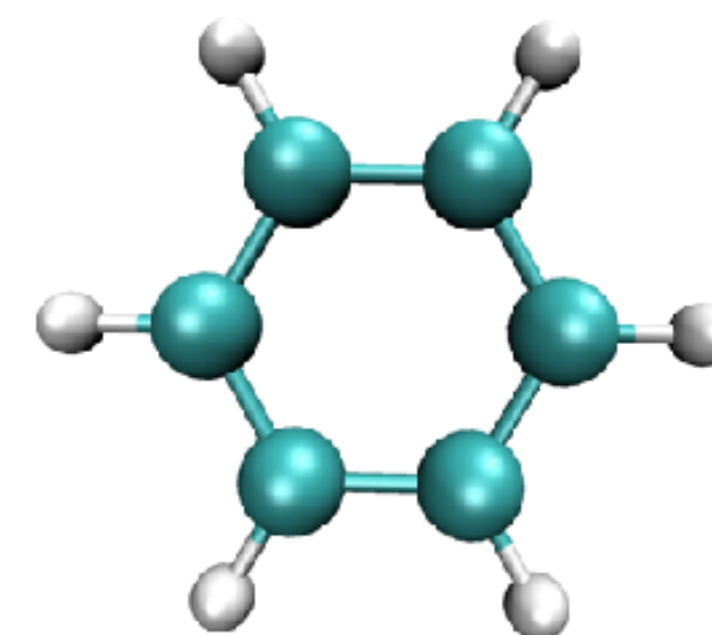
Use rules above to connect atoms



(use fragments to simplify)



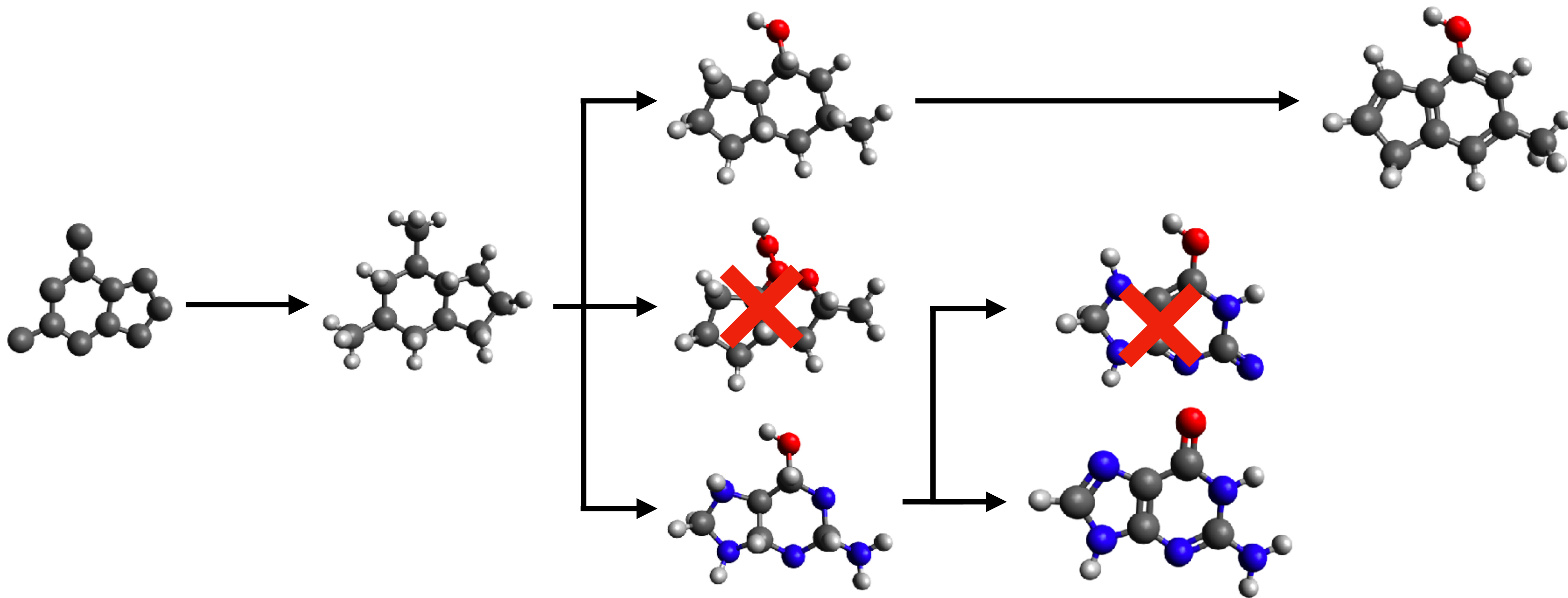
Optimize the Geometry (FF or QM)





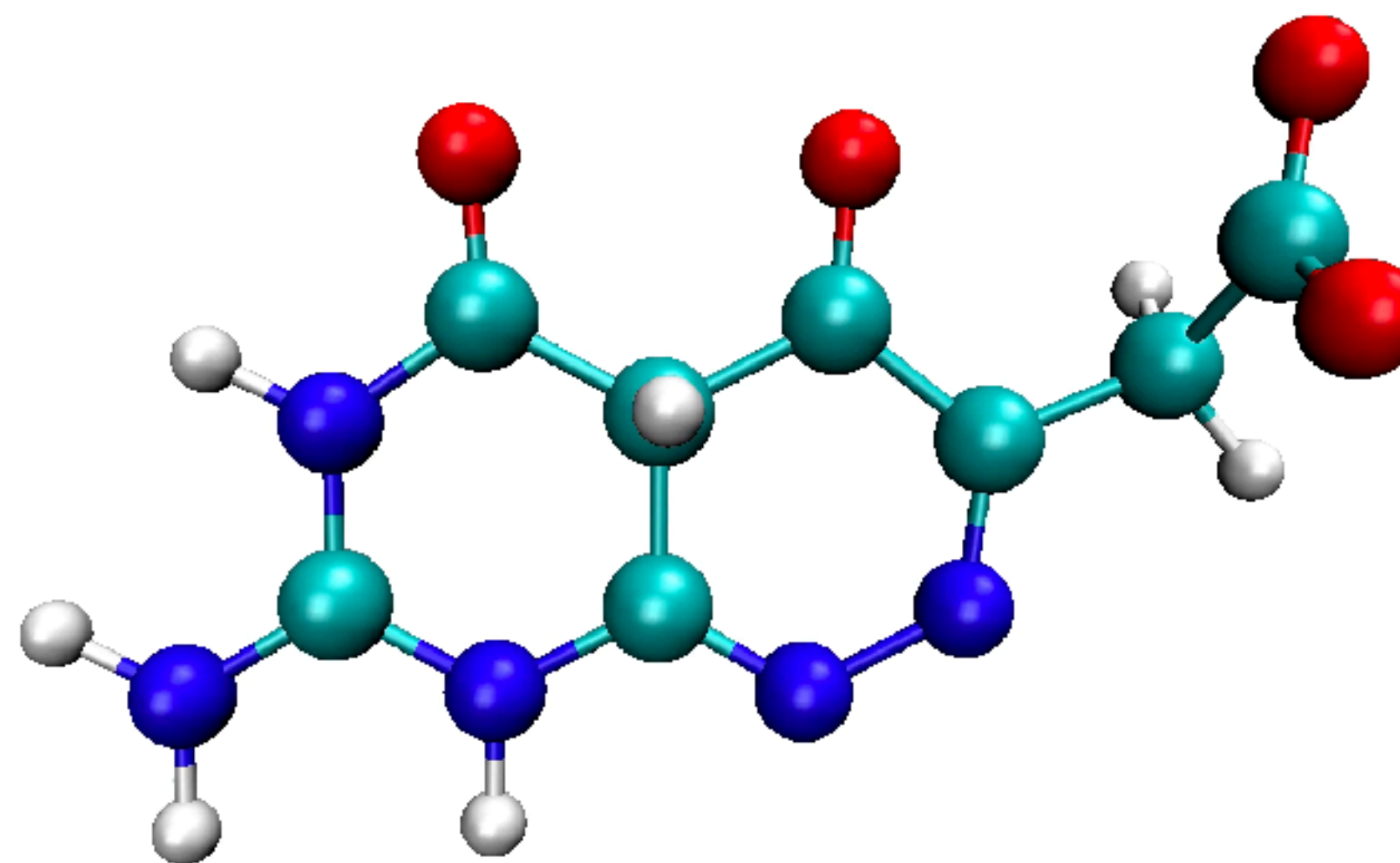
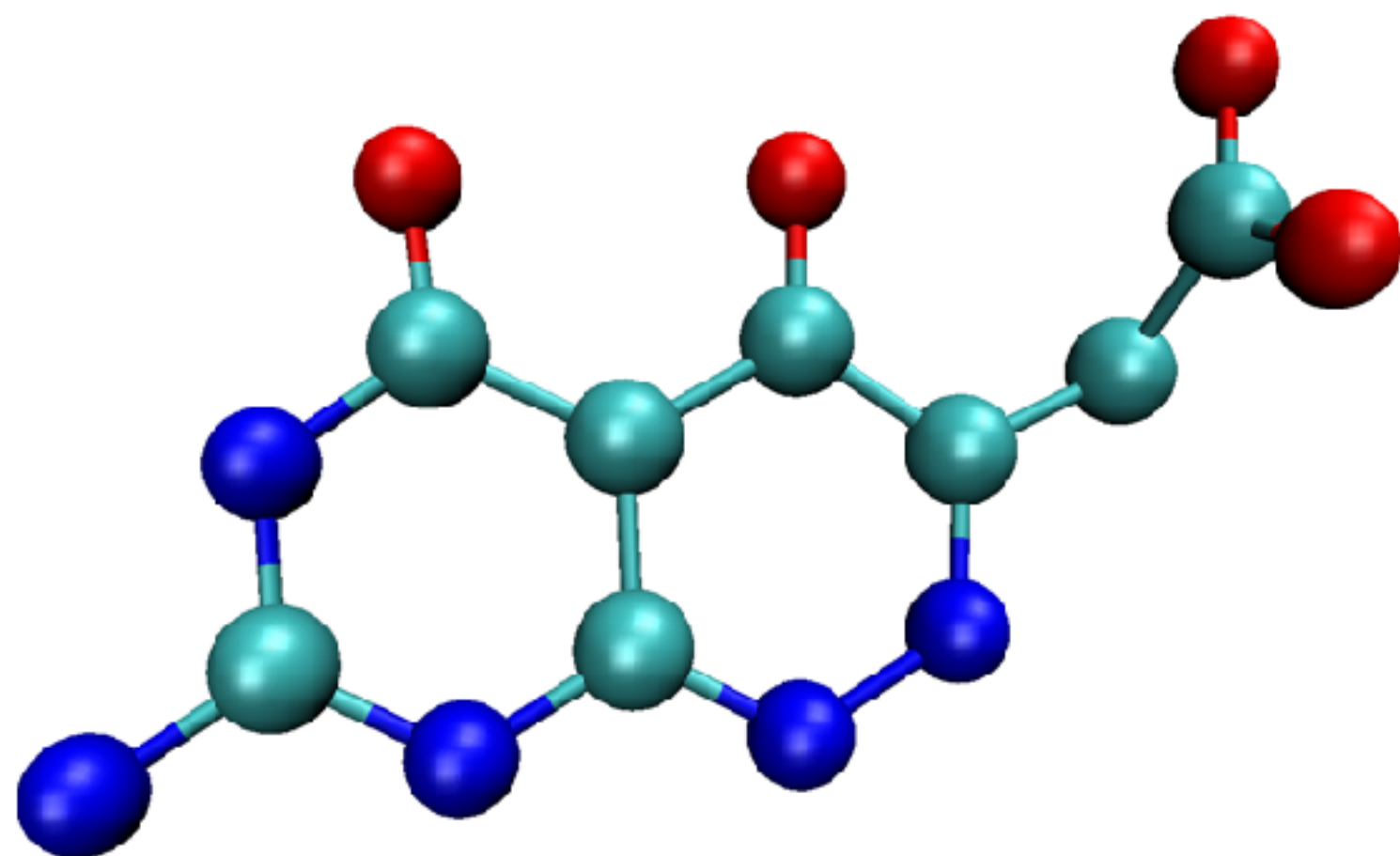
# Automated 3D-Structure Generation

QUANTUM-MACHINE.ORG

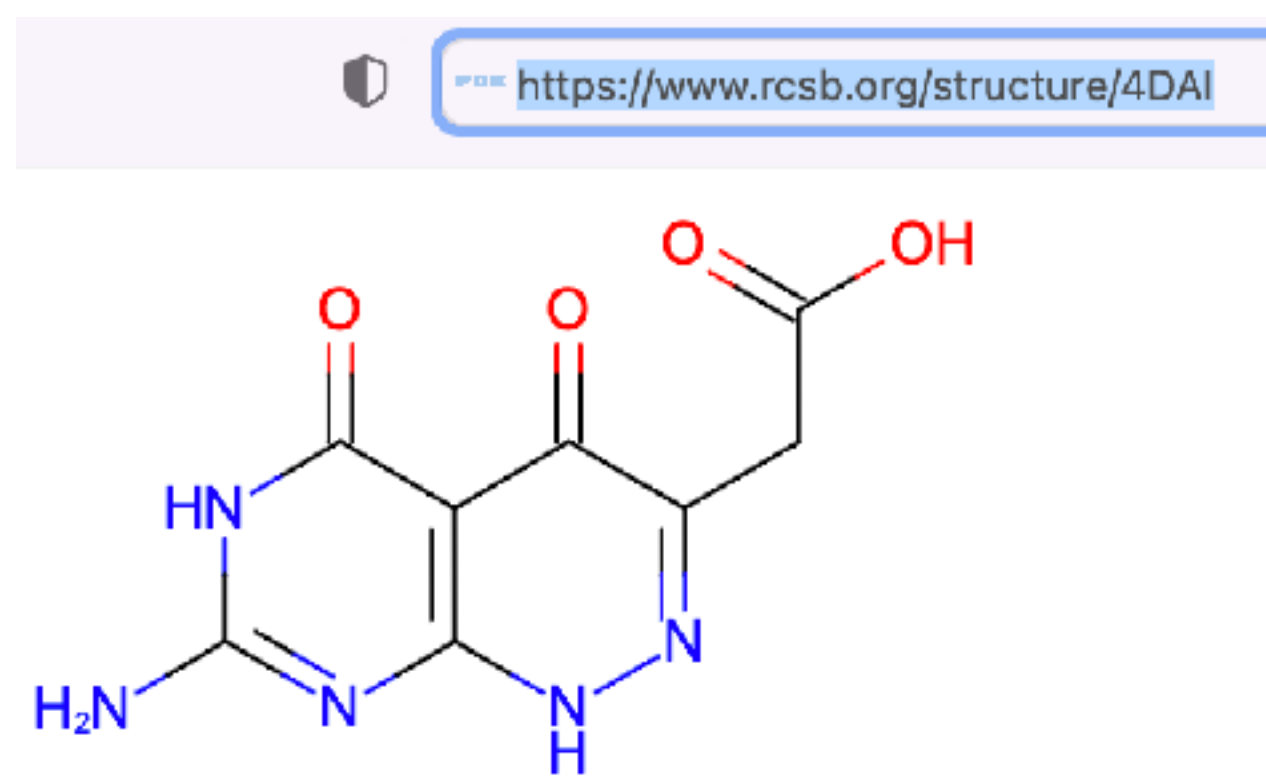




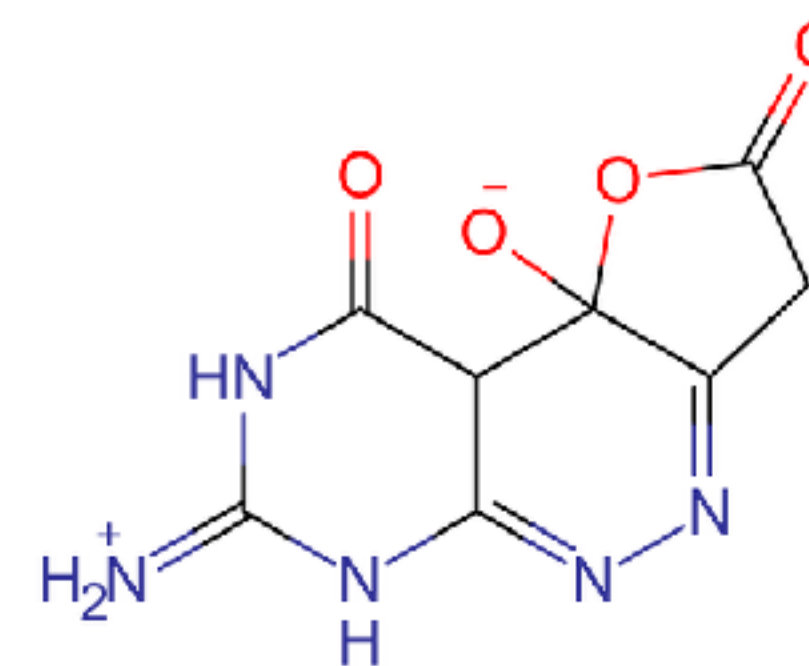
# What can go wrong... will go wrong



PDBbind

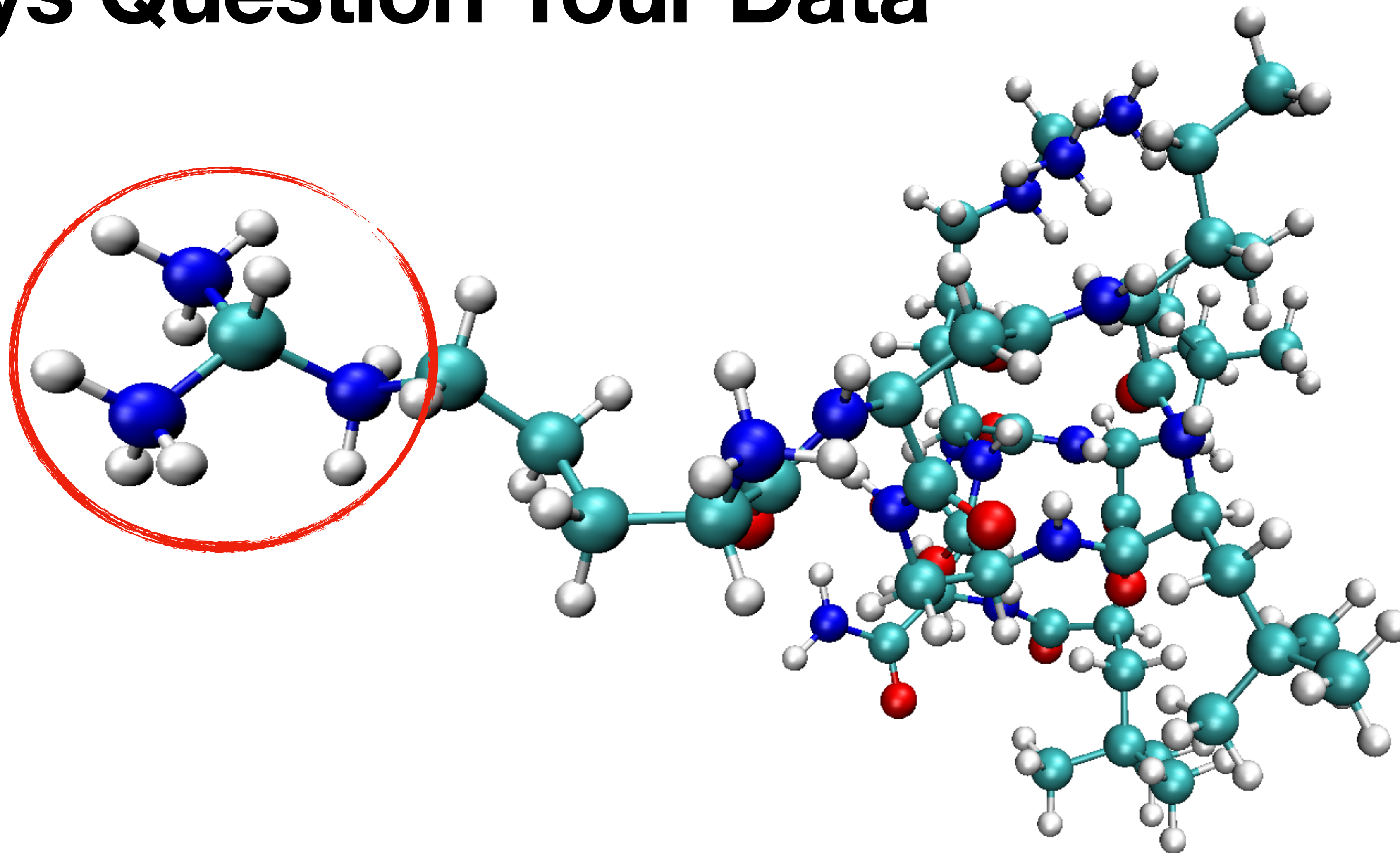


4DAI

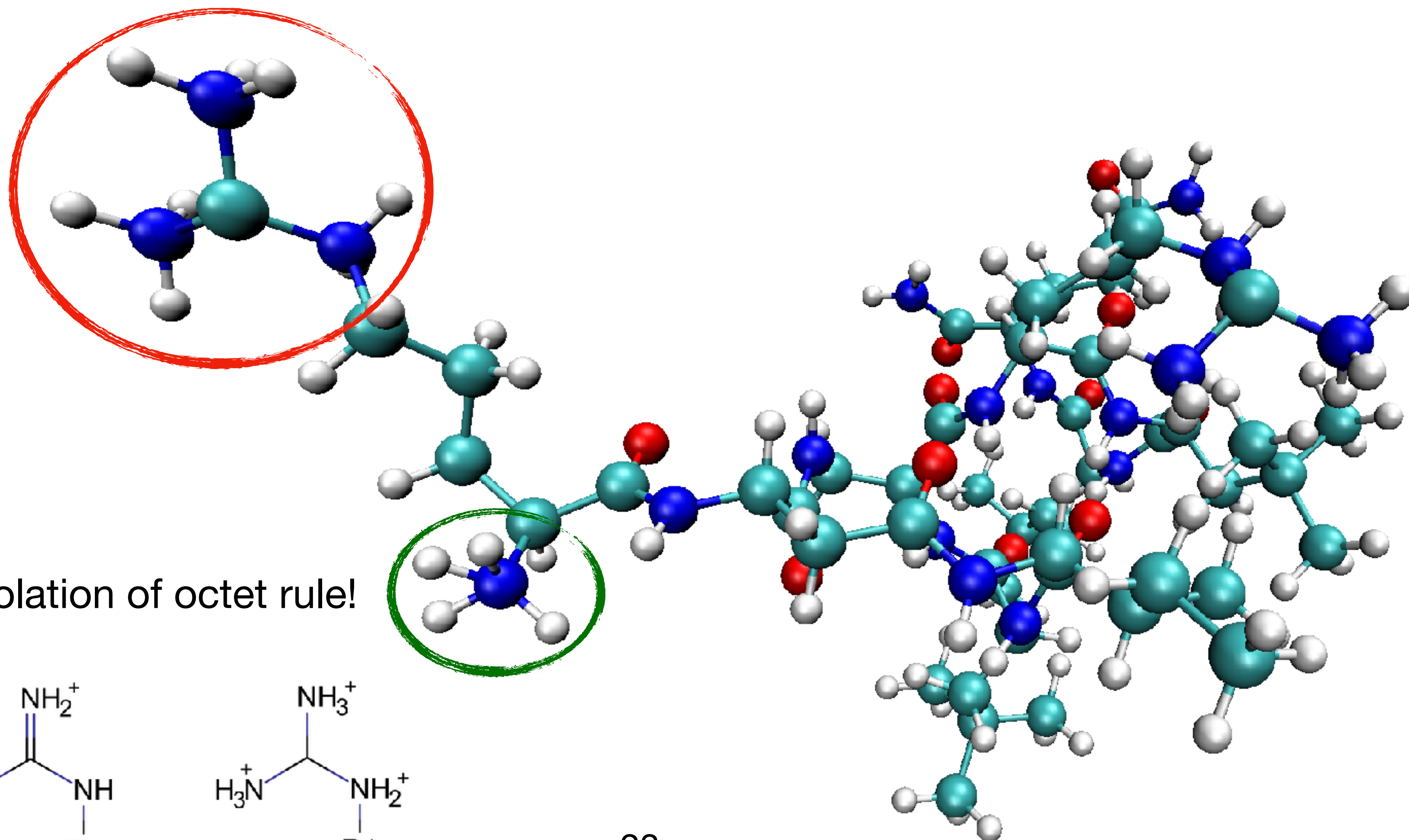




# Always Question Your Data

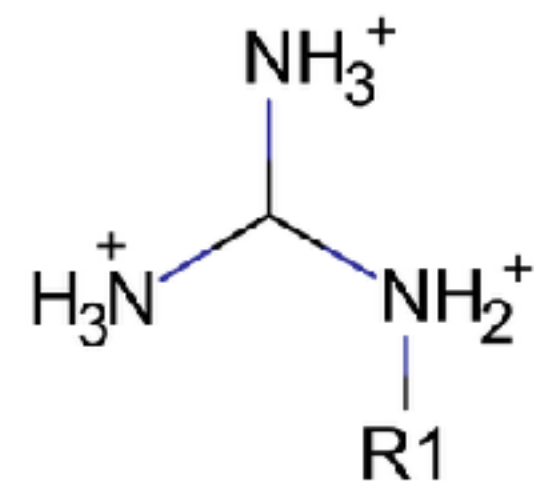
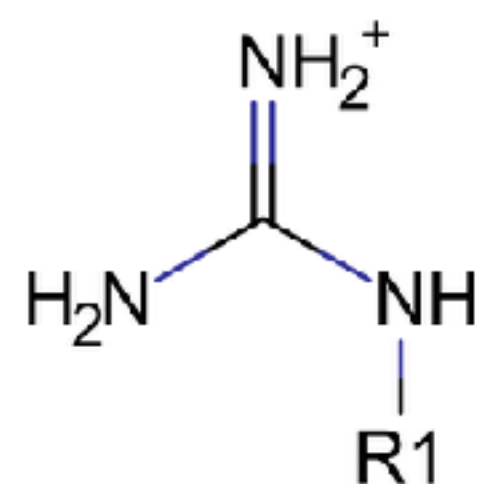


# So fix it with widely used software



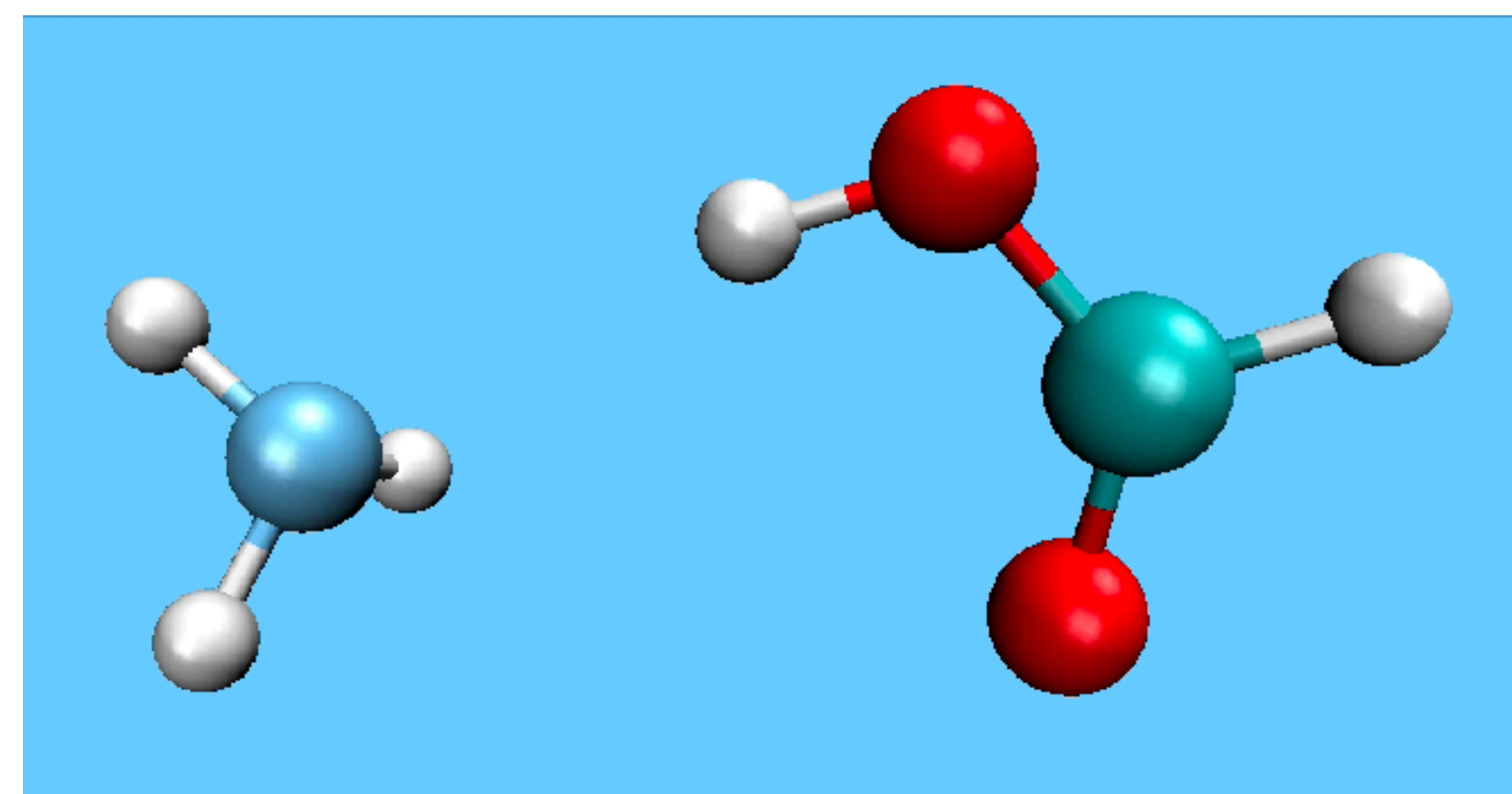
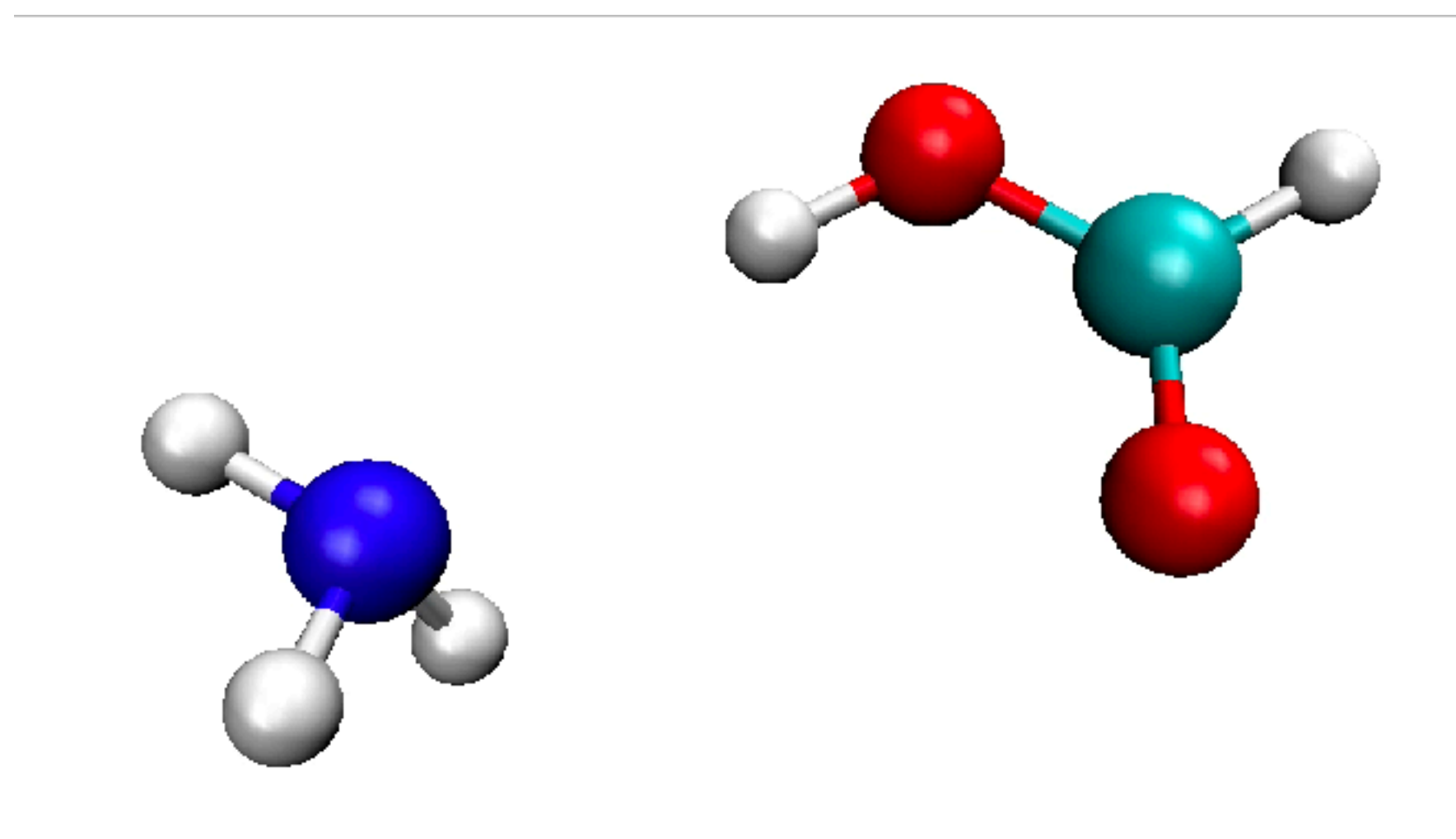
Explicit violation of octet rule!

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 website interface. It features a search form with fields for Compound Name, Molecular Formula, Molecular Weight, CAS Registry No., and SDBS No. There are also sections for Atoms (C, H, N, O, F, Cl, Br, I, S, P, Si) and Spectrum (MS, IR, <sup>13</sup>C NMR, <sup>1</sup>H NMR, ESR, Raman). The IR Peaks section includes a range and transmittance filter. The <sup>13</sup>C NMR and <sup>1</sup>H NMR sections include shift ranges and peak lists. A search button and a 'Hit' indicator are at the bottom.

**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:**

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

Numbers between left and right columns.

**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-

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

**<sup>1</sup>H NMR Shift(ppm):**  Allowance ± 0.2  
"/" is the separator for multiple shifts, eg. 110.78,...

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

**Search** **Clear** **Hit:** 20Hit **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 PubChem website interface. It features a search bar and a navigation menu with links to About, Blog, Submit, and Contact. The main section is titled 'Explore Chemistry' and includes a search bar and a list of search results. The bottom section shows statistics for Compounds, Substances, Bioactivities, and 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

**Draw Structure** **Upload ID List** **Browse Data** **Periodic Table**

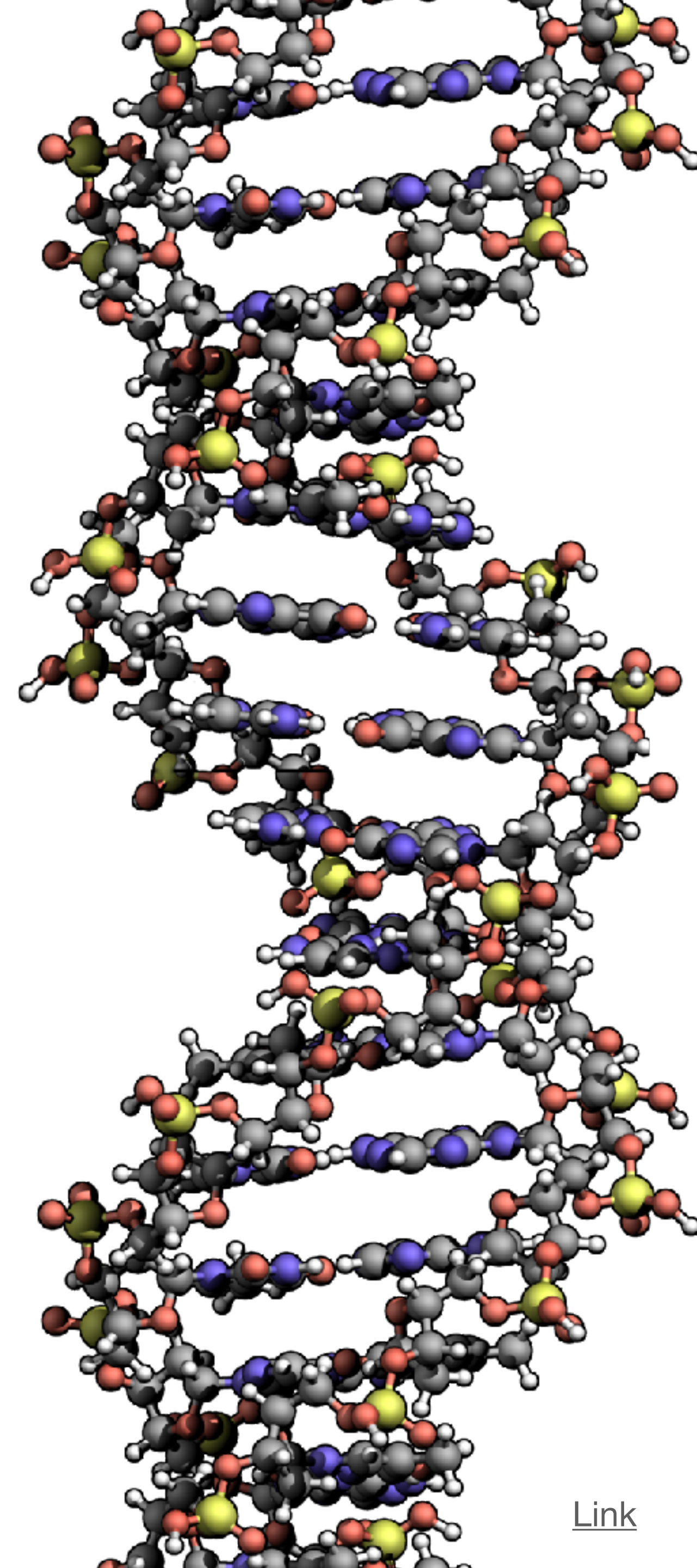
**110M** Compounds **277M** Substances **293M** Bioactivities **843** Data Sources

[Link](#)



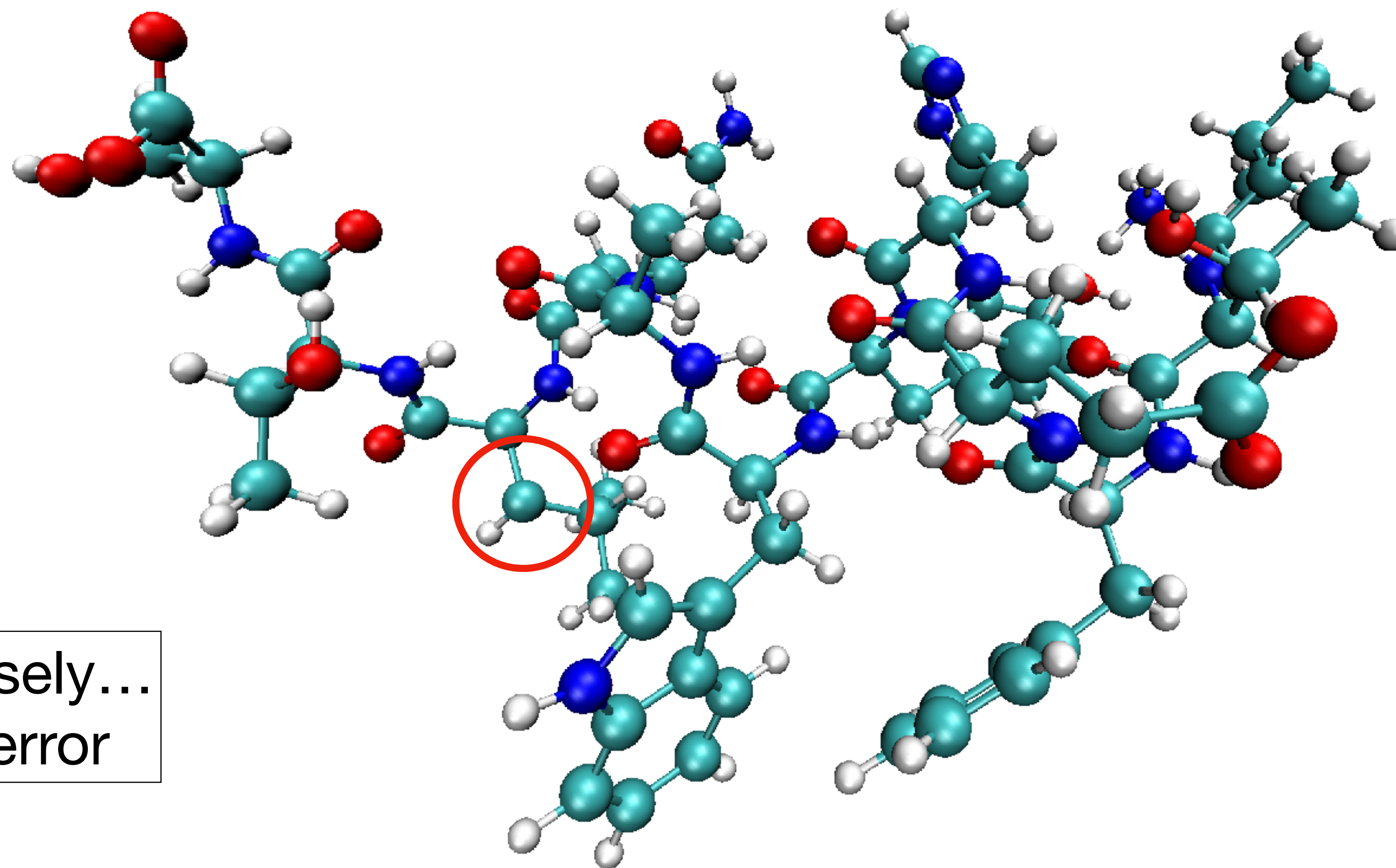
# Part 2

## Structure Refinement





# How to find problems in structures 101

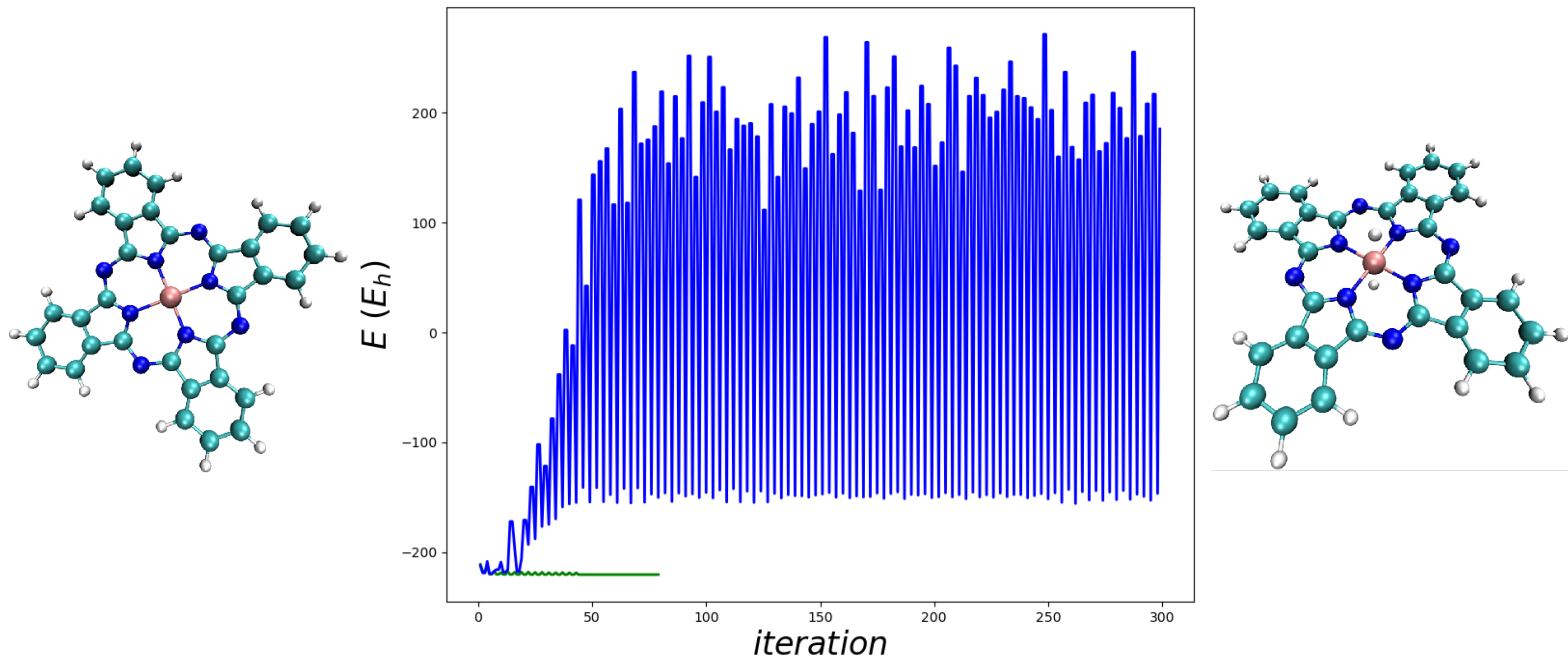


You look very closely...  
and you find the error

# If You don't Know, Ask Physics

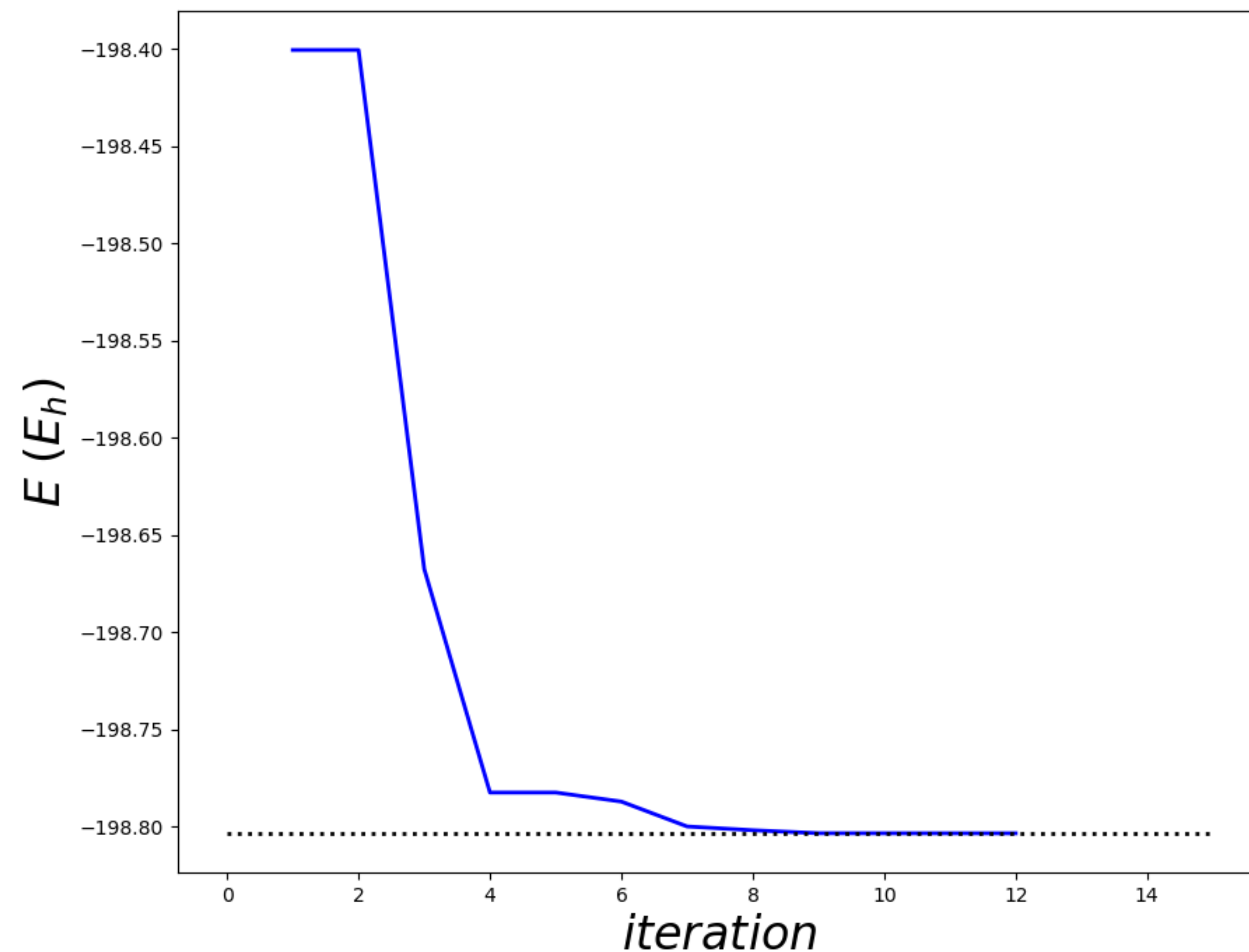
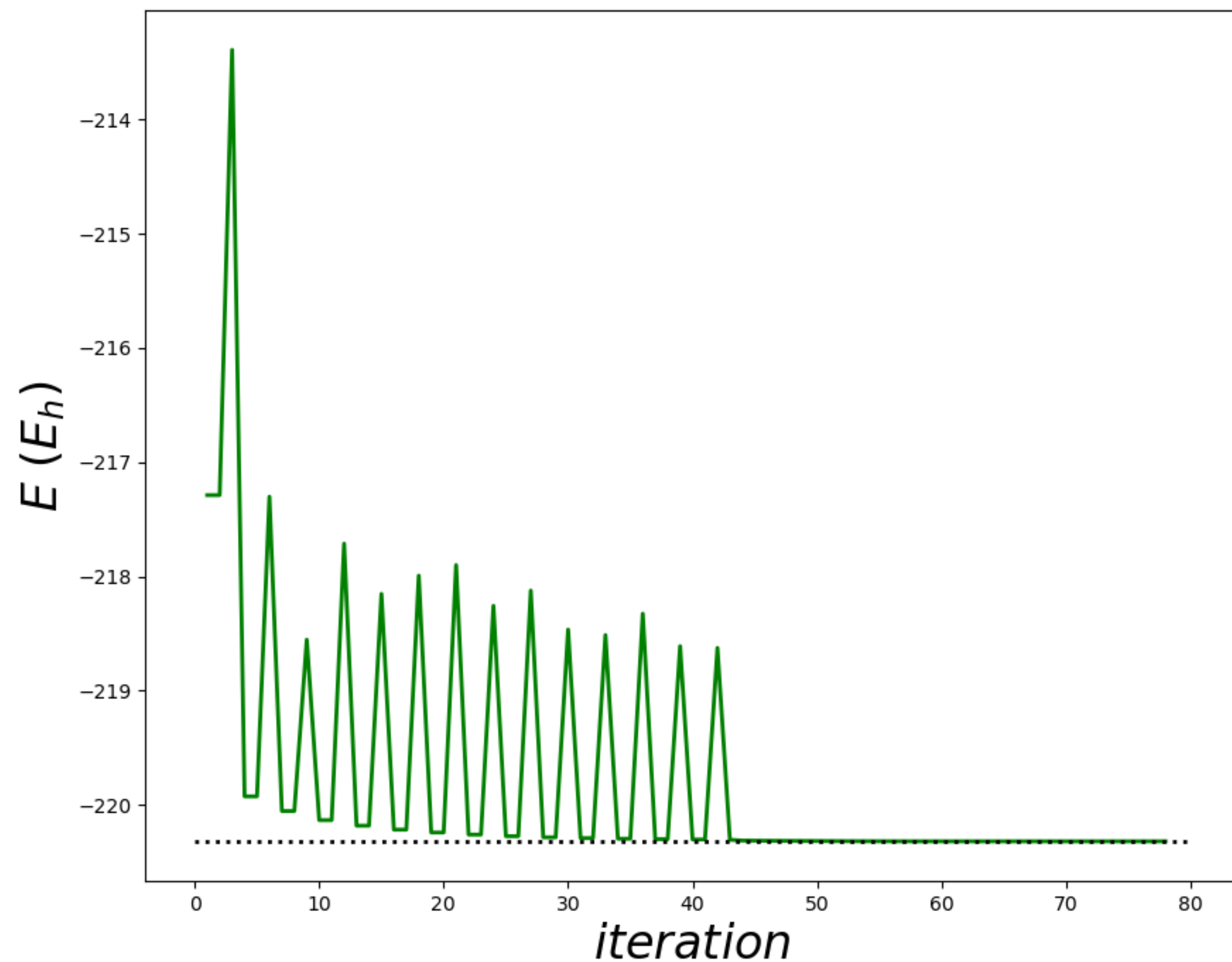
- Physically wrong structures (and electron counting) affects:
  - Force Field Parametrization (fails).
  - Quantum Mechanics convergence (zero to none).
- But also good structures have problems:
  - Dancing Charge problem.
- Population Analysis.
- Geometry Optimization or Dynamics.

# Example of Convergence of Wavefunction



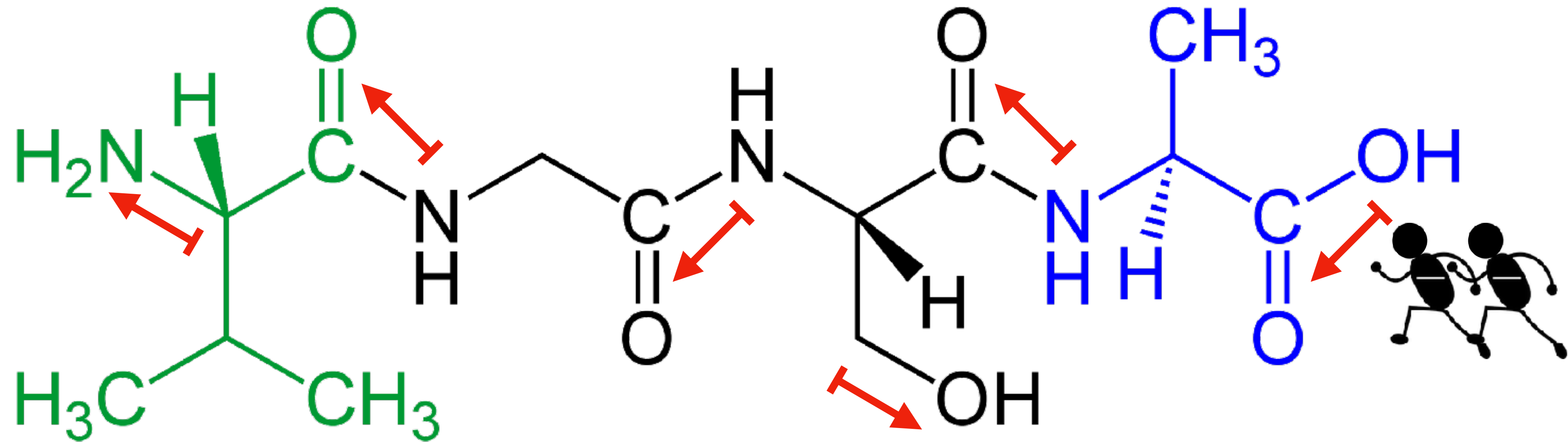


# Example of Convergence of Wavefunction

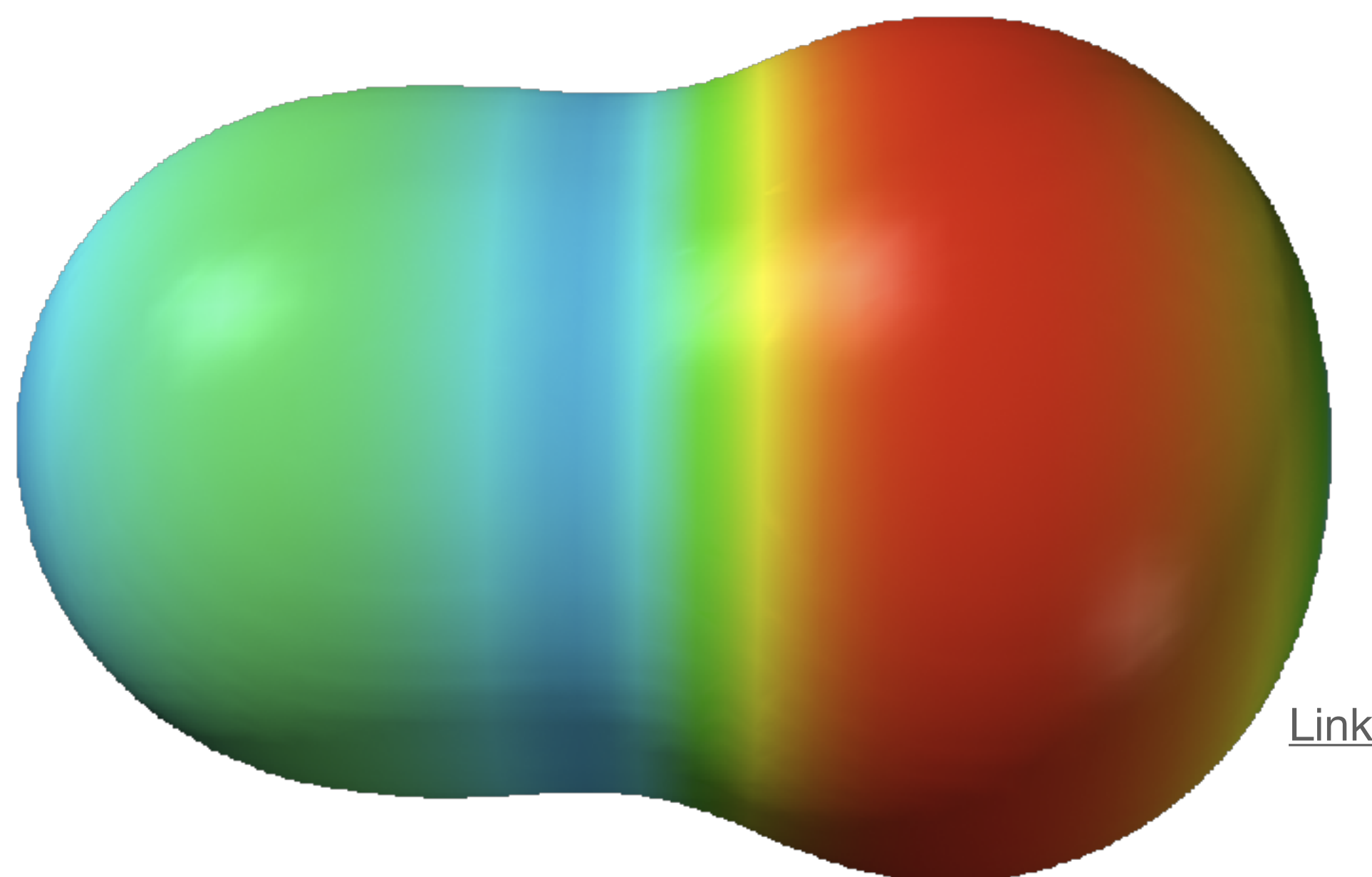
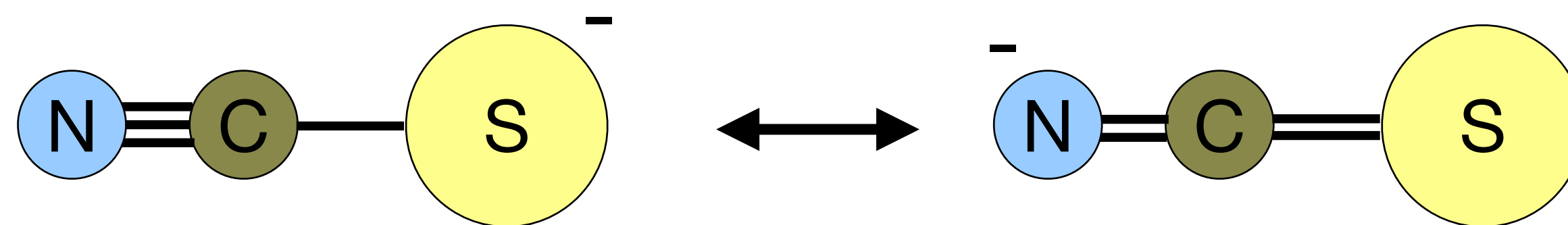




# Peptides and Dancing Charges



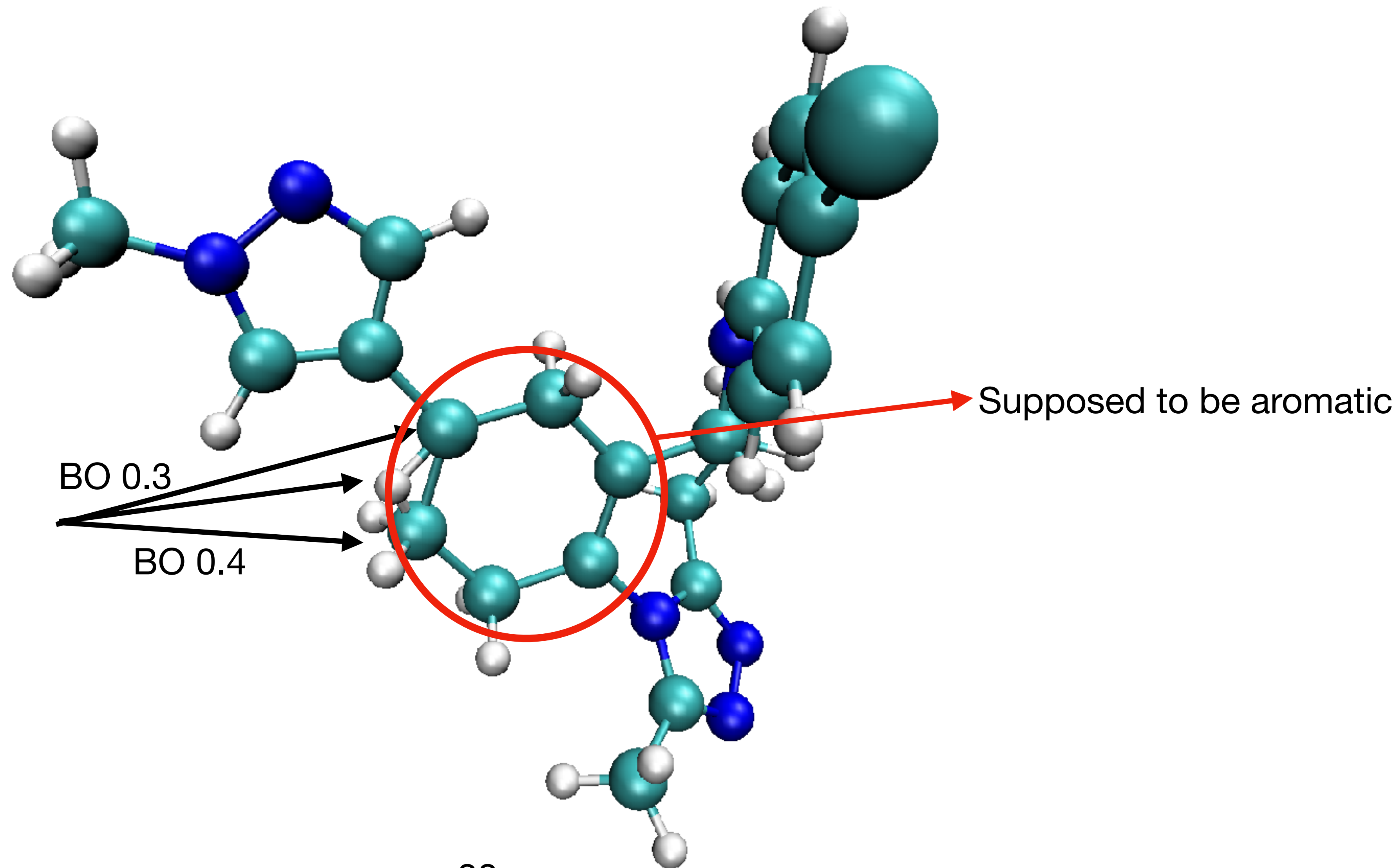
# Population Analysis and Bond Orders



[Link](#)

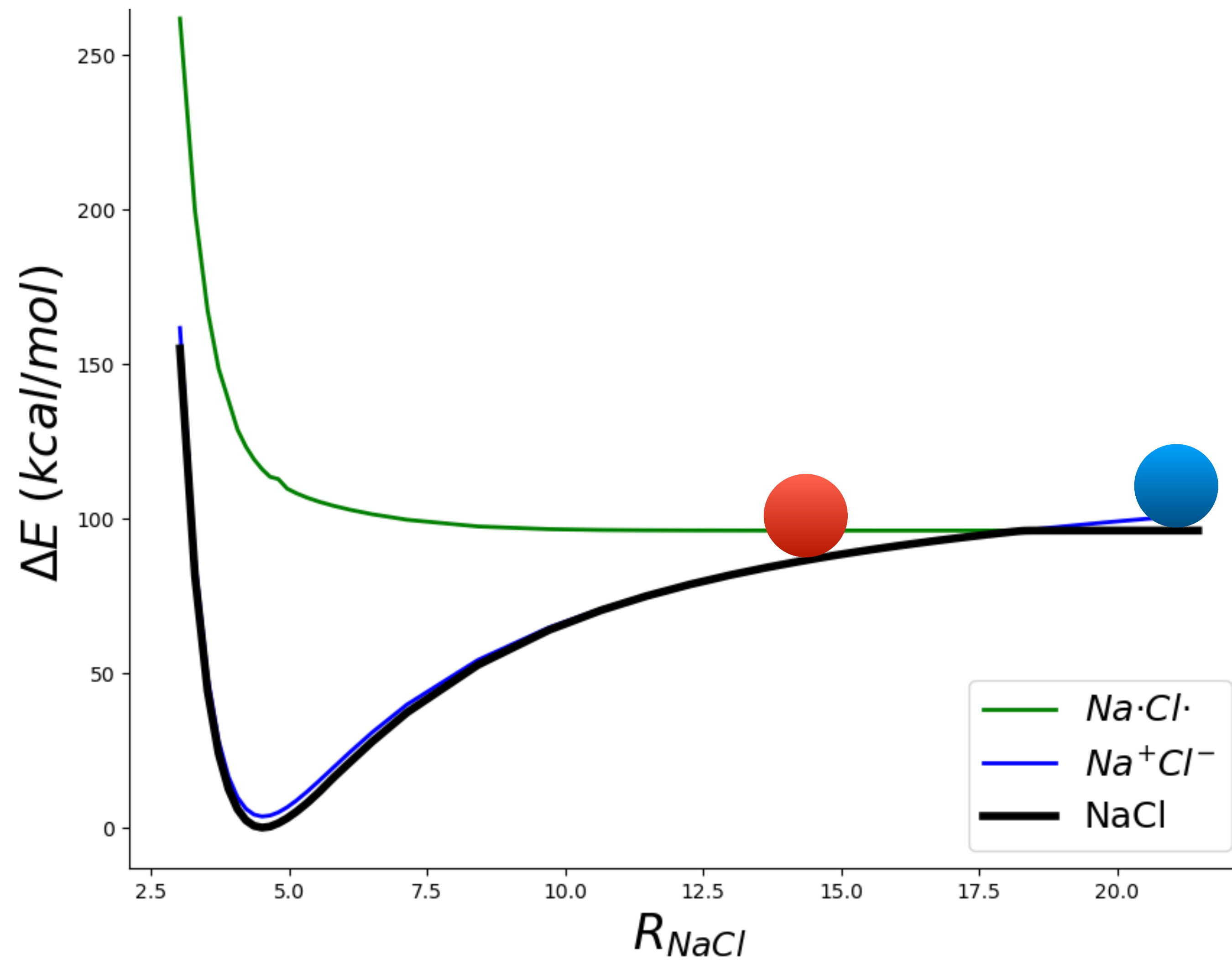
|   | N     | C     | S     |
|---|-------|-------|-------|
| N | 0.000 | 2.555 | 0.357 |
| C | 2.555 | 0.000 | 1.385 |
| S | 0.357 | 1.385 | 0.000 |

# Bond Order Matrices



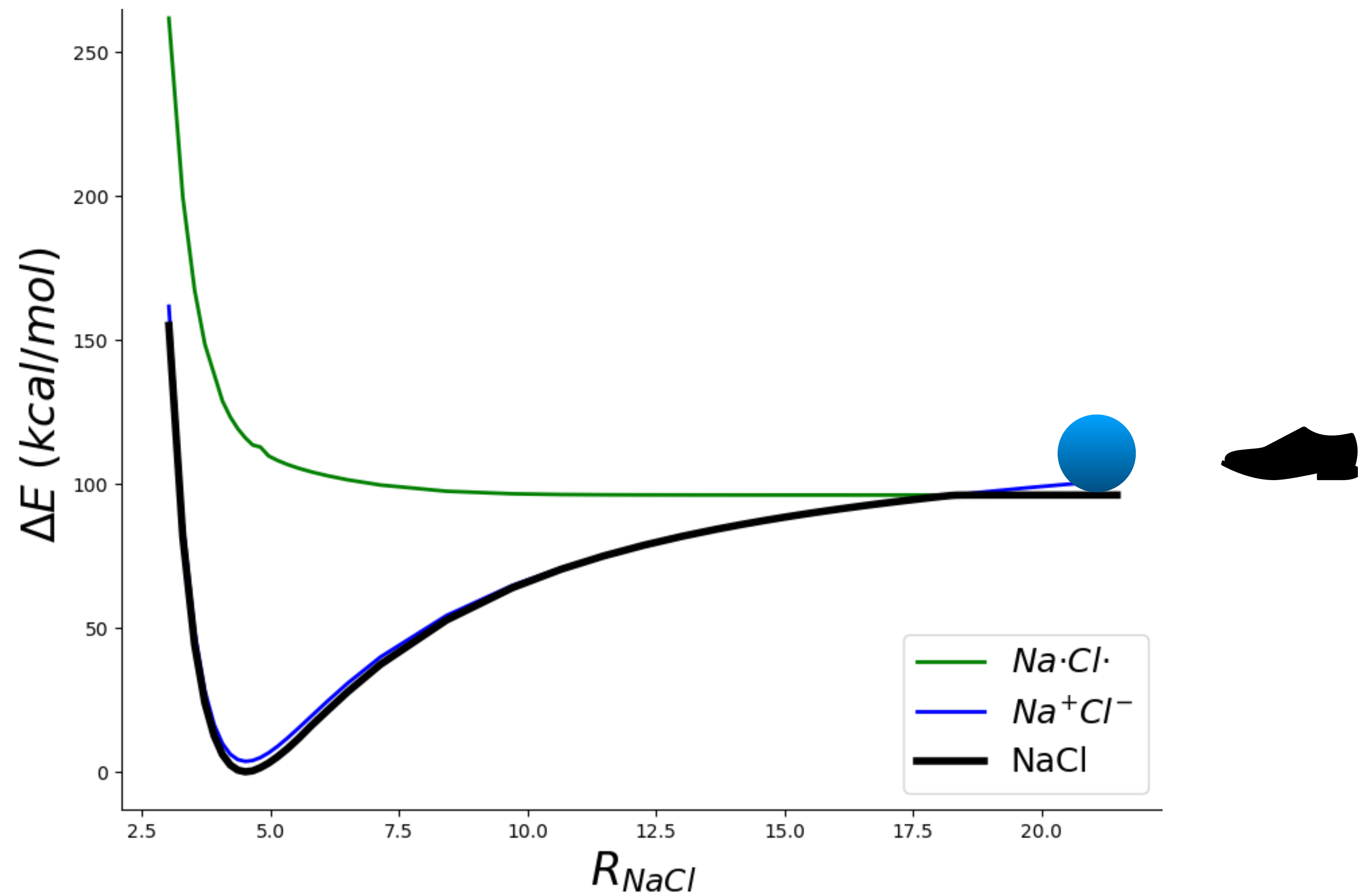


# Optg and MD

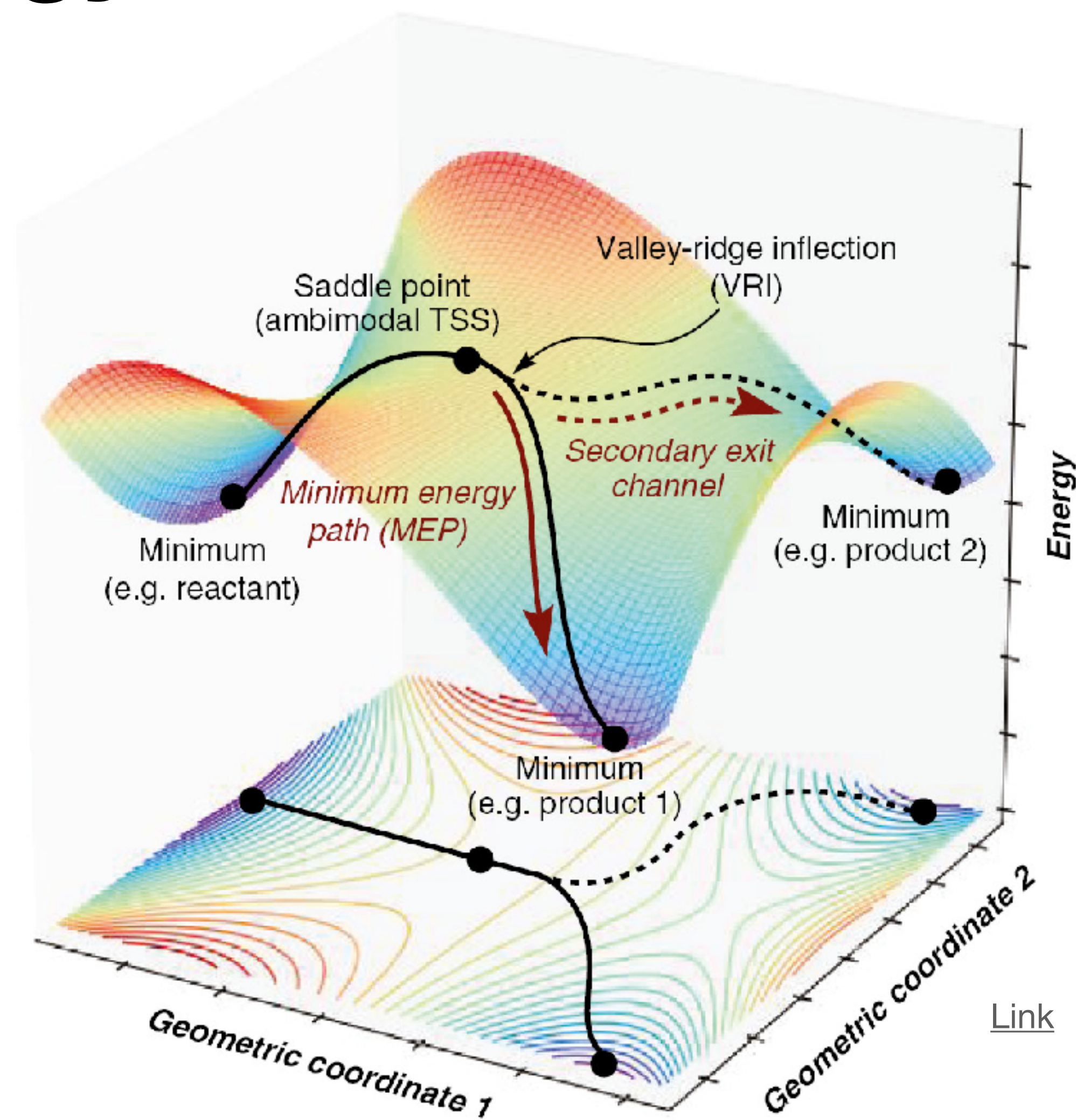




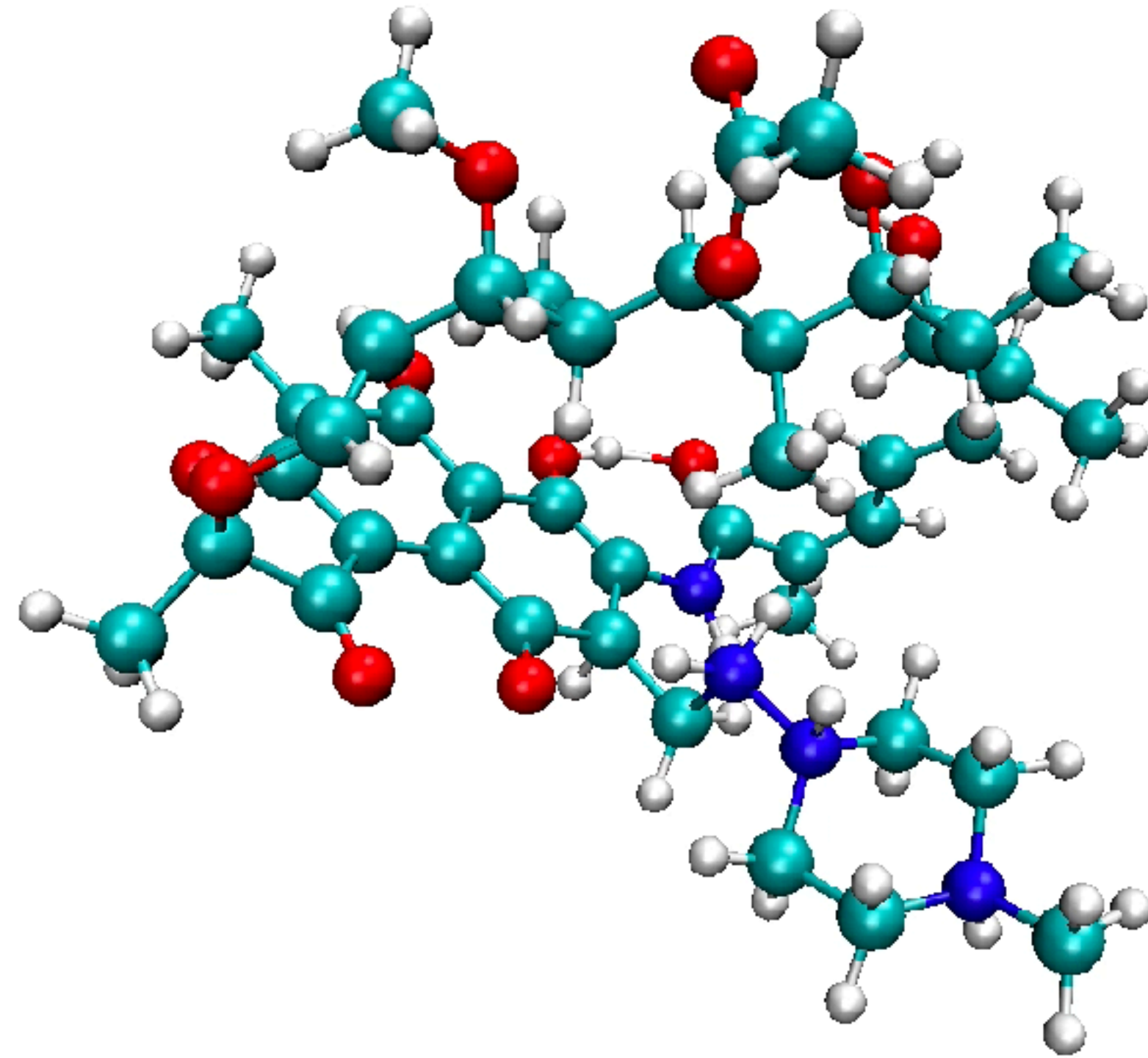
# Optg and MD



# Potential Energy Surface

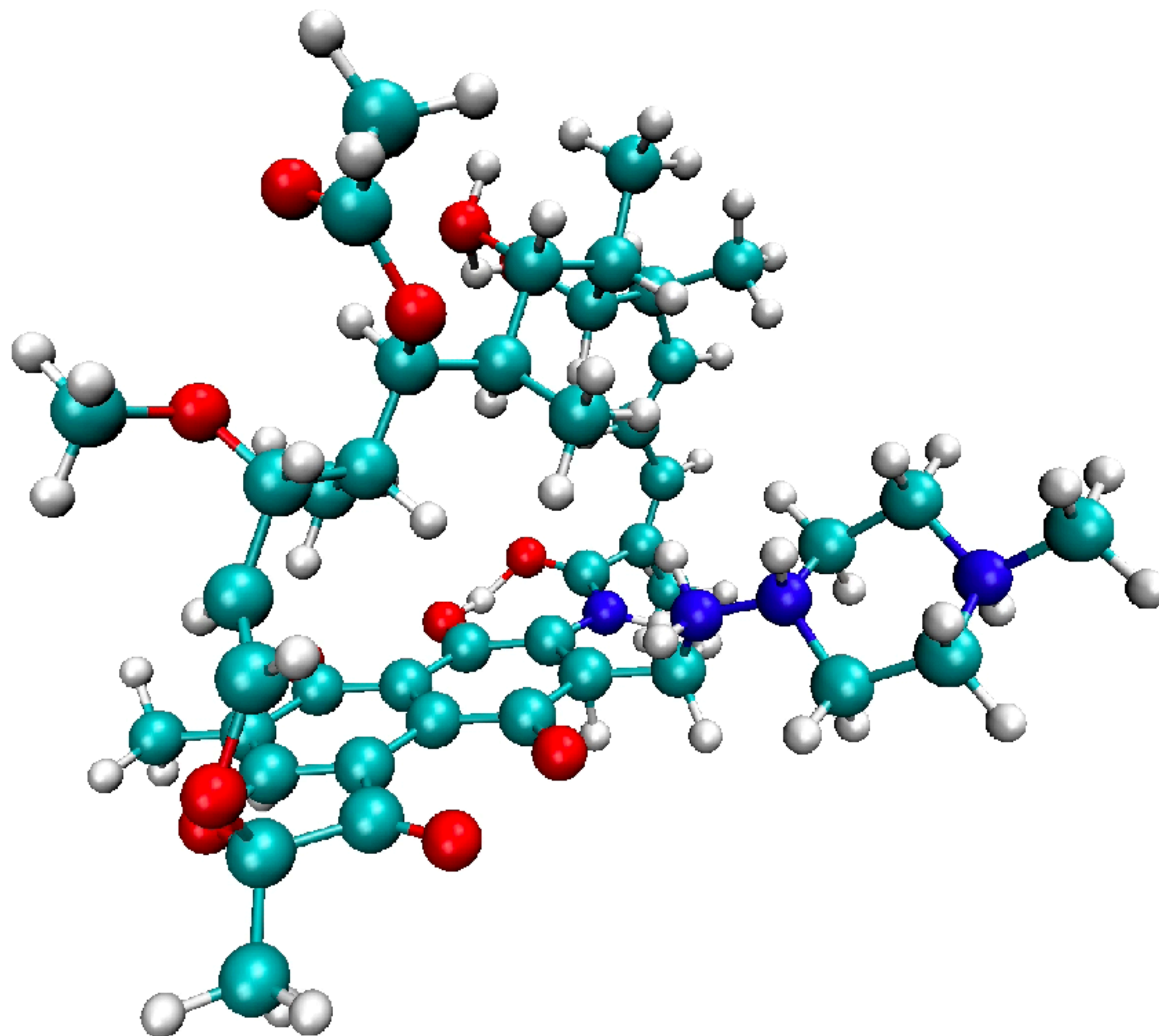


# MD of Incorrect Structure





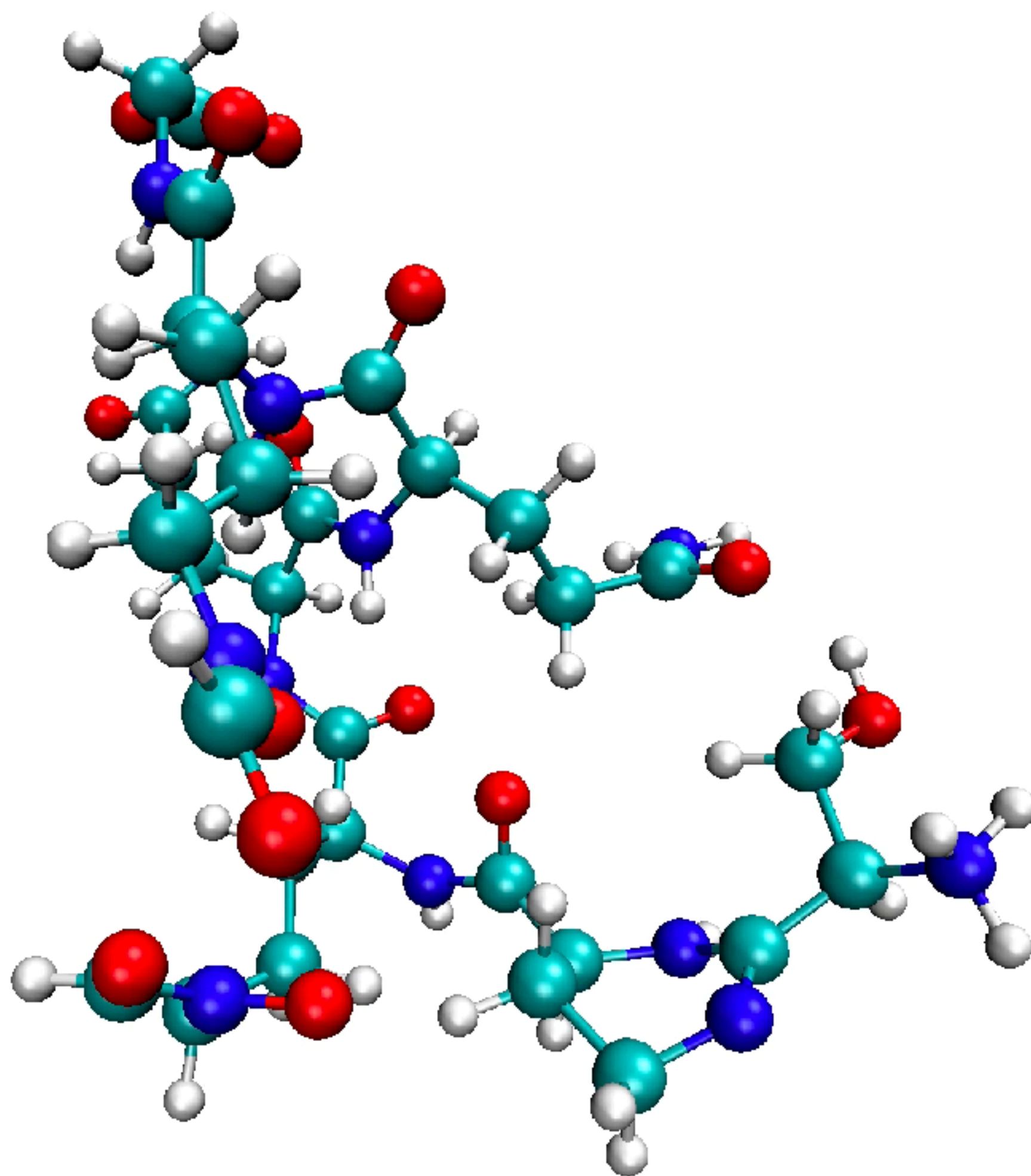
# Optg of (same) Incorrect Structure





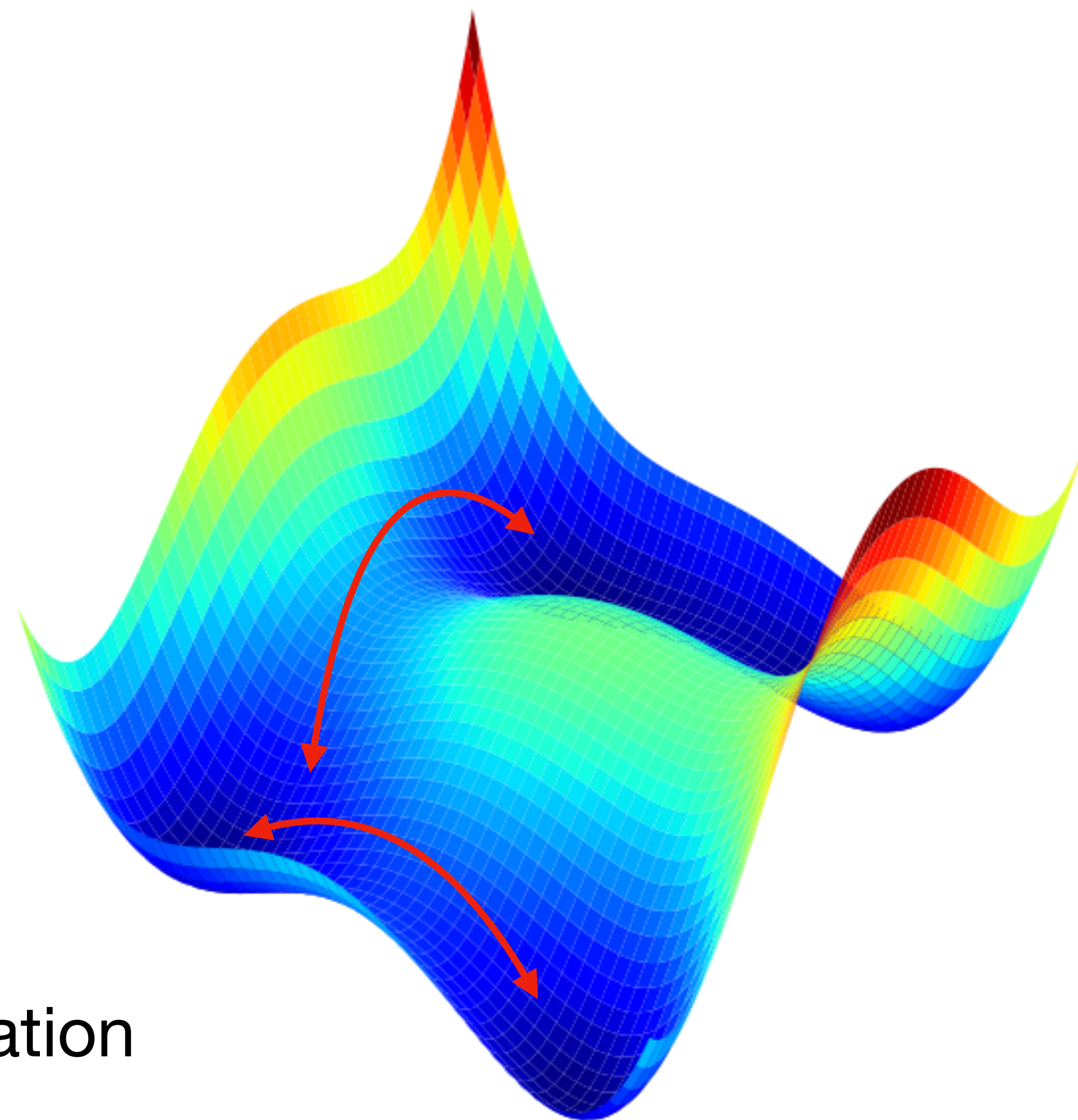
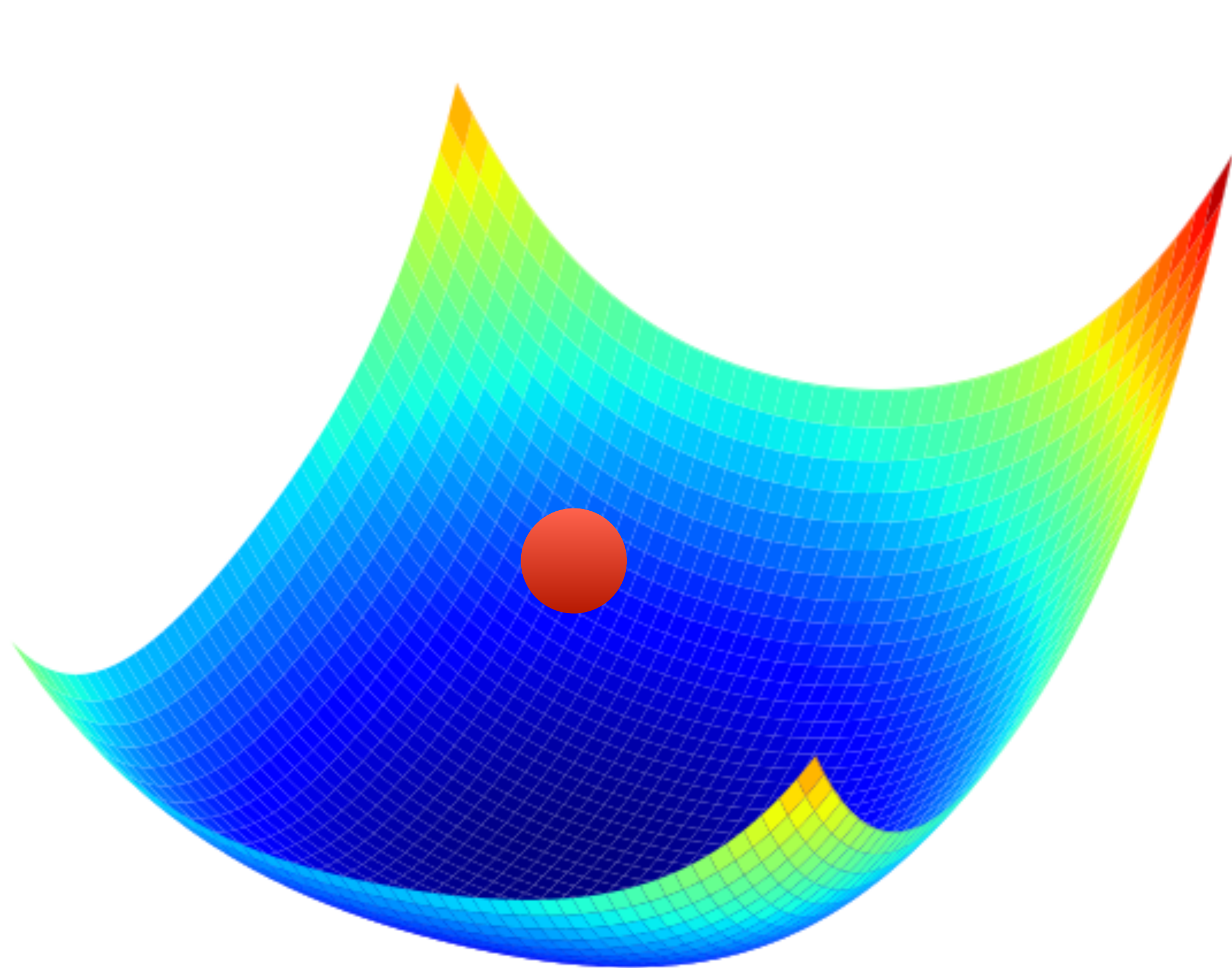
# Radicals Force Their Way

---



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

# Molecules as Minima in Surfaces

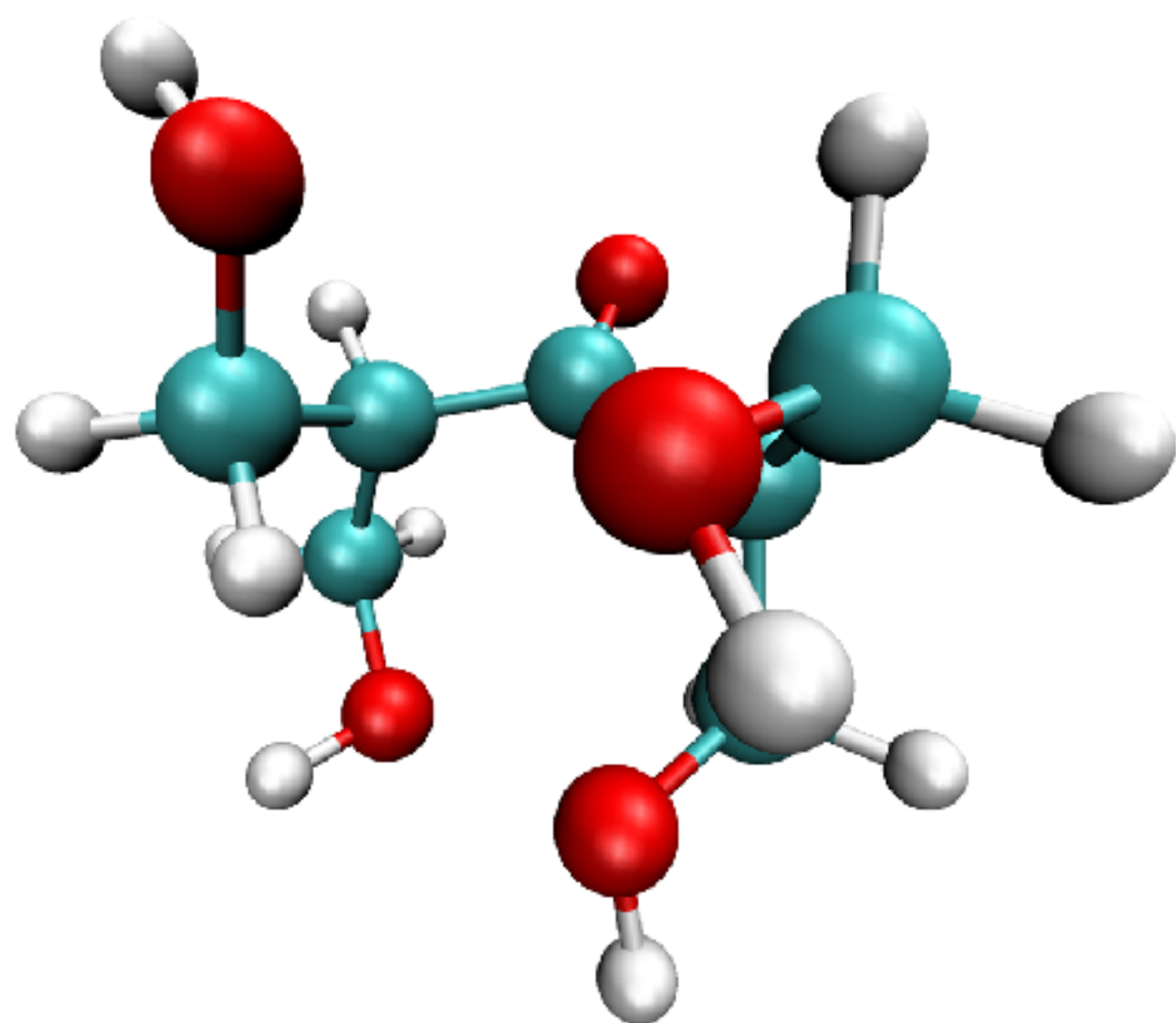


Conformers: different minima obtained by rotation  
along single bond

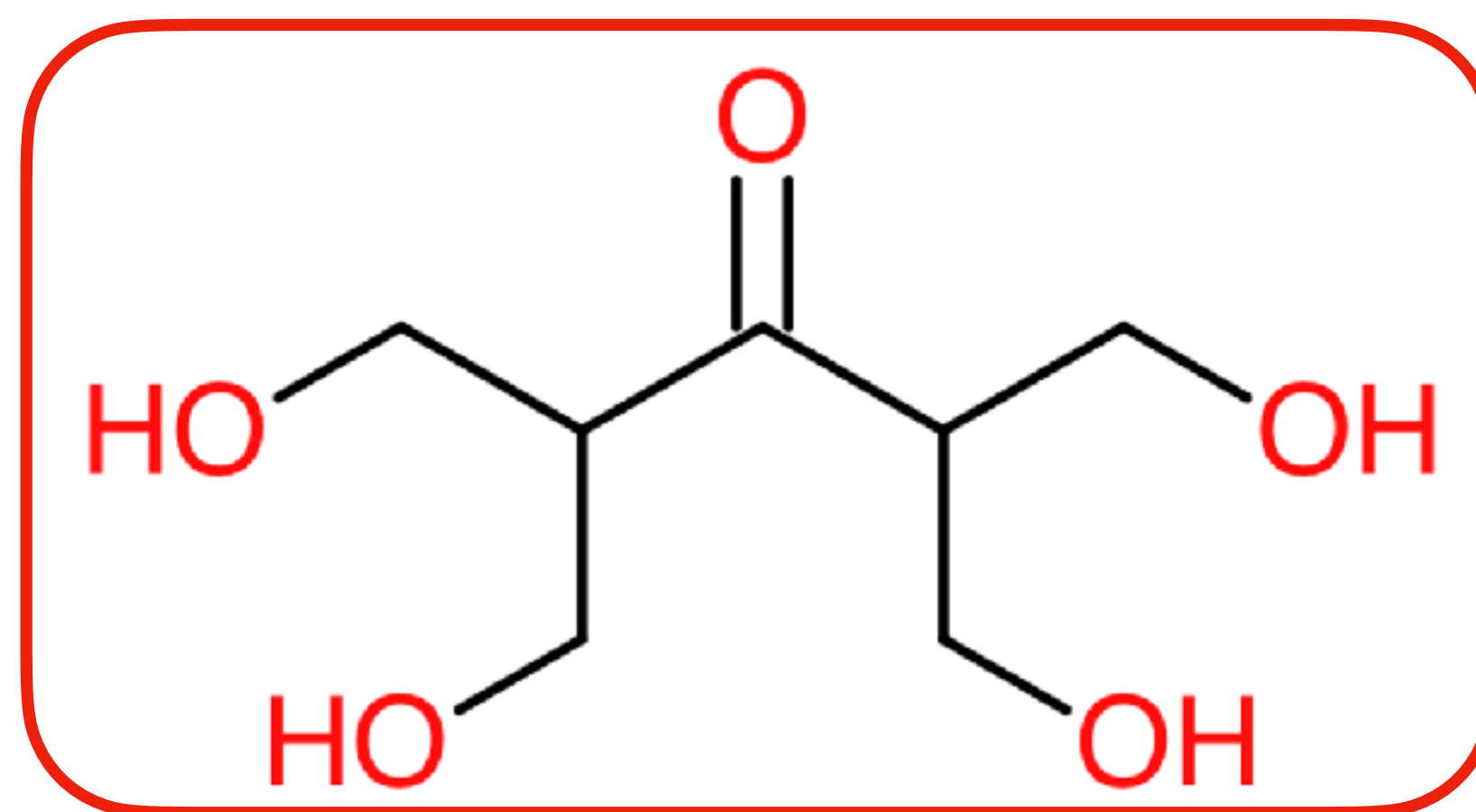


# All Conformers are equal... are they?

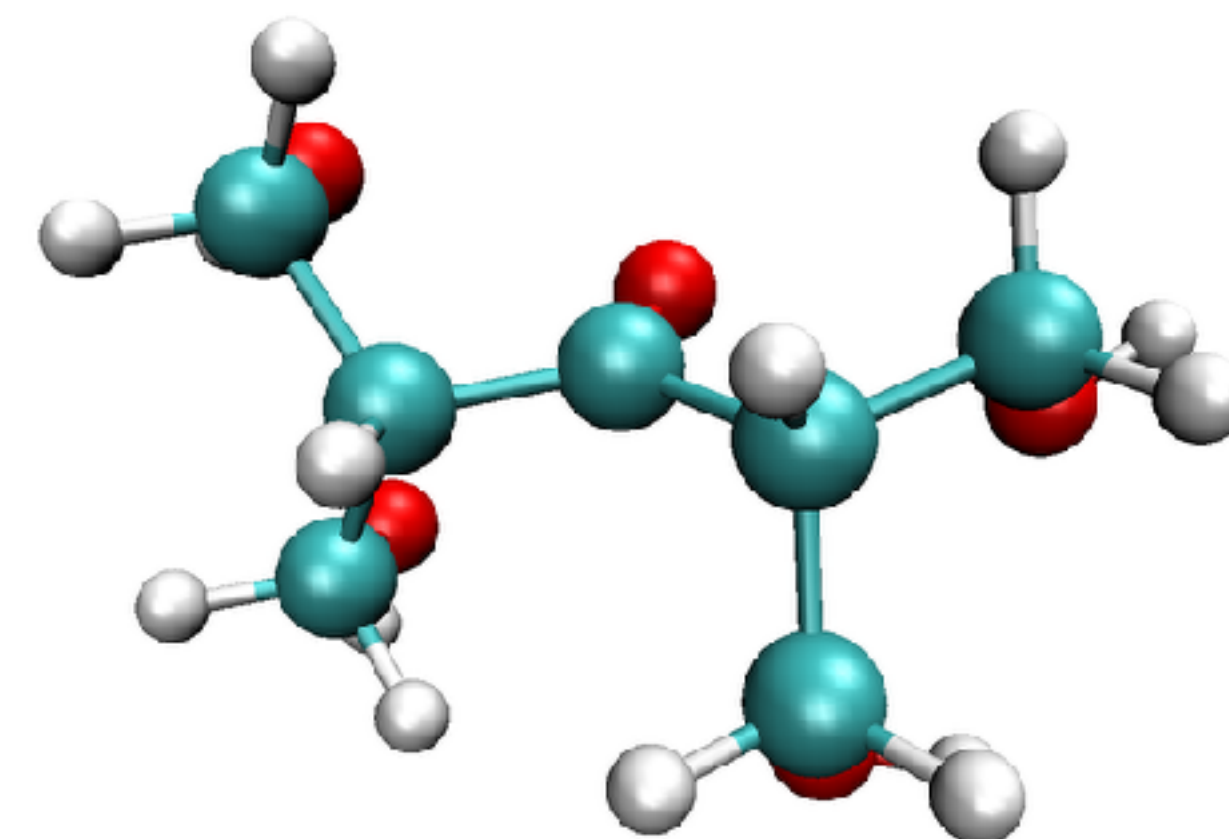
Conformer A



Solvent 1

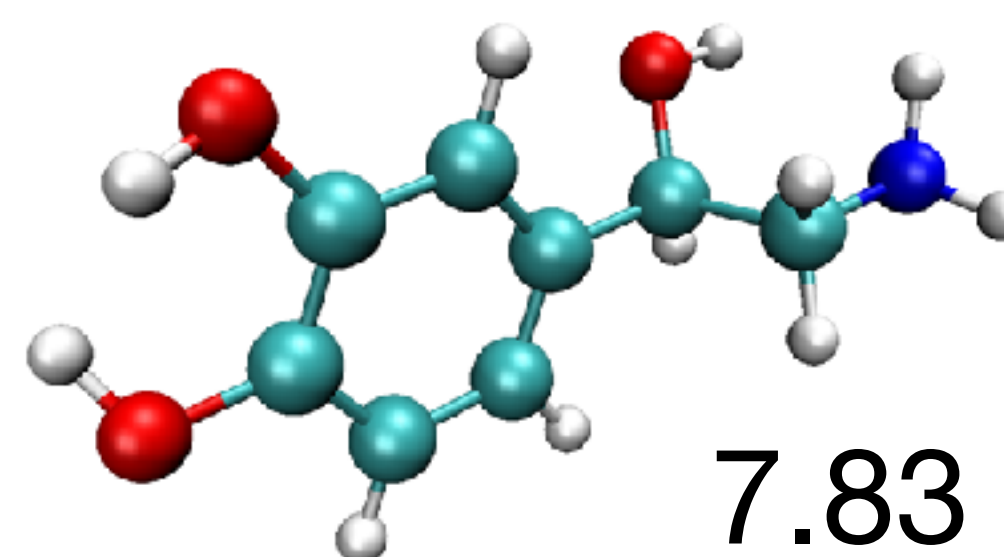
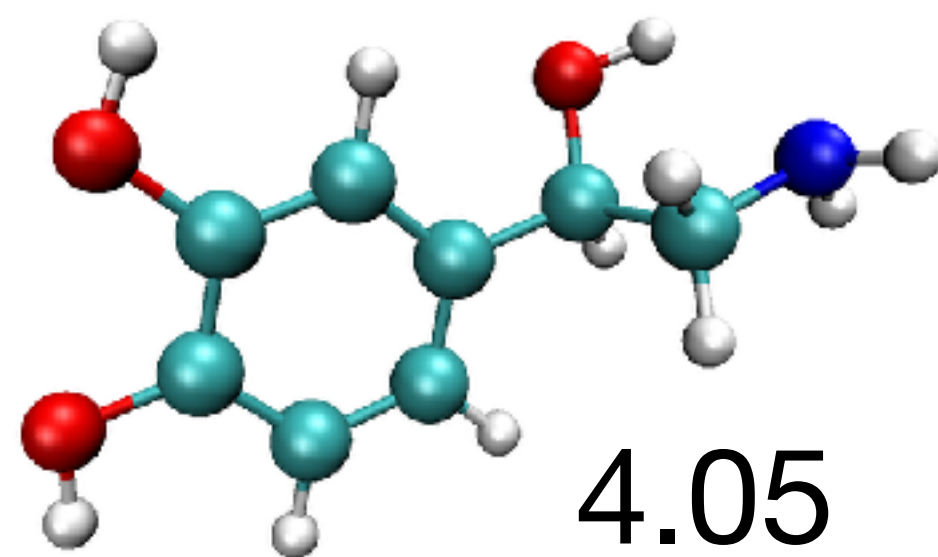
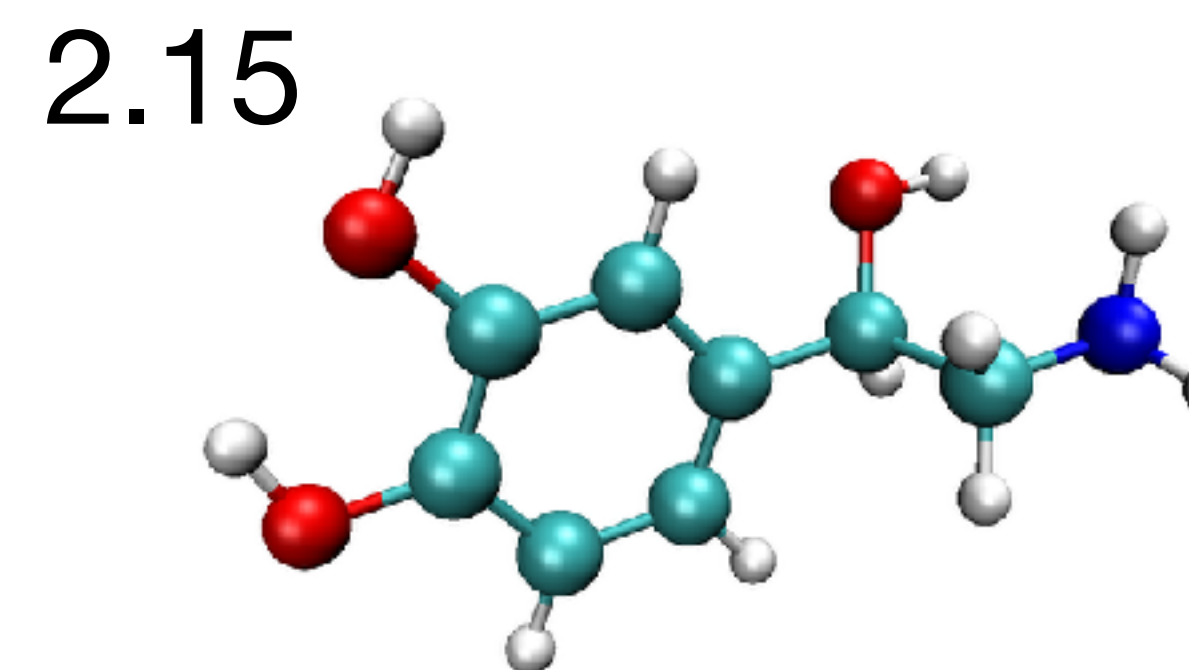
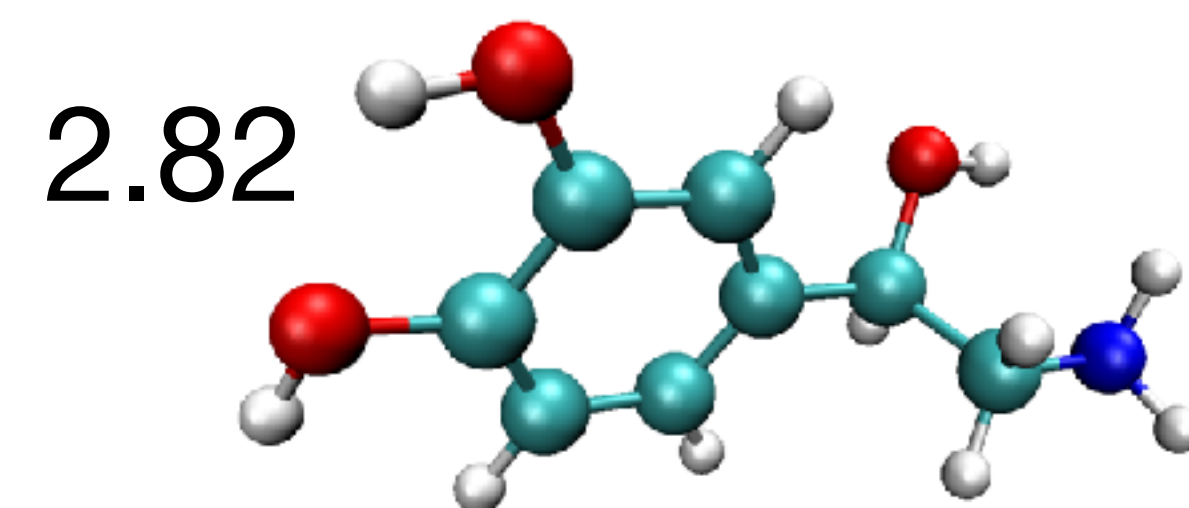
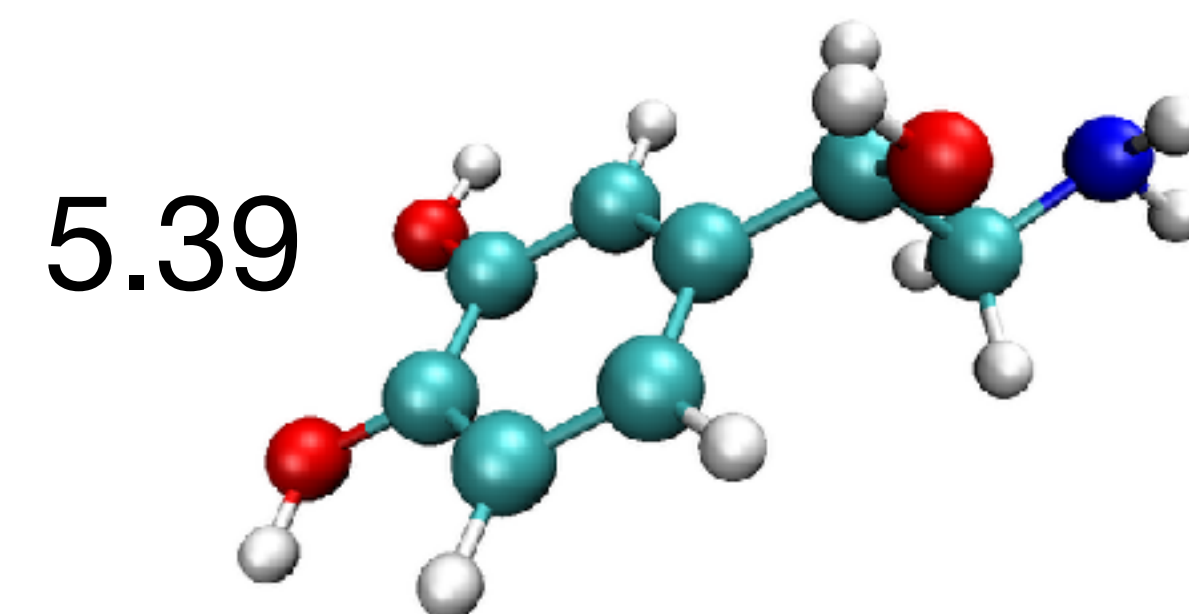
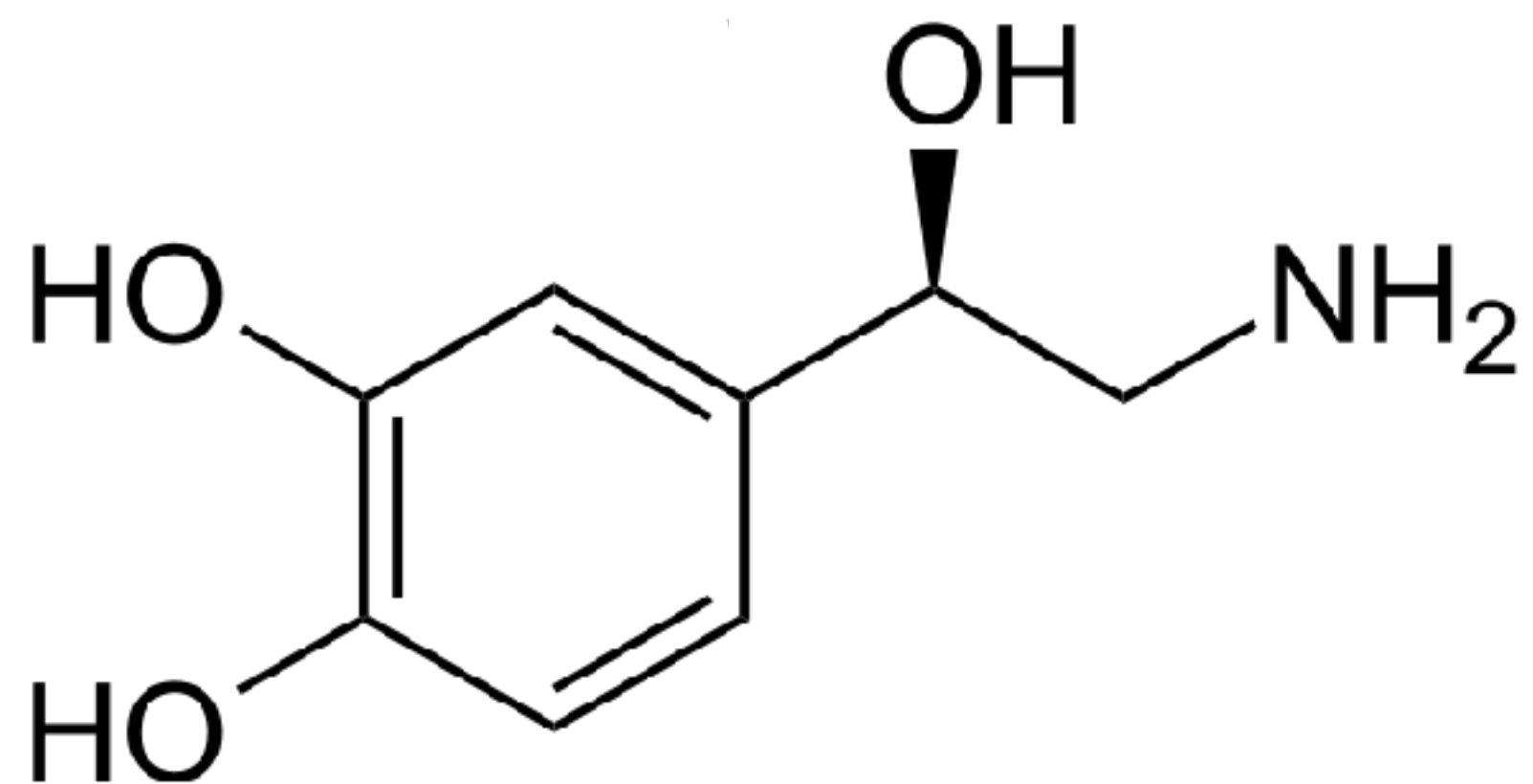
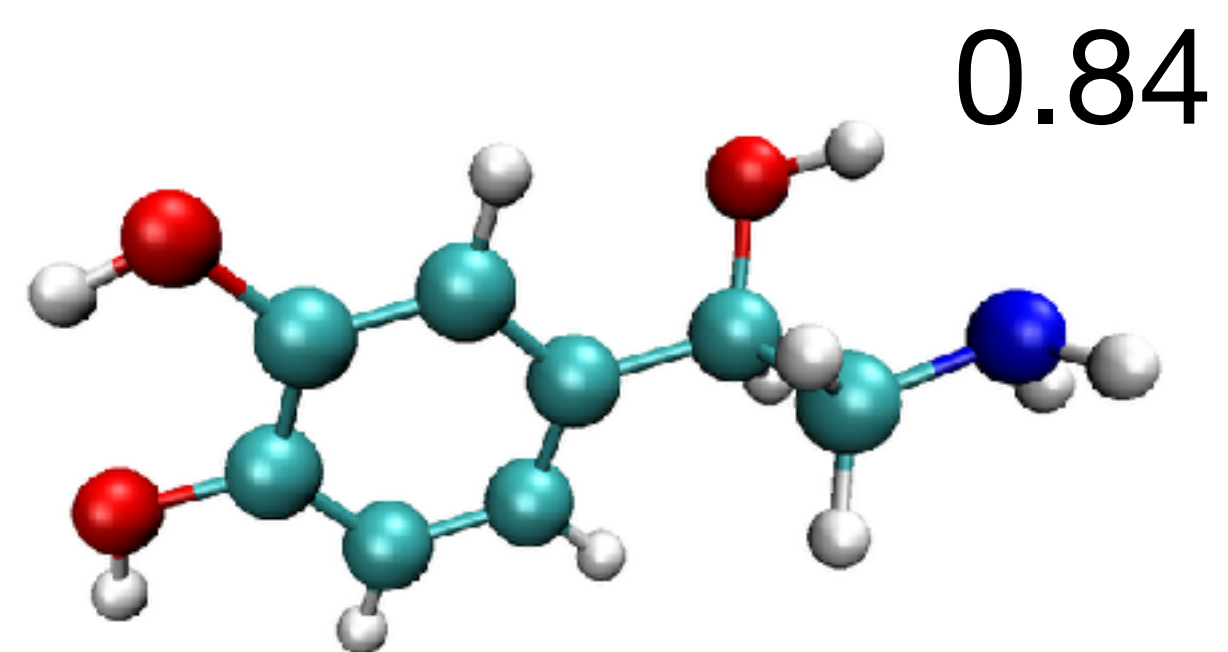
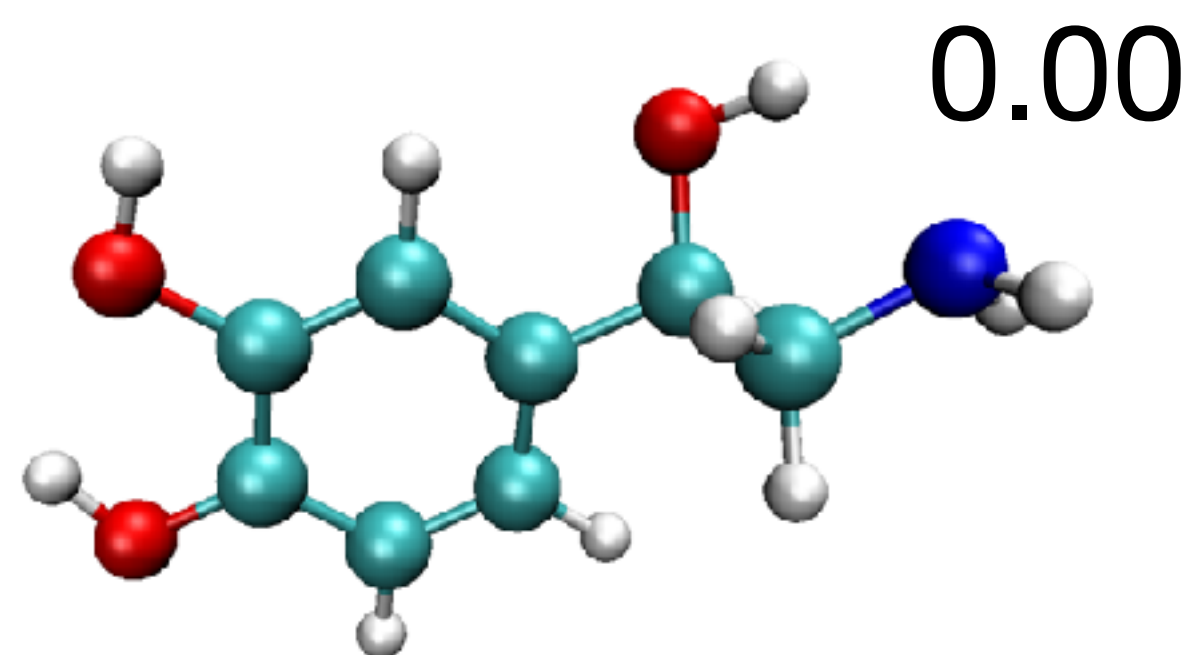
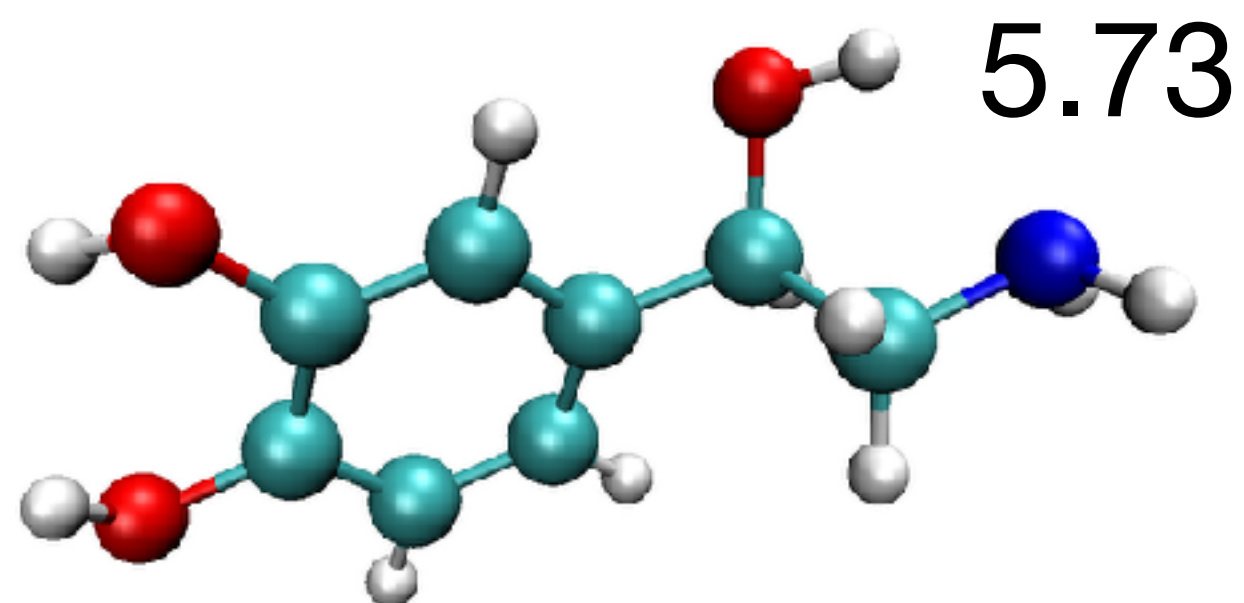


Conformer B



Solvent 2

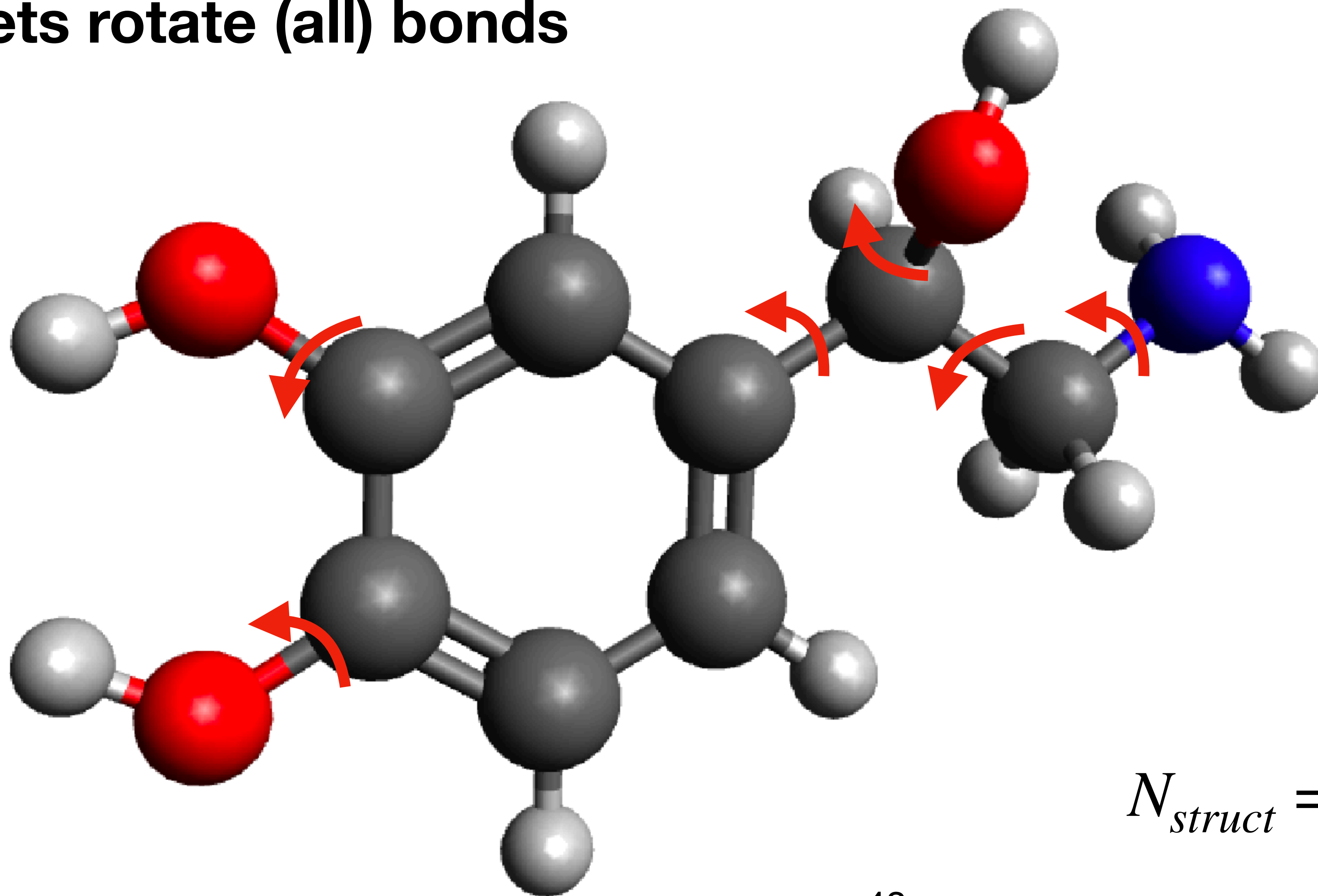
# All Conformers are equal...





# Simple Algorithms for Finding Conformers

Lets rotate (all) bonds

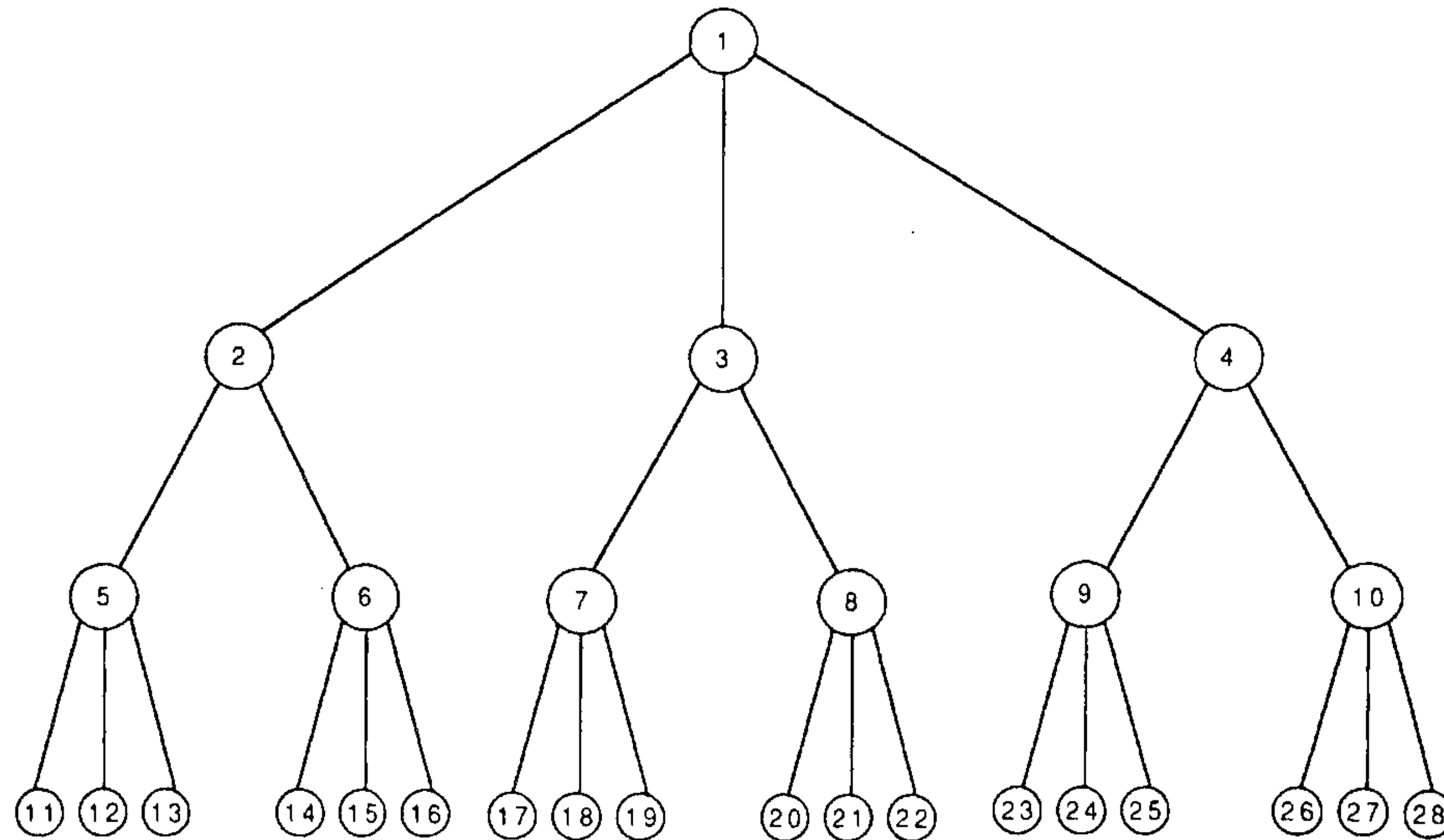


$$N_{struct} = \left( \frac{360}{\theta} \right)^N$$

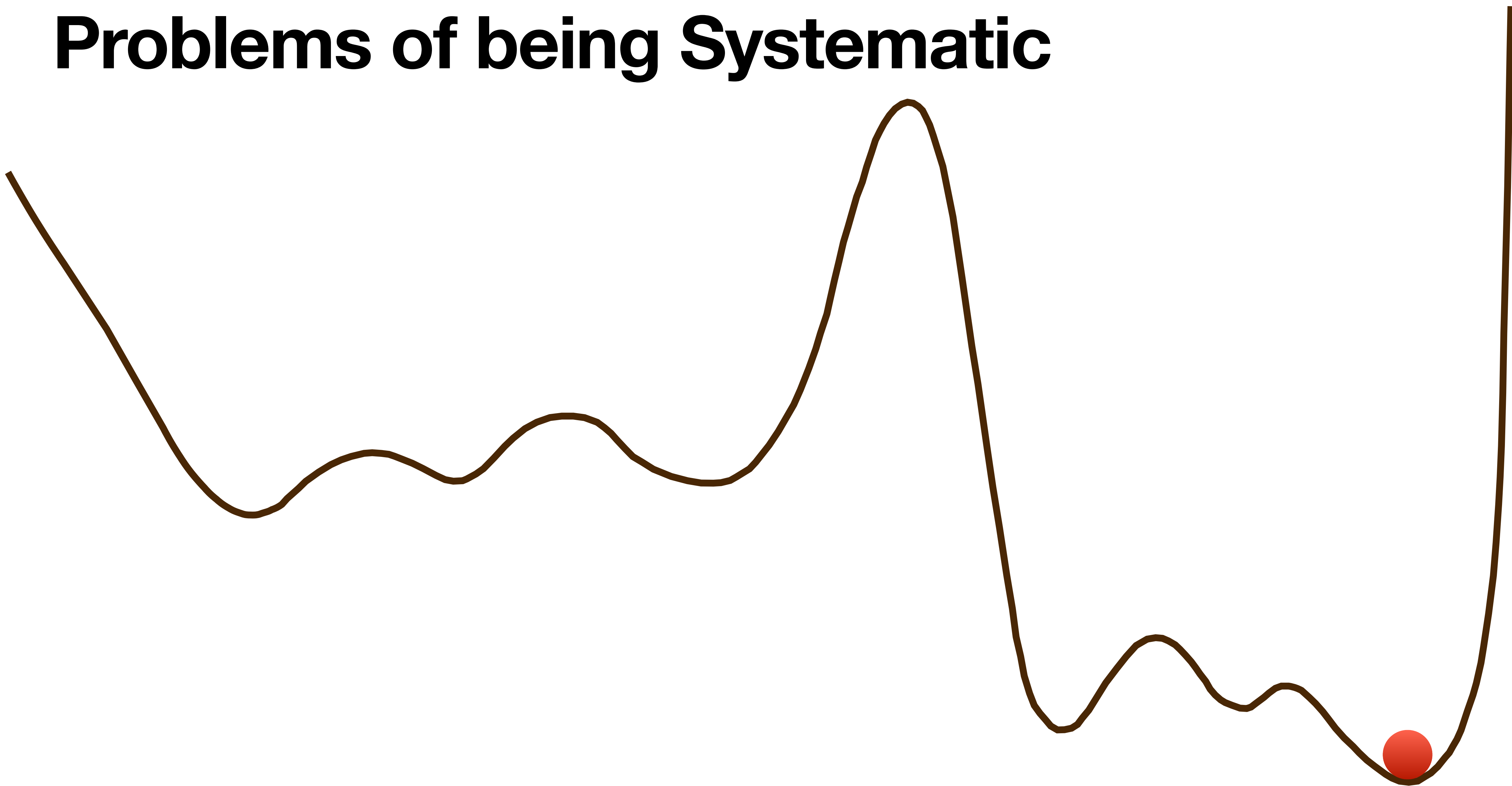
$$N_{struct} = \left( \frac{360}{45} \right)^6 = 262144$$

# Simple Algorithms for Finding Conformers

Lets use tree... searches

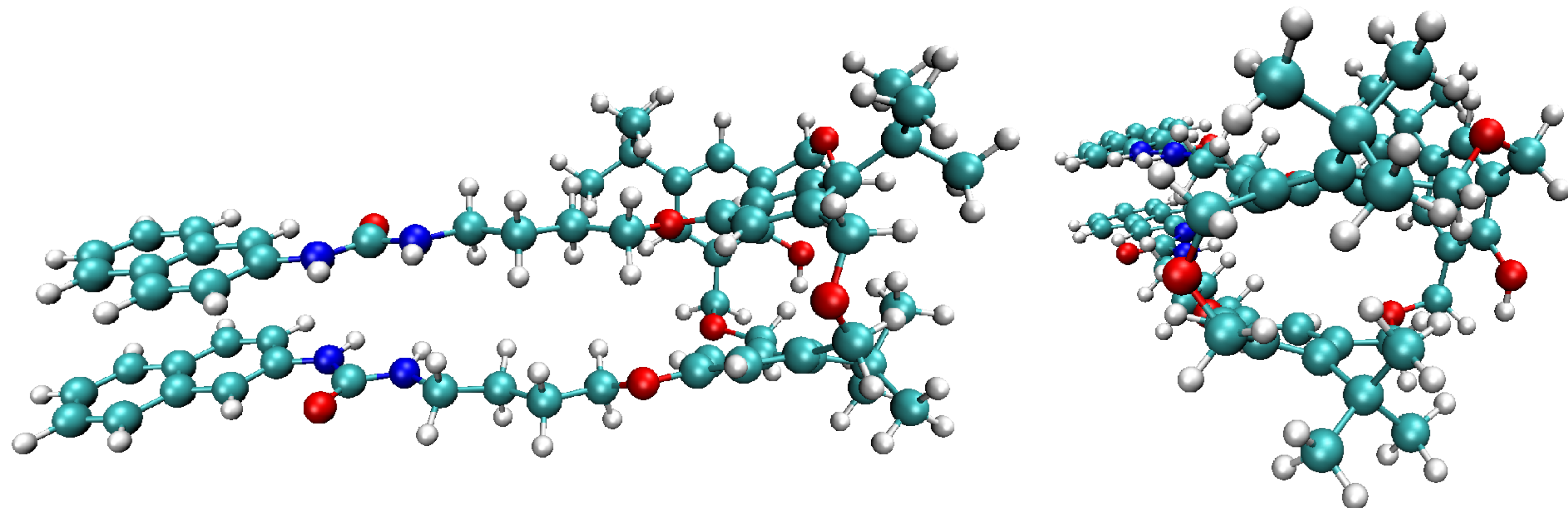


# Problems of being Systematic

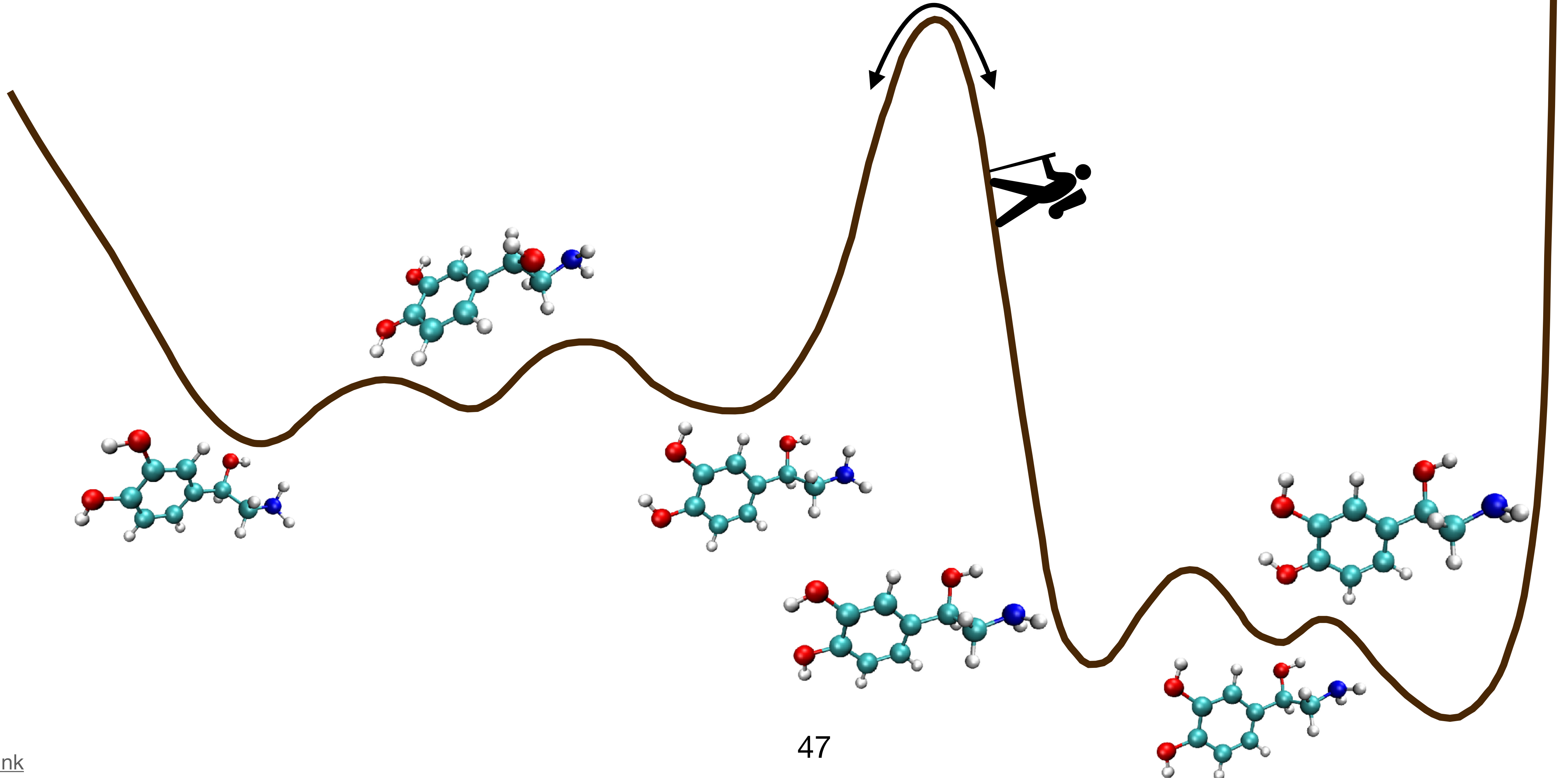




# Problems of being Systematic

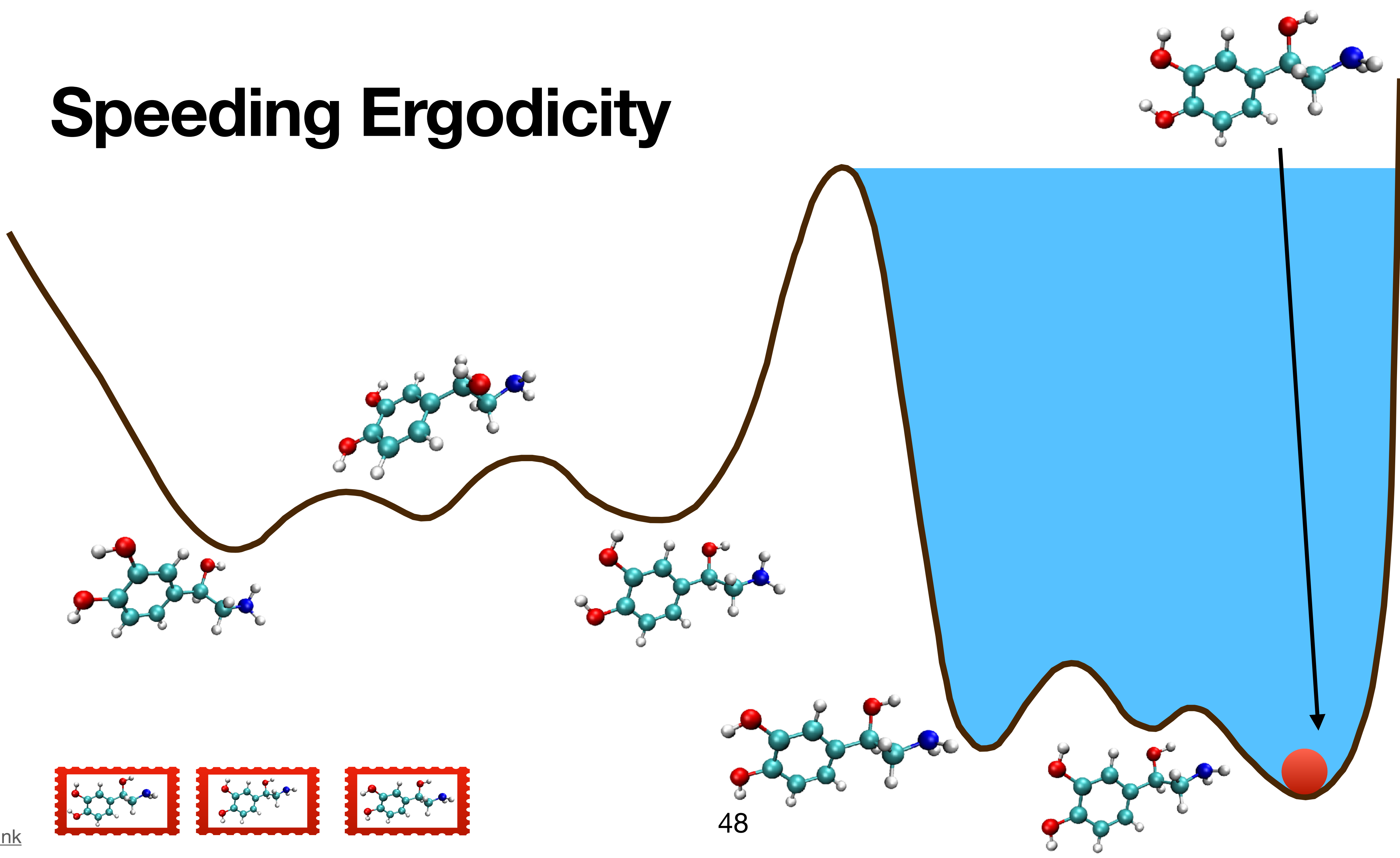


# Walk Over PES and Ergodicity





# Speeding Ergodicity

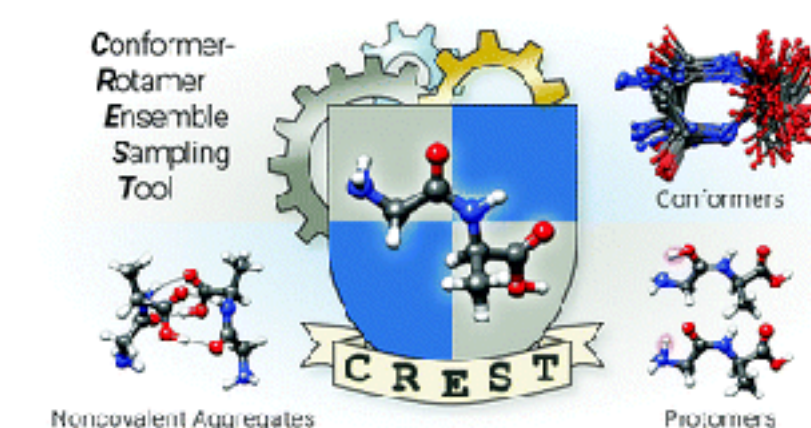
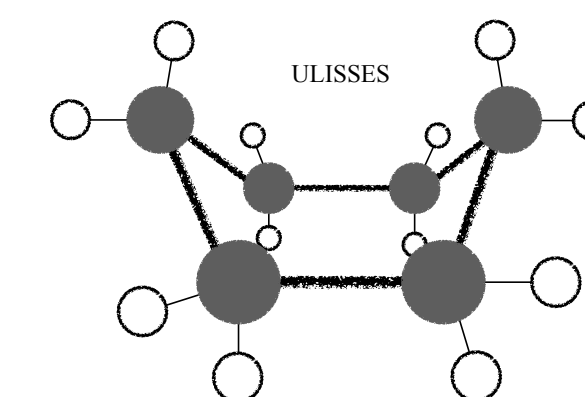
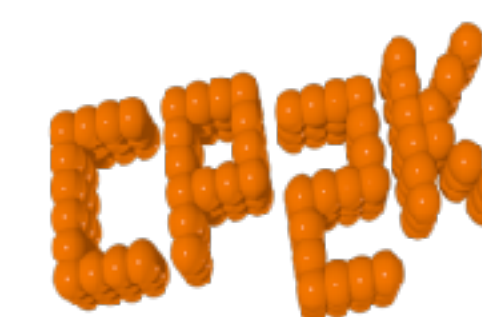


# Metadynamics Conformer Search

- 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.

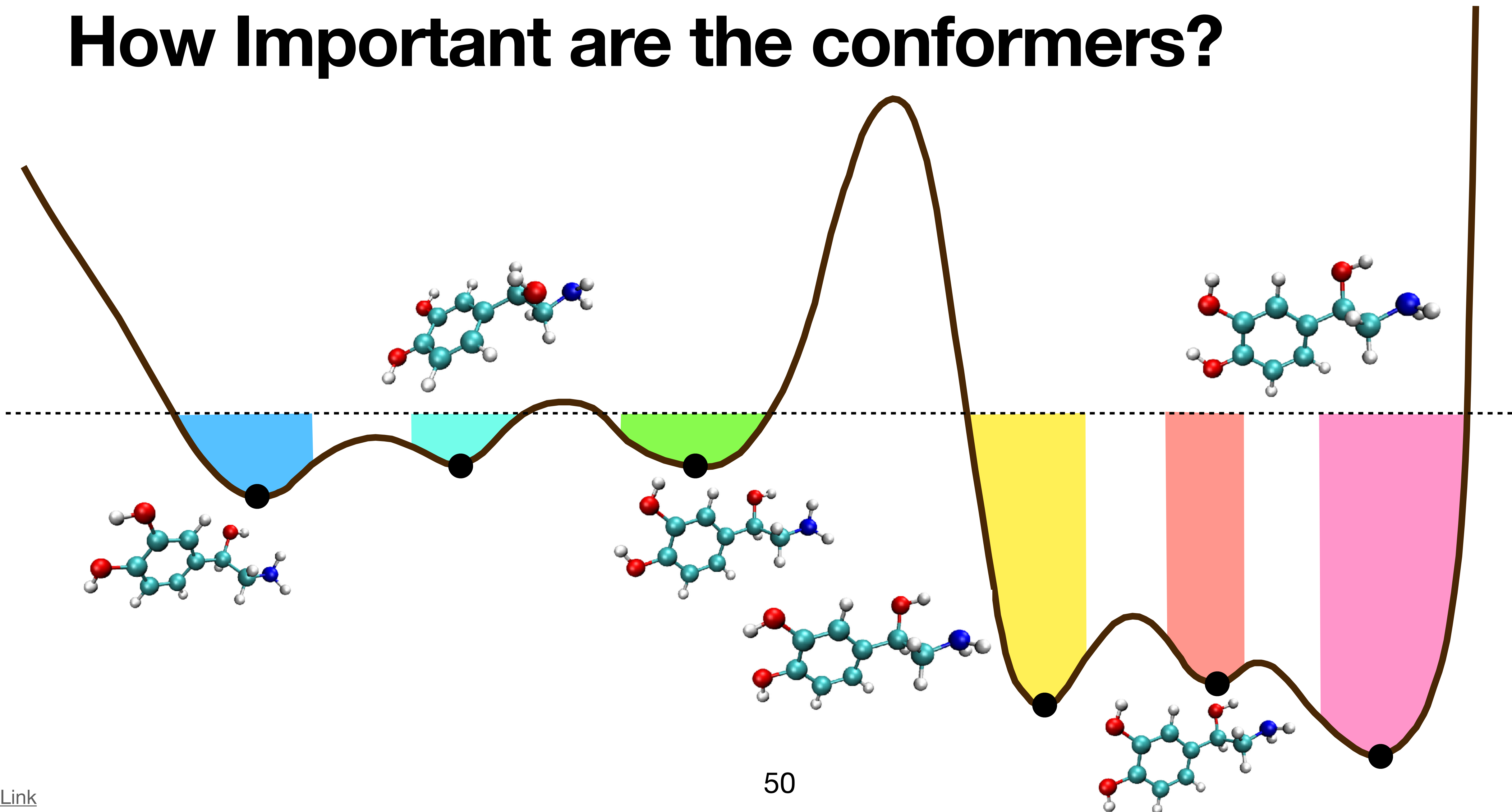
$$FF < SEQM < DFT \leq \Psi$$

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





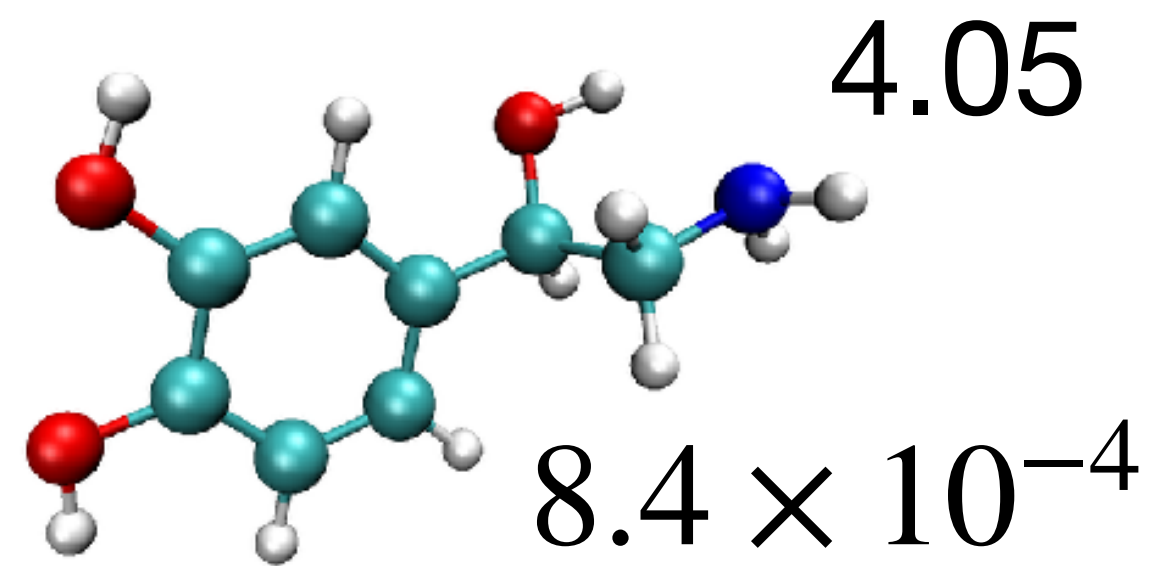
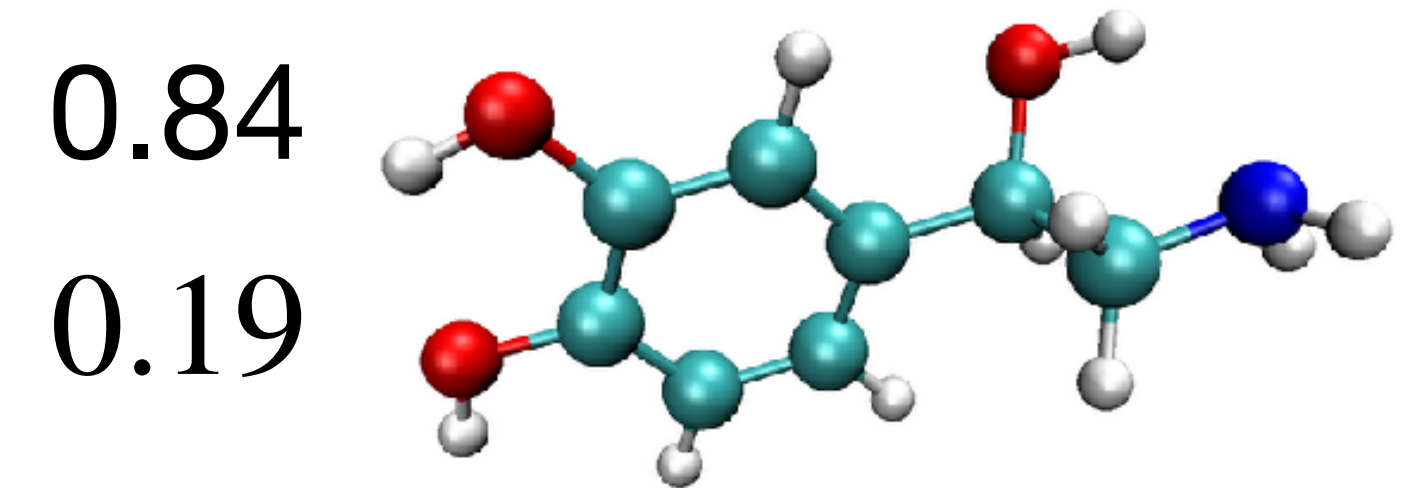
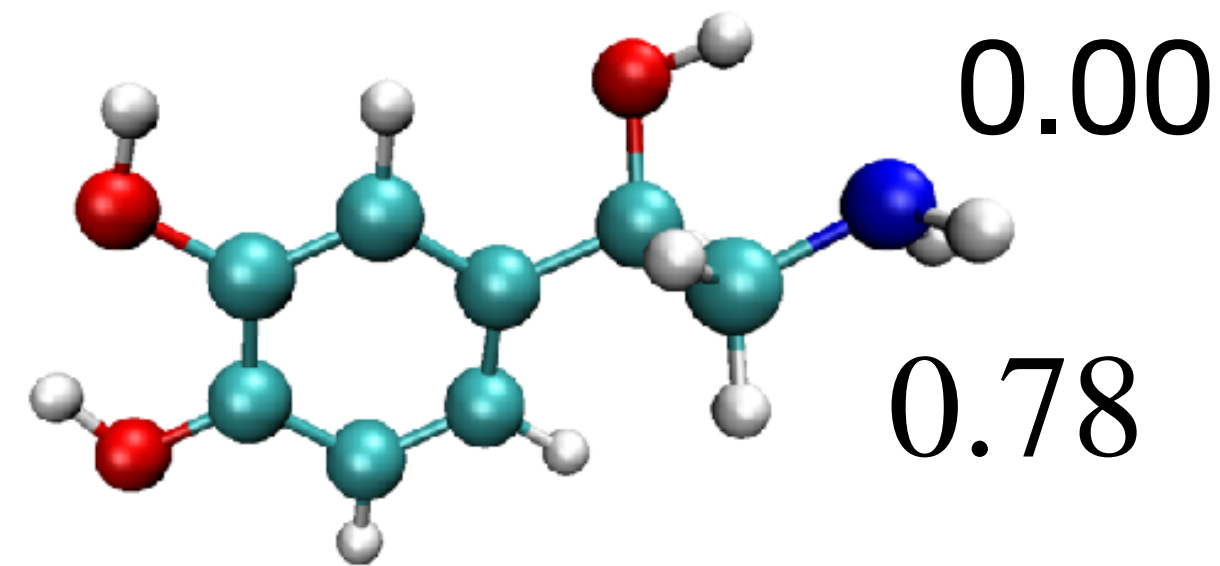
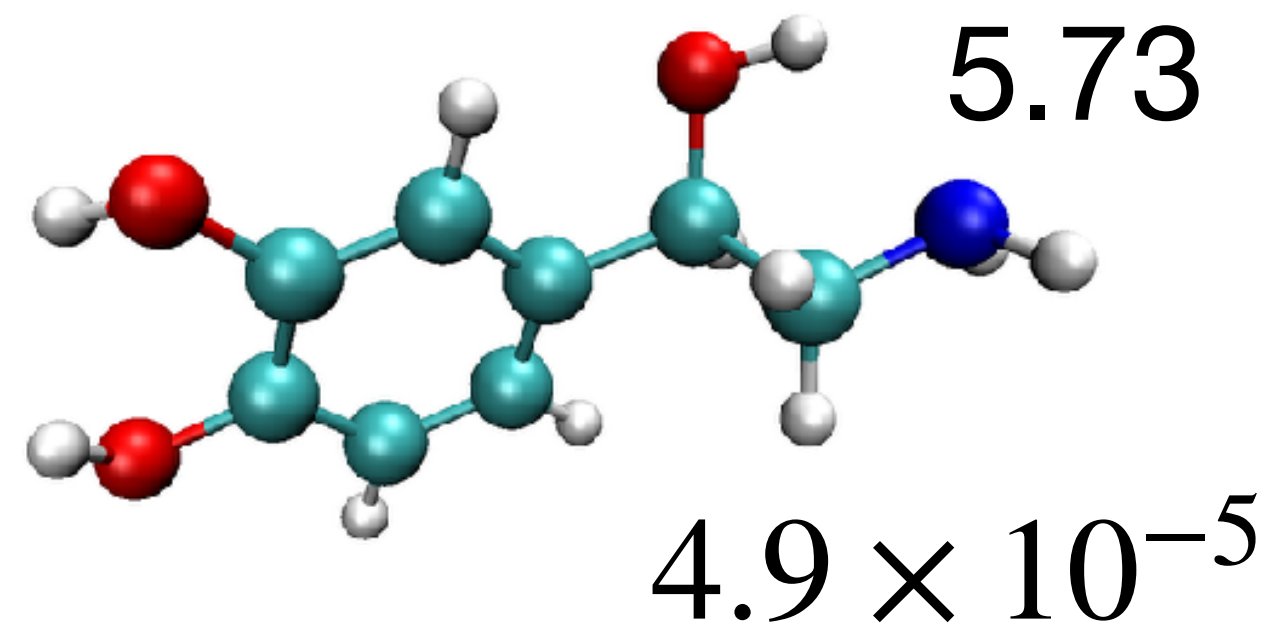
# How Important are the conformers?



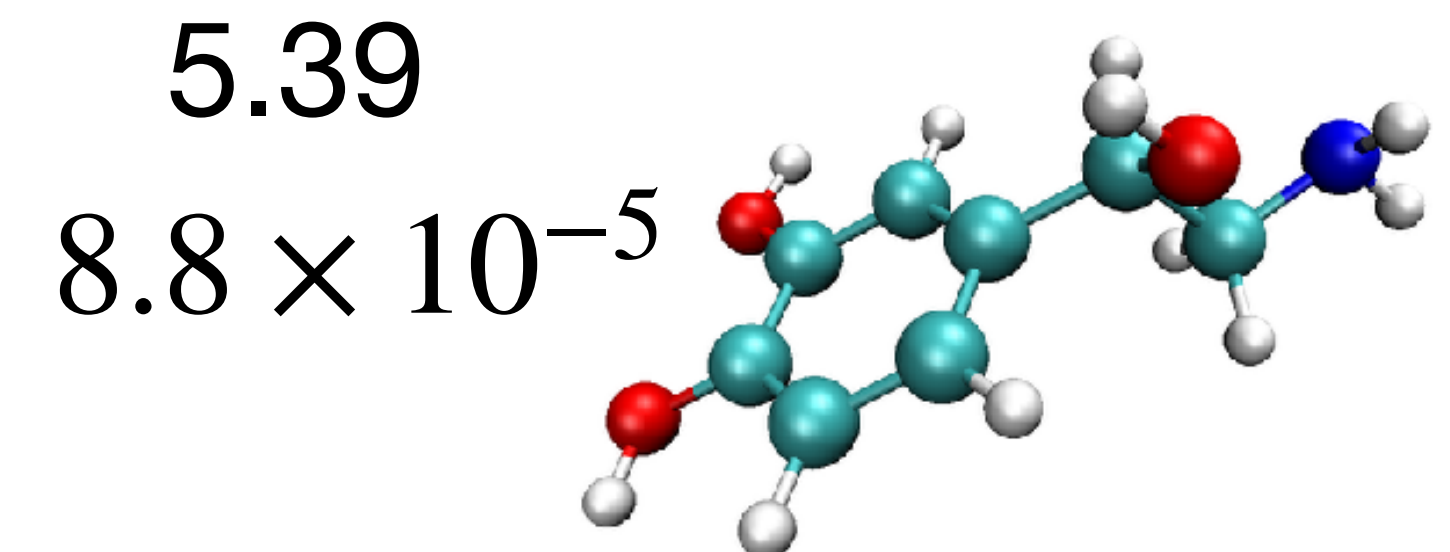
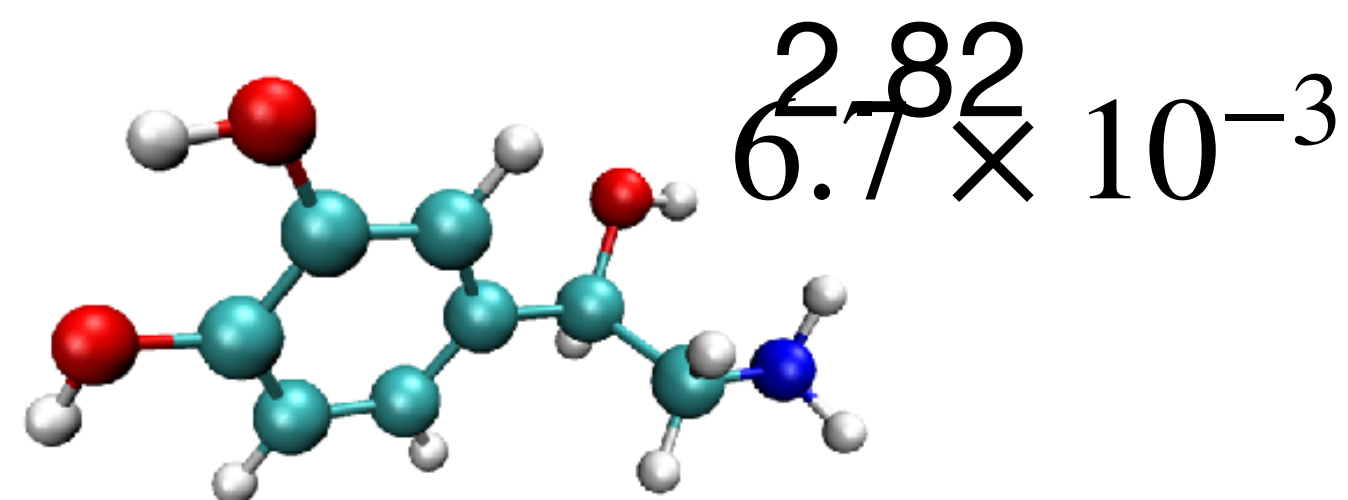
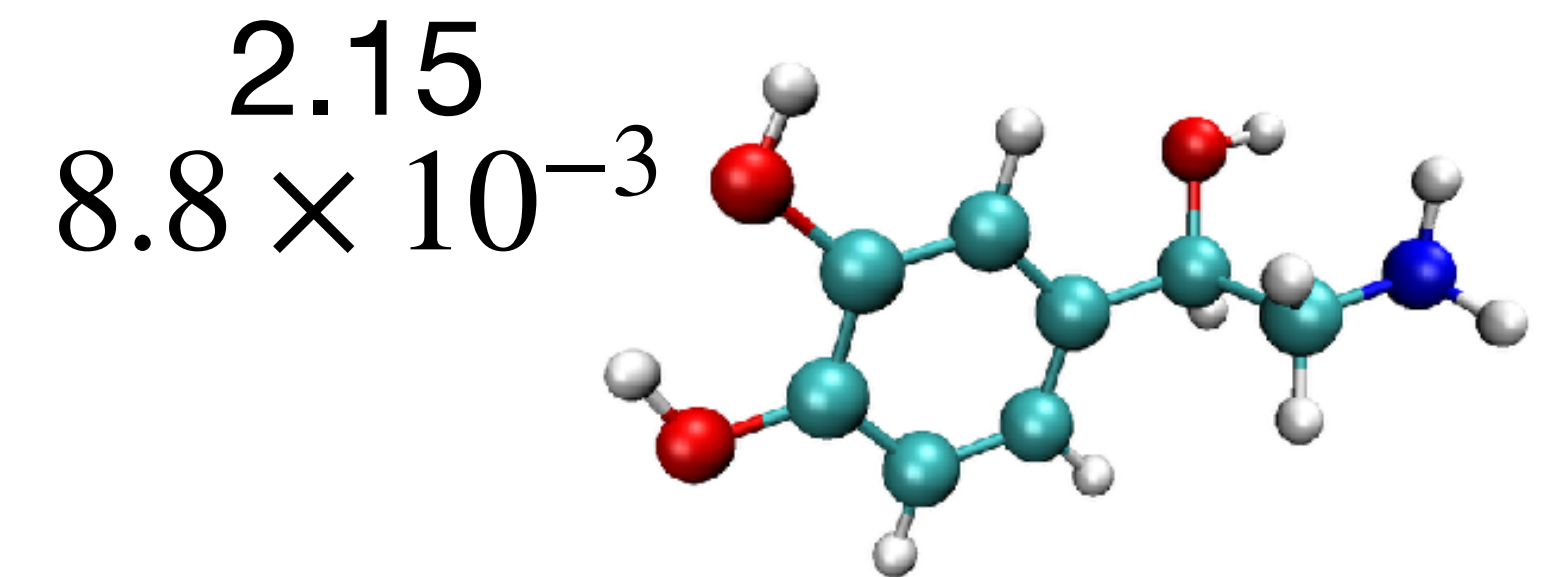
# Which Conformers are relevant?



Ludwig Boltzmann



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



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

# Which Conformers are relevant?



Ludwig Boltzmann

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

## **VALIDITY:**

- Gas phase
- All structures similar  
(Similar vibrations)

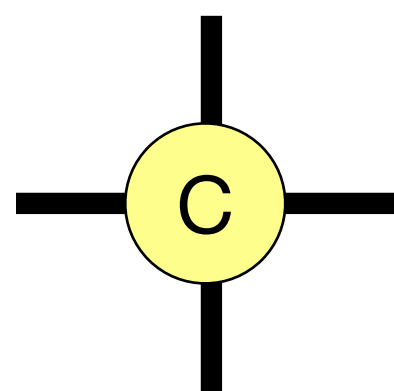
$$P_i = \frac{g_i \exp\left(-\frac{\Delta G_i}{k_B T}\right)}{\sum g_j \exp\left(-\frac{\Delta G_j}{k_B T}\right)}$$

## **VALIDITY:**

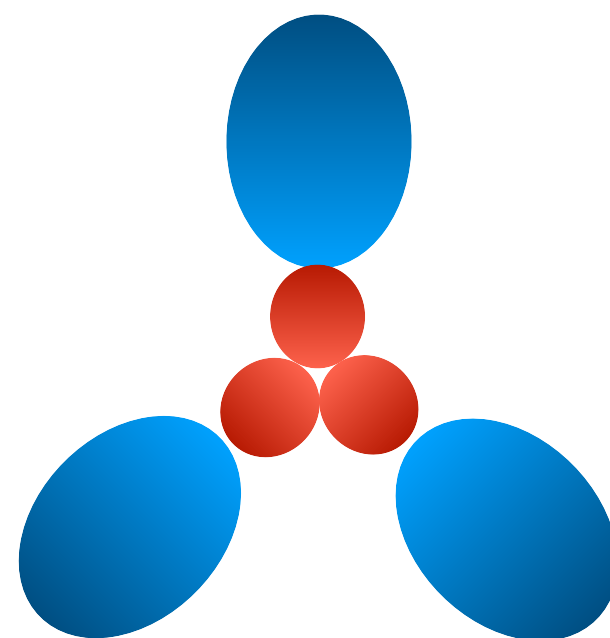
- Any phase
- Entropy is considered



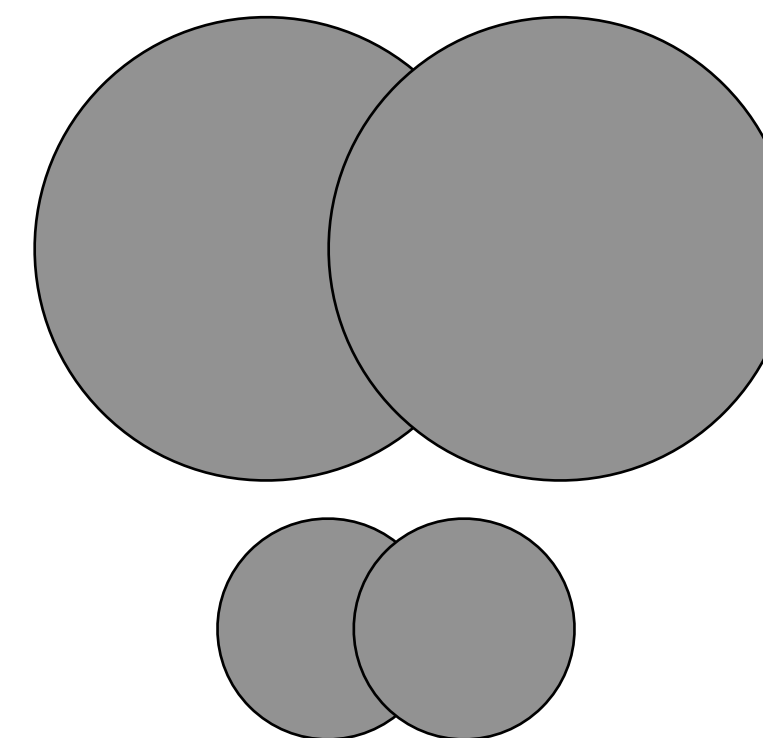
# Summary



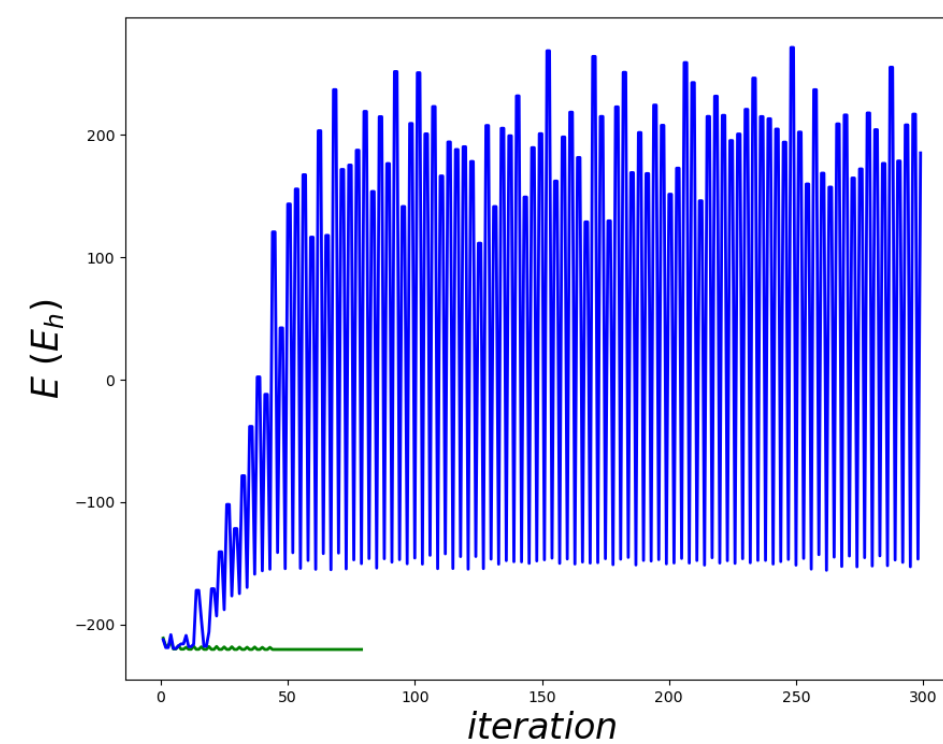
Octet Rule



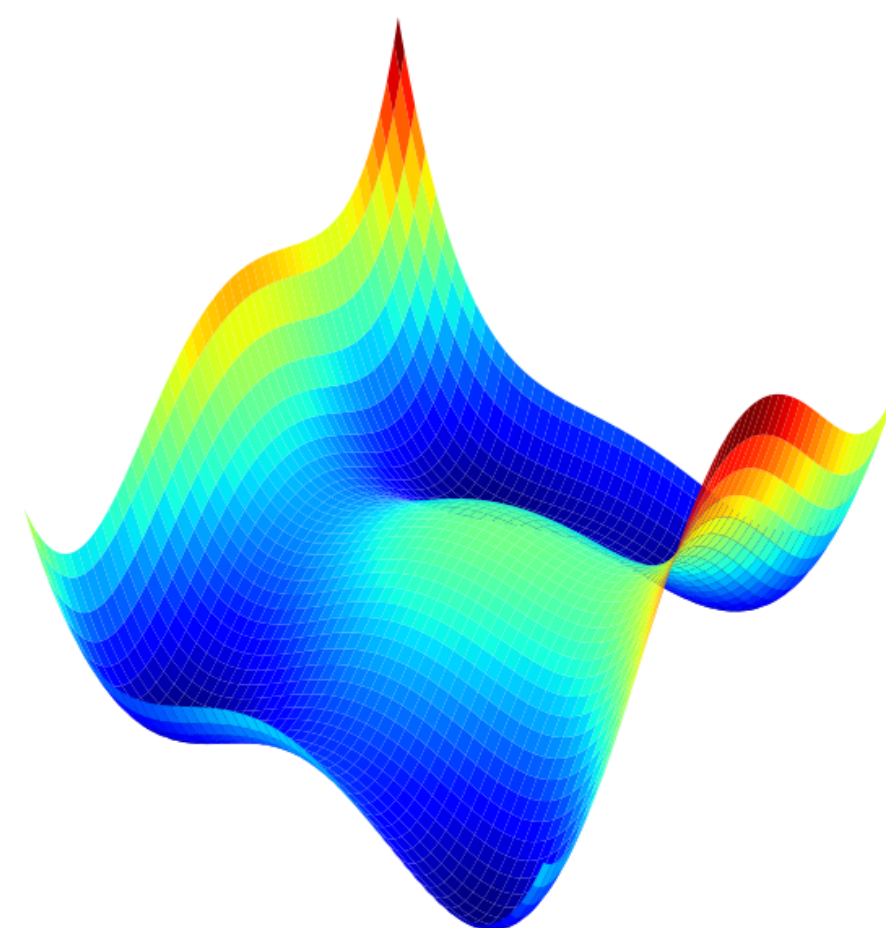
Hybridization



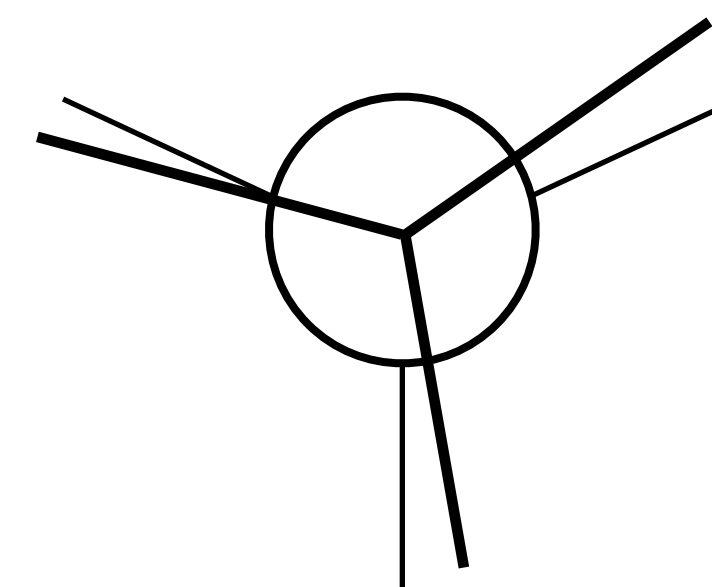
Hard-Soft



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

HELMHOLTZ  
MUNICH 

Institute of  
Structural Biology

Bayerisches NMR Zentrum 

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

*The End*

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



# Conformational Entropy and Ergodicity

$$\Delta S \geq 0$$

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

# Conformational Entropy and Ergodicity

