Revision 3

C2115
Practical introduction to supercomputing

Lesson 2

Petr Kulhánek

kulhanek@chemi.muni.cz

National Centre for Biomolecular Research, Faculty of Science
Masaryk University, Kamenice 5, CZ-62500 Brno

Content

- **Computational Chemistry Group**

  overview of solved projects

- **Model problems and systems for exercises**

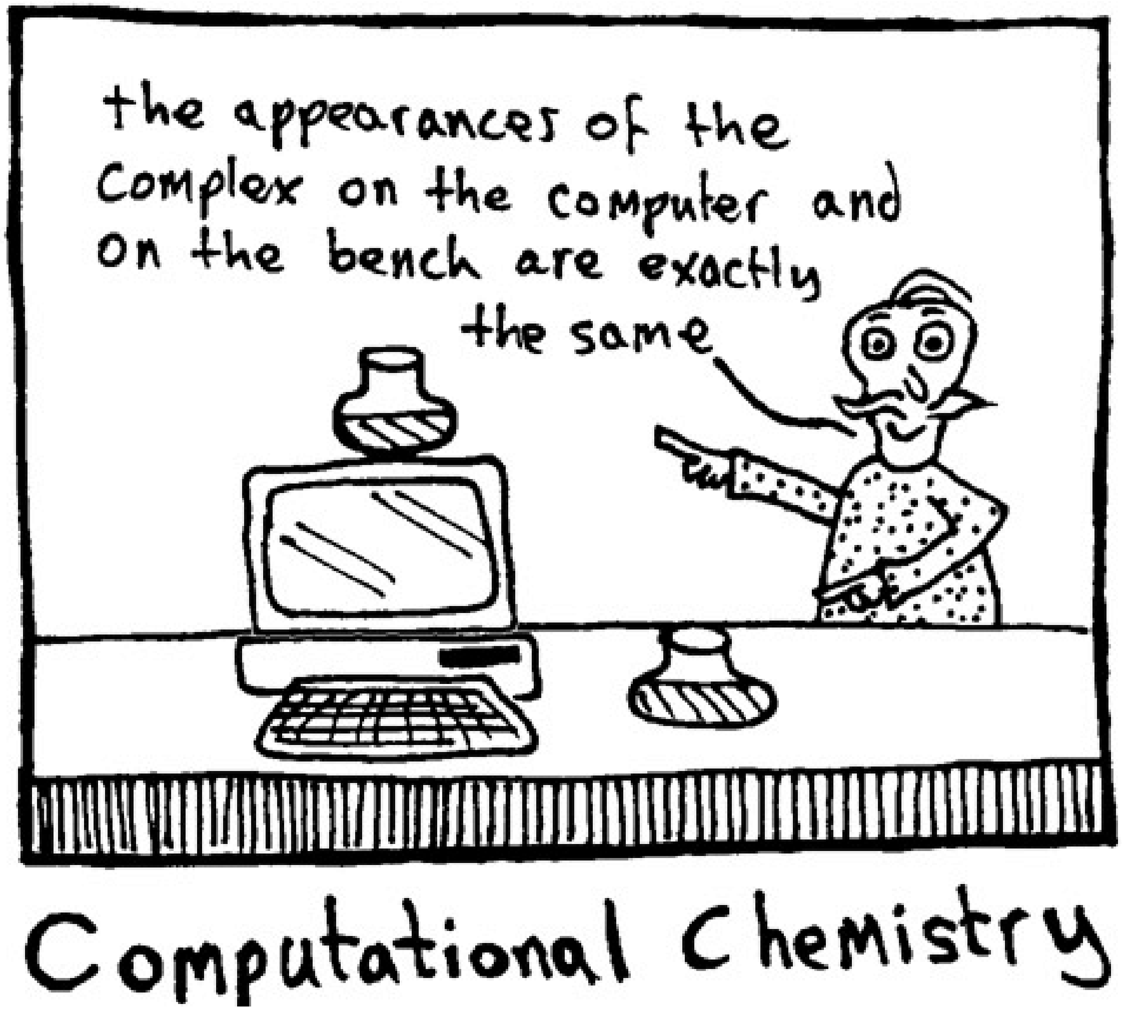
  matrix multiplication, numerical integration, QM and MD calculations

Computational chemistry group

overview of solved projects

group leader: prof. RNDr. Jaroslav Koča, DrSc.

Computational chemistry


the appearances of the Complex on the computer and on the bench are exactly the same
Computational Chemistry

http://www.ninger.com/images/comp.jpg

Computational chemistry

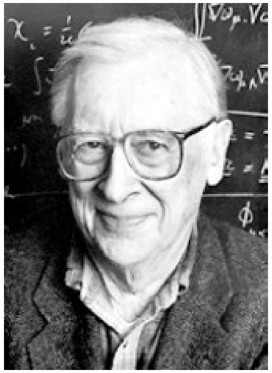
Computational chemistry (computer chemistry)

Computational chemistry is a branch of chemistry that uses **computer** simulation to assist in solving **chemical problems**. It uses methods of **theoretical chemistry**, incorporated into efficient **computer programs**, to calculate the structures and properties of molecules and solids. While computational results normally complement the information obtained by chemical experiments, it can in some cases predict hitherto unobserved chemical phenomena. It is widely used in **the design of new drugs** and **materials**.

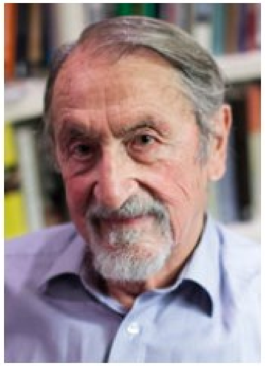
www.wikipedia.org

Nobel Prize in Chemistry 1998/2013

Walter Kohn



John A. Pople



© Harvard University
Martin Karplus

Photo: © S. Fisch
Michael Levitt

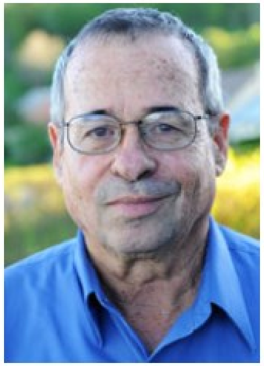


Photo: Wikimedia Commons
Arieh Warshel

The Nobel Prize in Chemistry 1998 was divided equally between

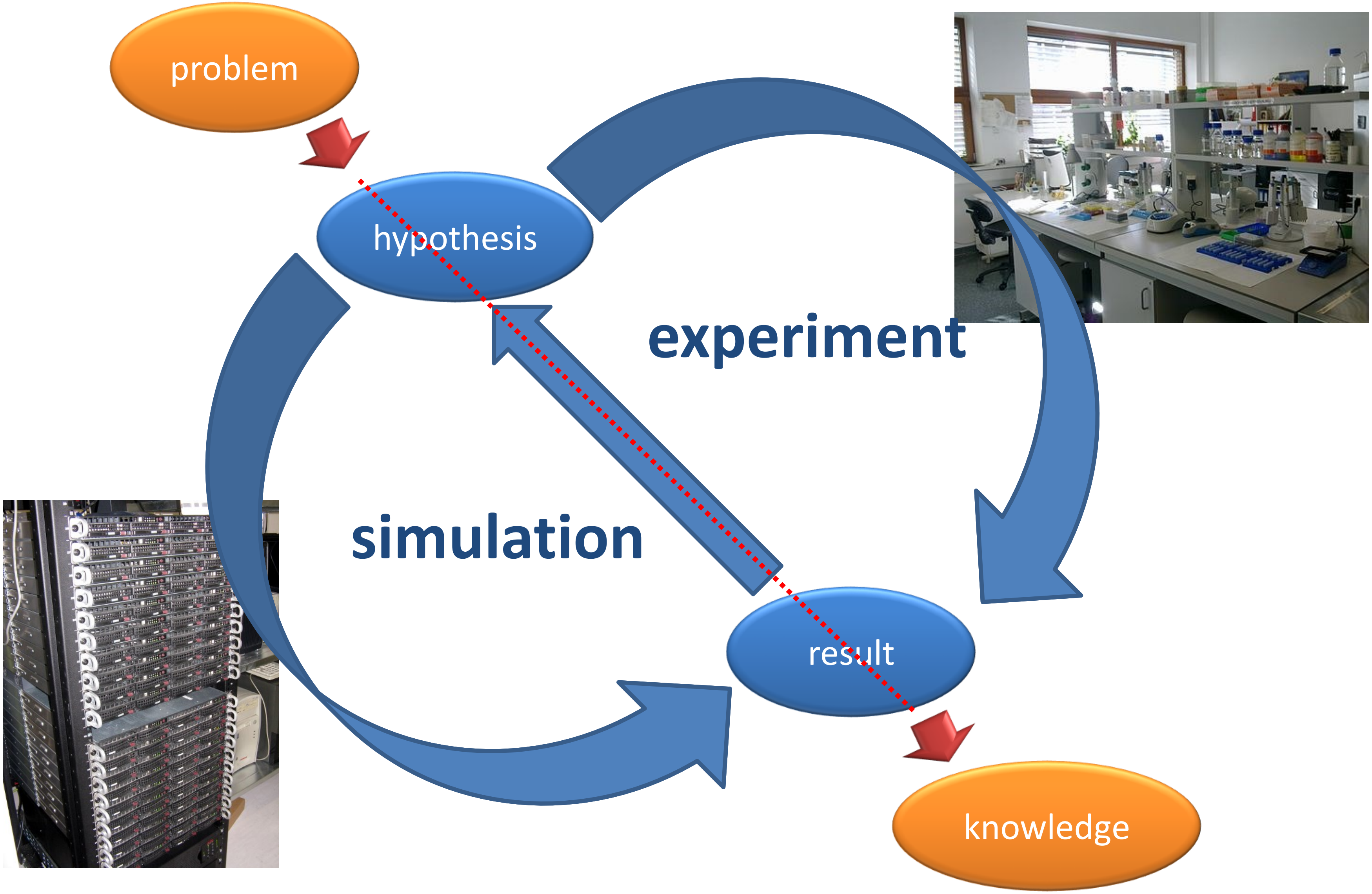
Walter Kohn "for his development of the **density-functional theory**“ and

John A. Pople "for his development of **computational methods in quantum chemistry**"

Development of Multiscale Models for Complex Chemical Systems

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1998/
http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/

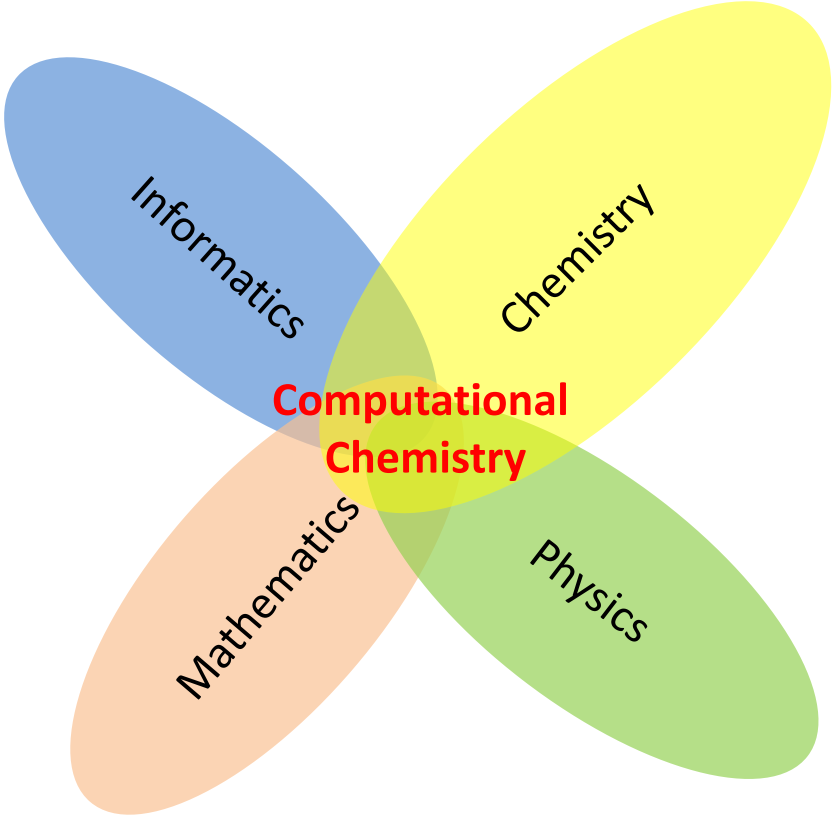
Experiment vs computational chemistry



Interdisciplinary field

algorithms, CPU/GPU,
cluster/grid,
symbolic calculations

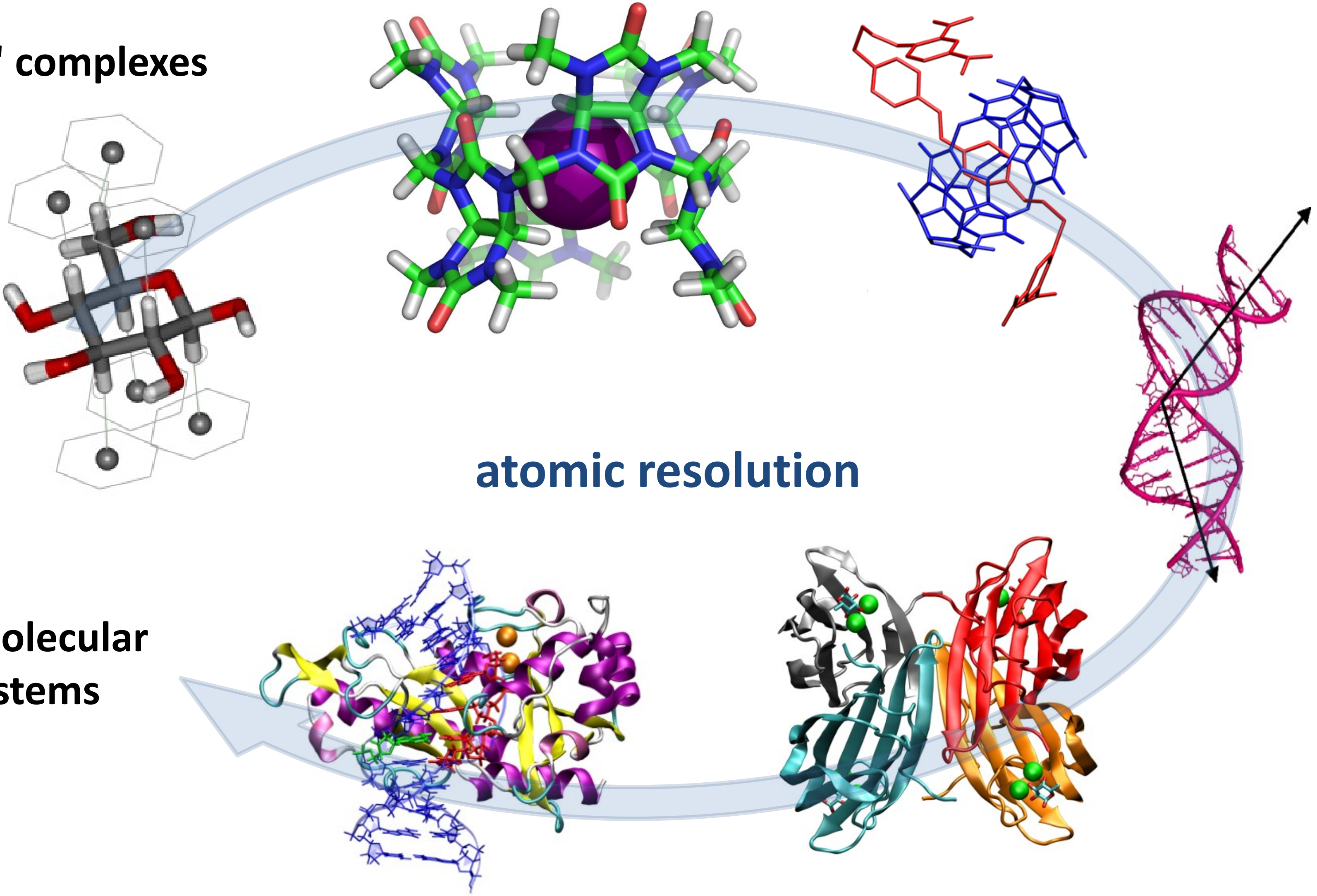
(bio)chemical problems,
experiments,
verification



theory, approximation

analytical solutions,
numerical sulutions,
aproximations

What we study...



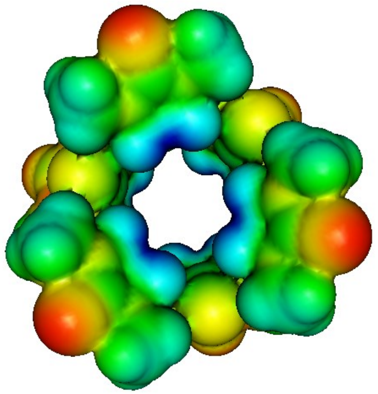
"small" complexes

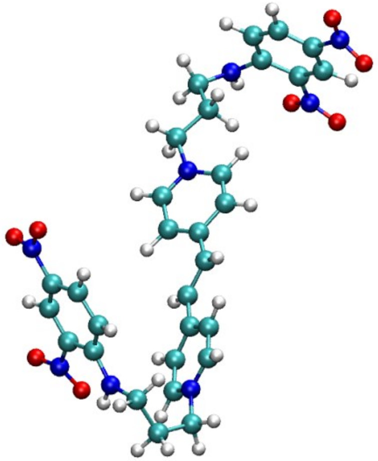
atomic resolution

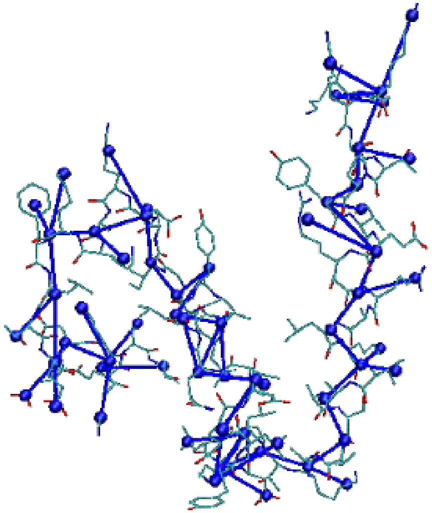
biomolecular systems

selected systems studied by the group of computational chemistry

Levels of theory



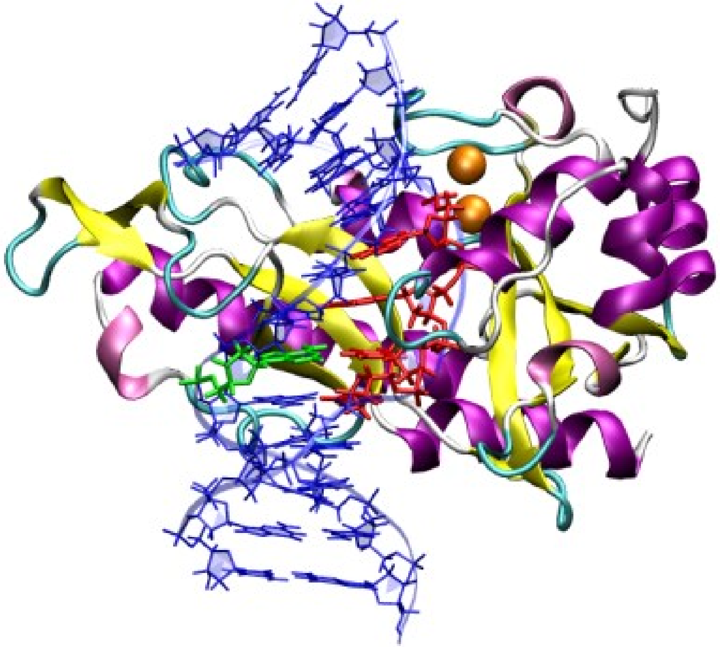




Quantum mechanics	Molecular mechanics	*Coarse-grained* mechanics
atomic resolution		bead resolution
reactivity	conformational movements	domain movement, folding
up to 1,000 atoms *	up to 1,000,000 atoms *	up to 1,000,000 beads *
up to 100 ps *	to 1 µs *	up to ms *

Projects

Study of (bio)molecular systems



Quantum chemical calculations

time independent Schrödinger equation

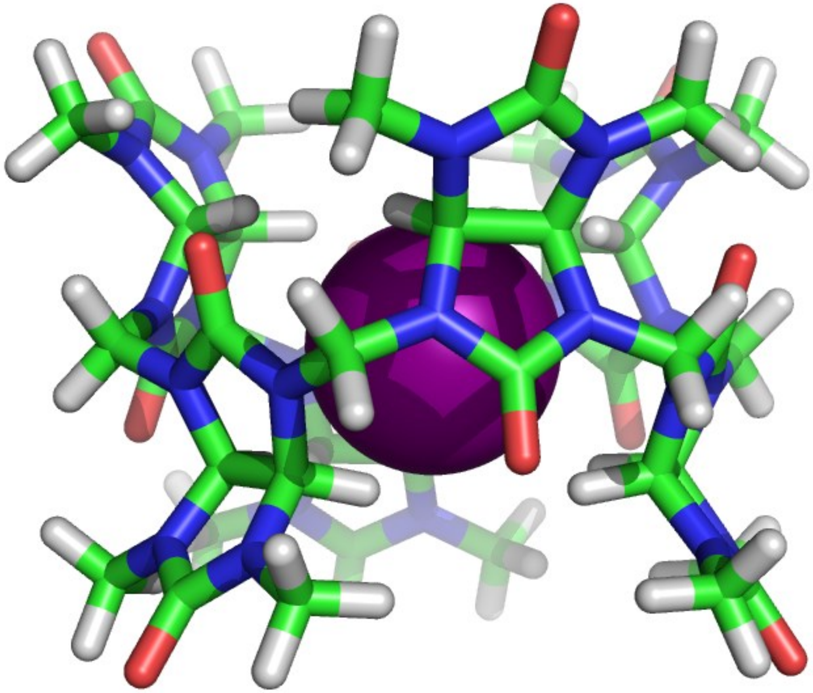
$$\hat{H}\psi_k(\mathbf{r}) = E_k\psi_k(\mathbf{r})$$

	Methods			
Formal scaling	**HF**	**CI methods**	**MP methods**	**CC methods**
N^4 -> N^2 -> N^1	HF, DFT			
N^5			MP2	CC2 (iterative)
N^6		CISD	MP3, MP4 (SDQ)	CCSD (iterative)
N^7			MP4	CCSD (T), CC3 (iterative)
N^8		CISDT	MP5	CCSDT
N^9			MP6	
N^{10}		CISDTQ	MP7	CCSDTQ (iterative)

Scaling, time demands: http://en.wikipedia.org/wiki/Time_complexity
HF - Hartree – Fock method, DFT - theory functional density,
CI - methods of configuration interaction, MP - Møllerova-Plesset perturbation theory,
CC - method of bound clusters, N - number of basis functions

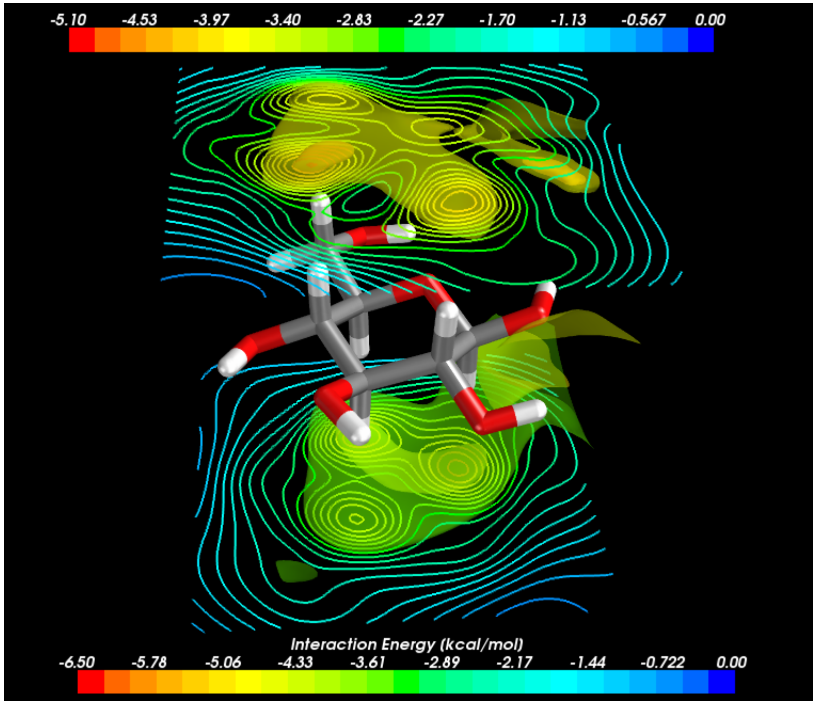
Jensen, F. Introduction to computational chemistry; 2nd ed .; John Wiley & Sons:Chichester, England; Hoboken, NJ, 2007.

Quantum chemical calculations



supramolecular complexes

carbohydrate binding capacity



Molecular mechanics

Schrödinger equation => quantum mechanical view

$$H_a\Psi(r) = \boxed{E(R)}\Psi(r)$$

approximations using classical physics
no explicit electron motion is considered
(movement is implicitly included in empirical parameters)

$$E(R) = \underbrace{E_{bonds} + E_{angles} + E_{torsions}}_{\text{covalent contributions}} + \underbrace{E_{el} + E_{vdw}}_{\text{non-covalent contributions}}$$

Classical physics => mechanical view

Formal scaling: N^2 -> N $\log_2$N
N - number of atoms

Moleculer dynamics

$$-\frac{\partial E(\mathbf{R})}{\partial \mathbf{R}} = \mathbf{F} \qquad \boxed{\mathbf{F}_i = m_i \mathbf{a}_i} \qquad \mathbf{a}_i = \frac{d^2 \mathbf{r}_i}{dt^2}$$

2nd Newton's law of motion (law of force)

$$-\frac{\partial E(\mathbf{R})}{\partial \mathbf{R}} = m_i \frac{d^2 \mathbf{r}_i}{dt^2}$$

system of second order differential equations requires a numerical solution

↓

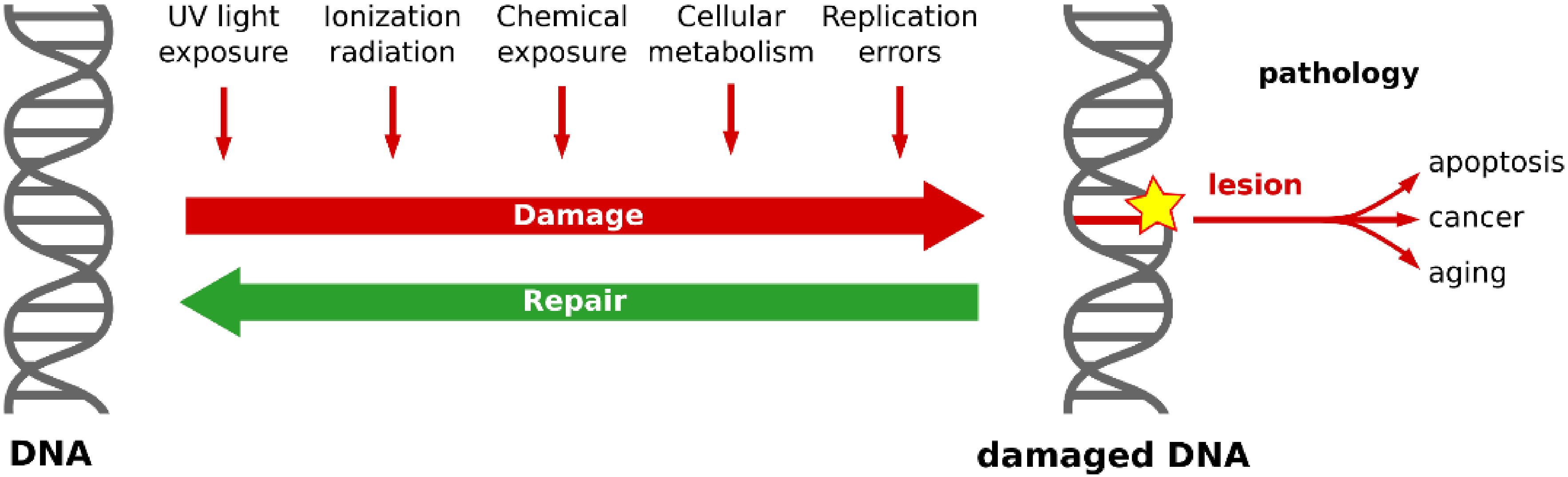
discretization of molecular motion in short time intervals

given by the fastest movement (vibration of bonds) → **1 fs** **typical integration step**

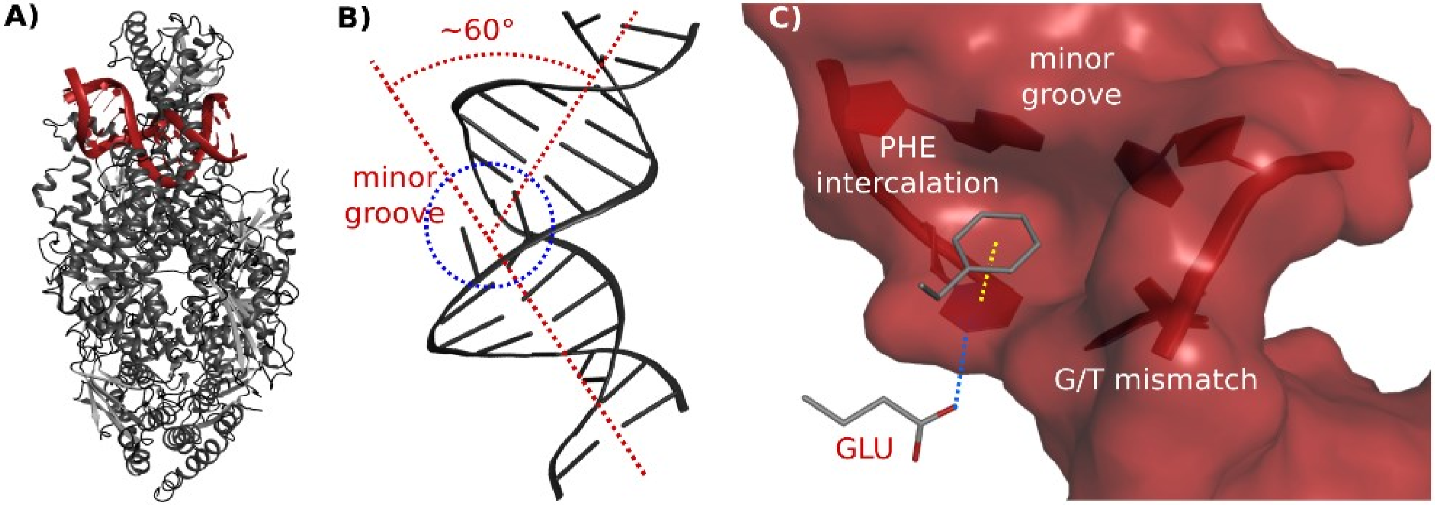
Imperfections in integration are removed by use of **thermostats** and **barostats**, which also provide the required simulation conditions.

Repair of damaged DNA

DNA is exposed to a number of factors that damage it. To avoid degradation of genetic information, damaged DNA is repaired by a number of mechanisms that work with different efficiencies. The aim of the project is to understand the method of detecting damage at the molecular level with a primary focus on the mechanical properties of damaged DNA.

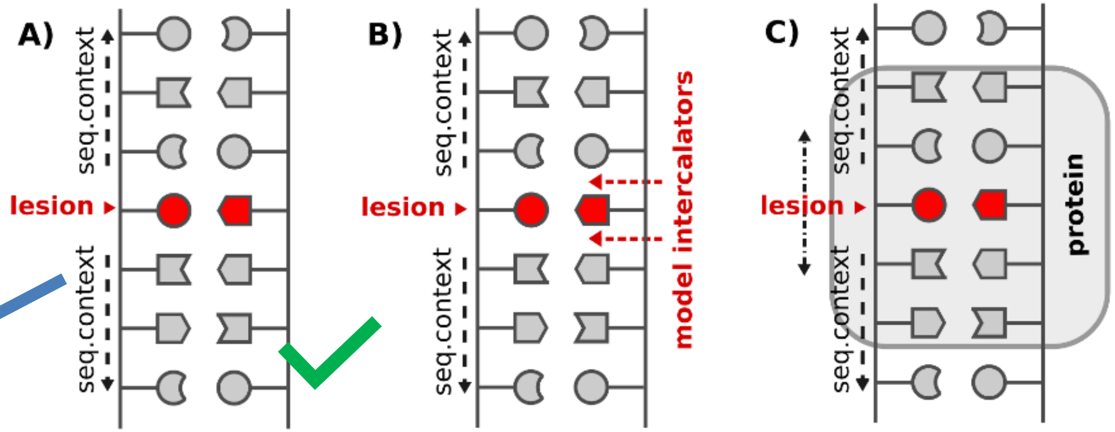


MMR - Mismatch Repair



Methods:

- **molecular dynamics**
- **free energy calculations**
- **quantum chemical calculations**



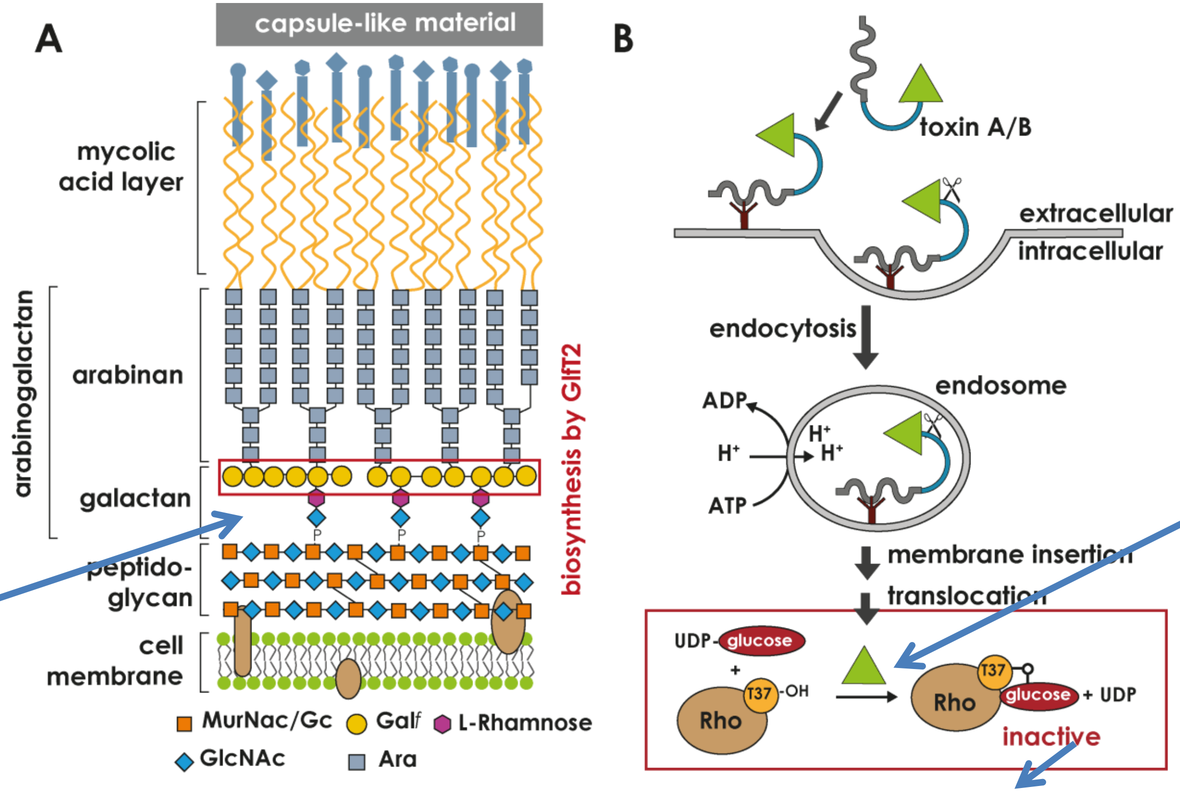
Bouchal, T .; Durník, I .; Illík, V .; Réblová, K .; Kulhánek, P. Importance of Base-Pair Opening for Mismatch Recognition. *Nucleic Acids Res.* **2020**. https://doi.org/10.1093/nar/gkaa896.

Glycosyltransferases

Glycosyltransferases are enzymes that **catalyze the transfer of activated sugar moiety** to oligosaccharides, proteins or other biomolecules. They are important in post-translational modification of proteins, regulation or creation of structural support.

Mycobacterium tuberculosis
(pathogenic bacteria)

Clostridium difficile
(pathogenic bacteria)

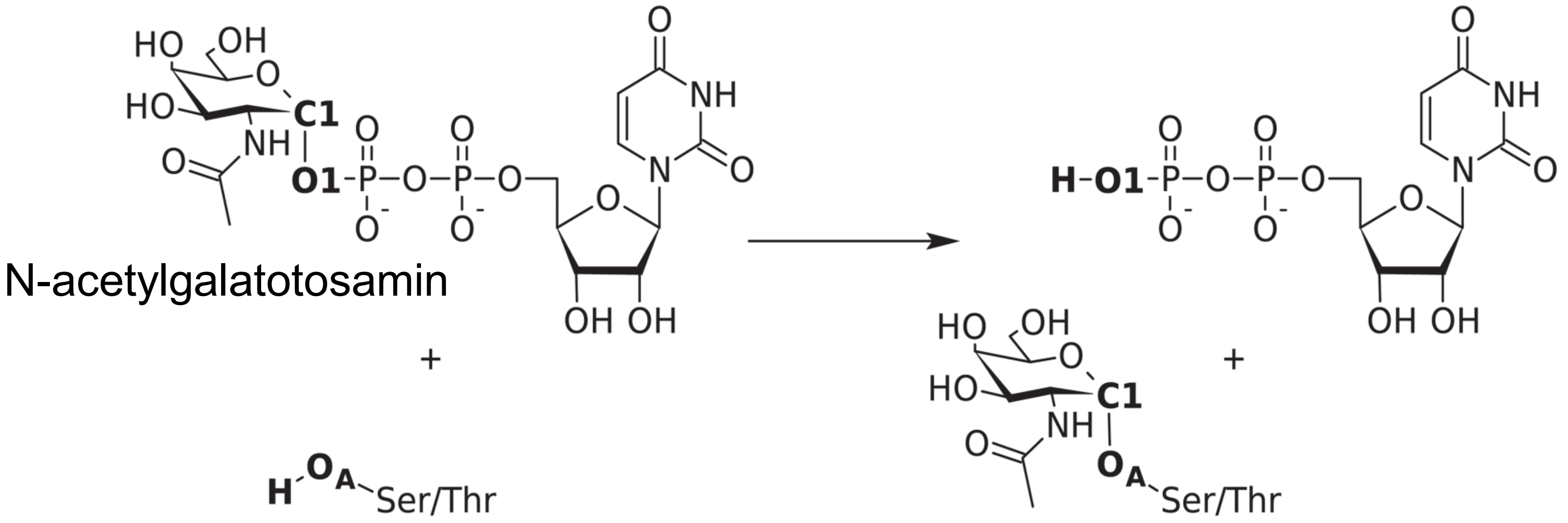


Motivation:
inhibitor of important membrane component synthesis
-> **antibiotic**

Motivation:
inhibitor of glycosyl-transferase toxin activity
-> **antidote**

cell death

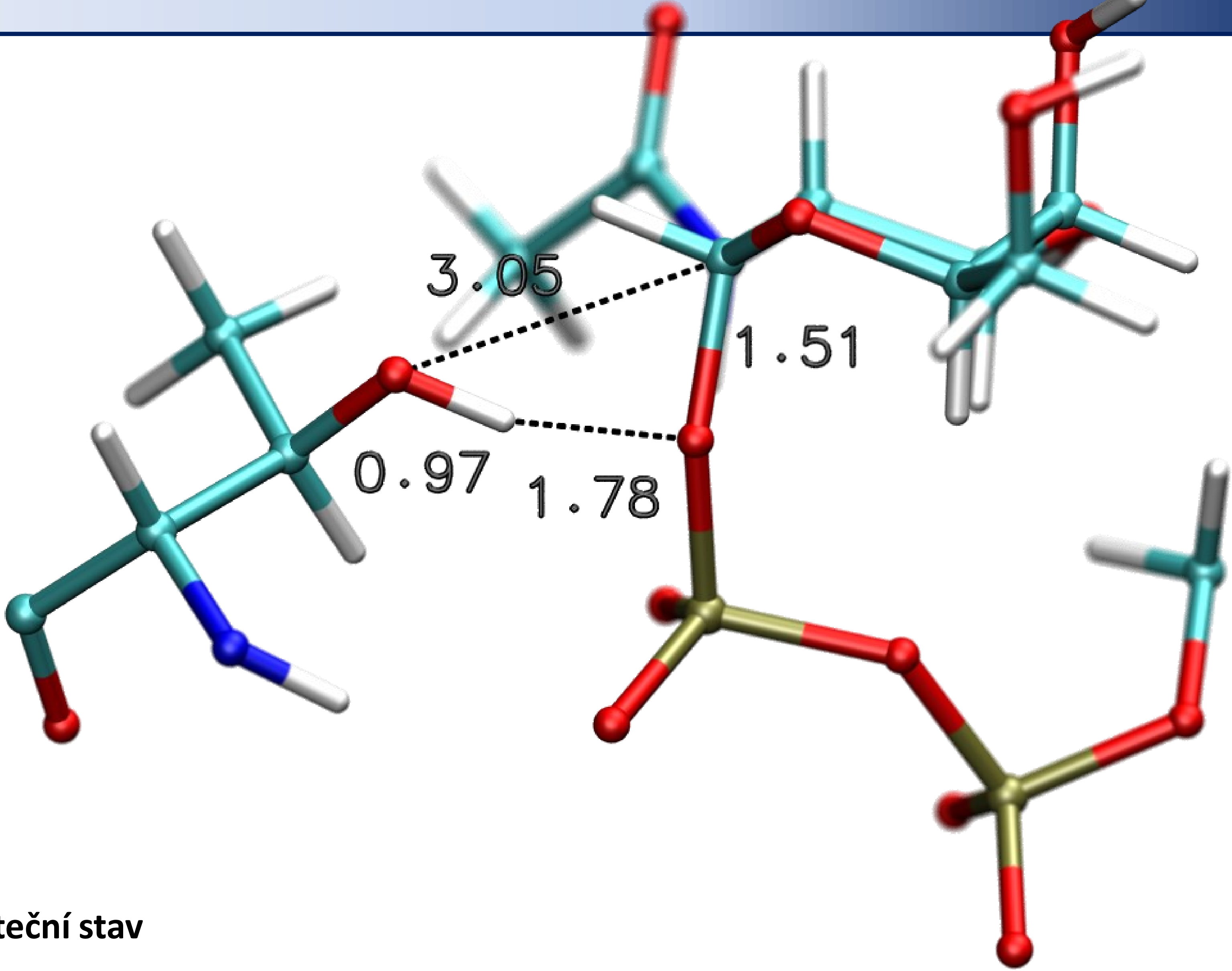
Example ppGalNAcT2 (QM/MM)



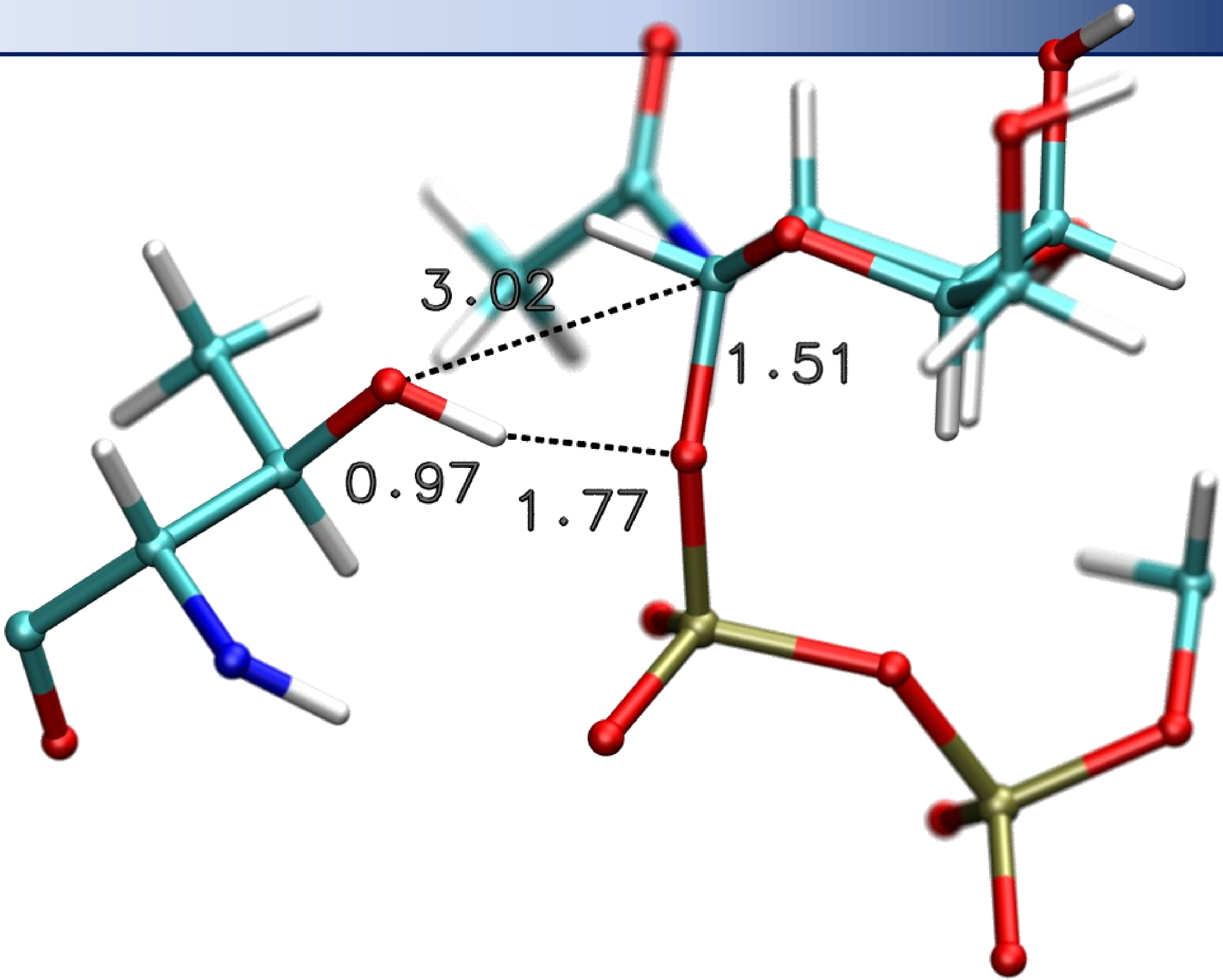
Supervisors or consultants:

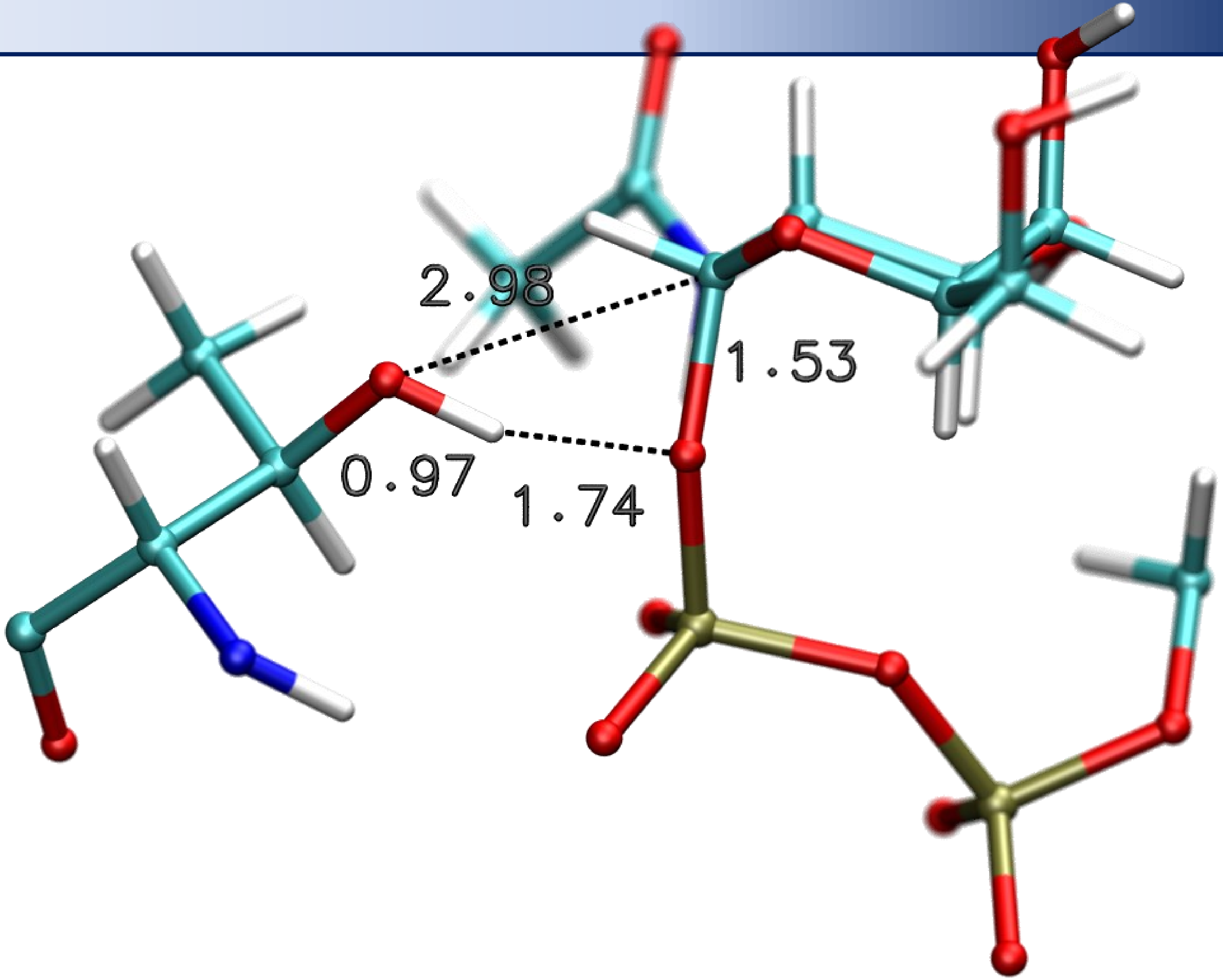
- prof. RNDr. Jaroslav Koča, DrSc.
  (Computational Chemistry - Center for Structural Biology - Central European Institute of Technology)
- Mgr. Stanislav Kozmon, Ph.D.
  (Institute of Chemistry, Slovak Academy of Sciences)
- Ing. Igor Tvaroška, DrSc.
  (Institute of Chemistry, Slovak Academy of Sciences)

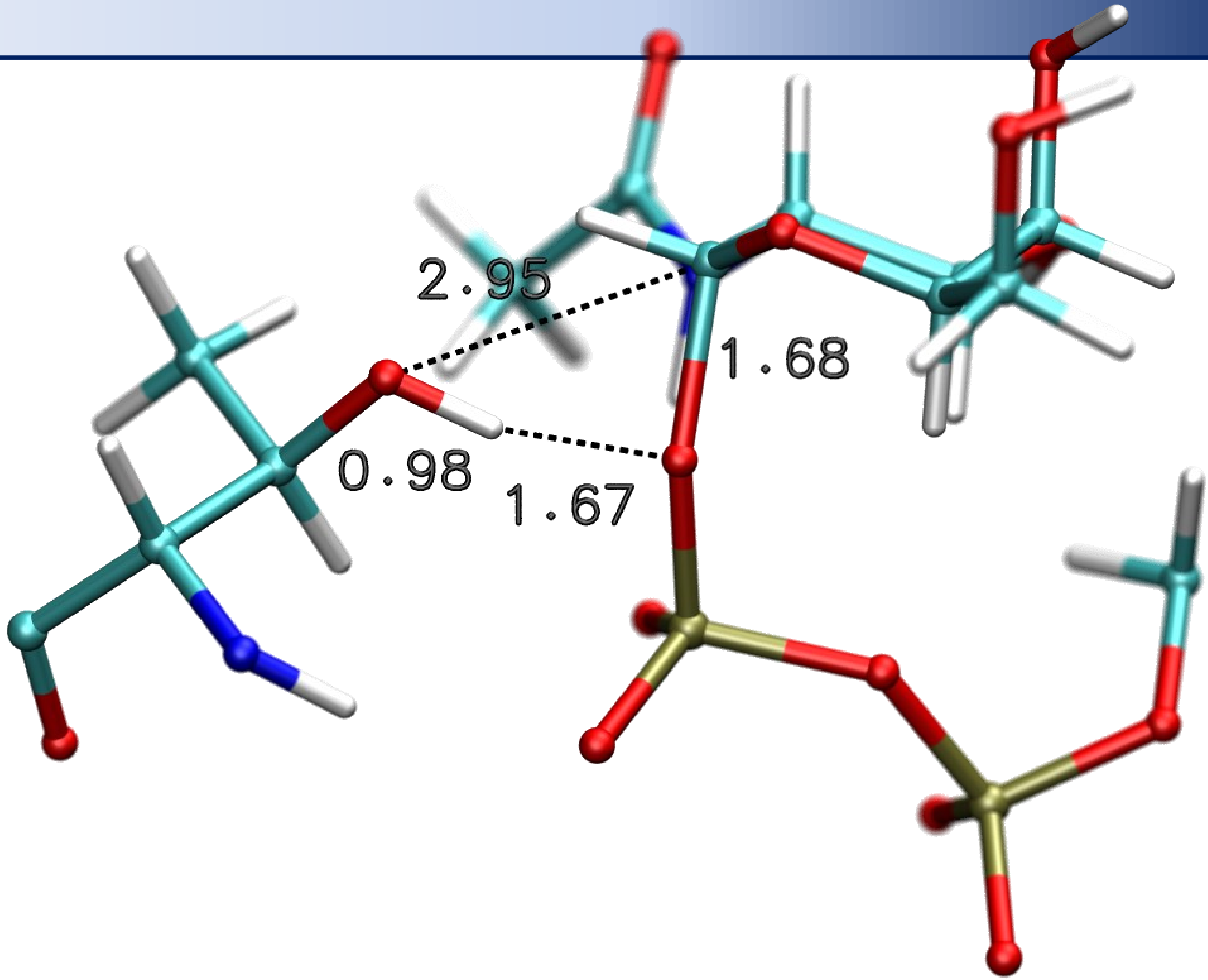
Janoš, P.; Trnka, T.; Kozmon, S.; Tvaroška, I.; Koča, J. Different QM/MM Approaches To Elucidate Enzymatic Reactions: Case Study on ppGalNAcT2. *J. Chem. Theory Comput.* **2016**, *12* (12), 6062–6076. https://doi.org/10.1021/acs.jctc.6b00531.

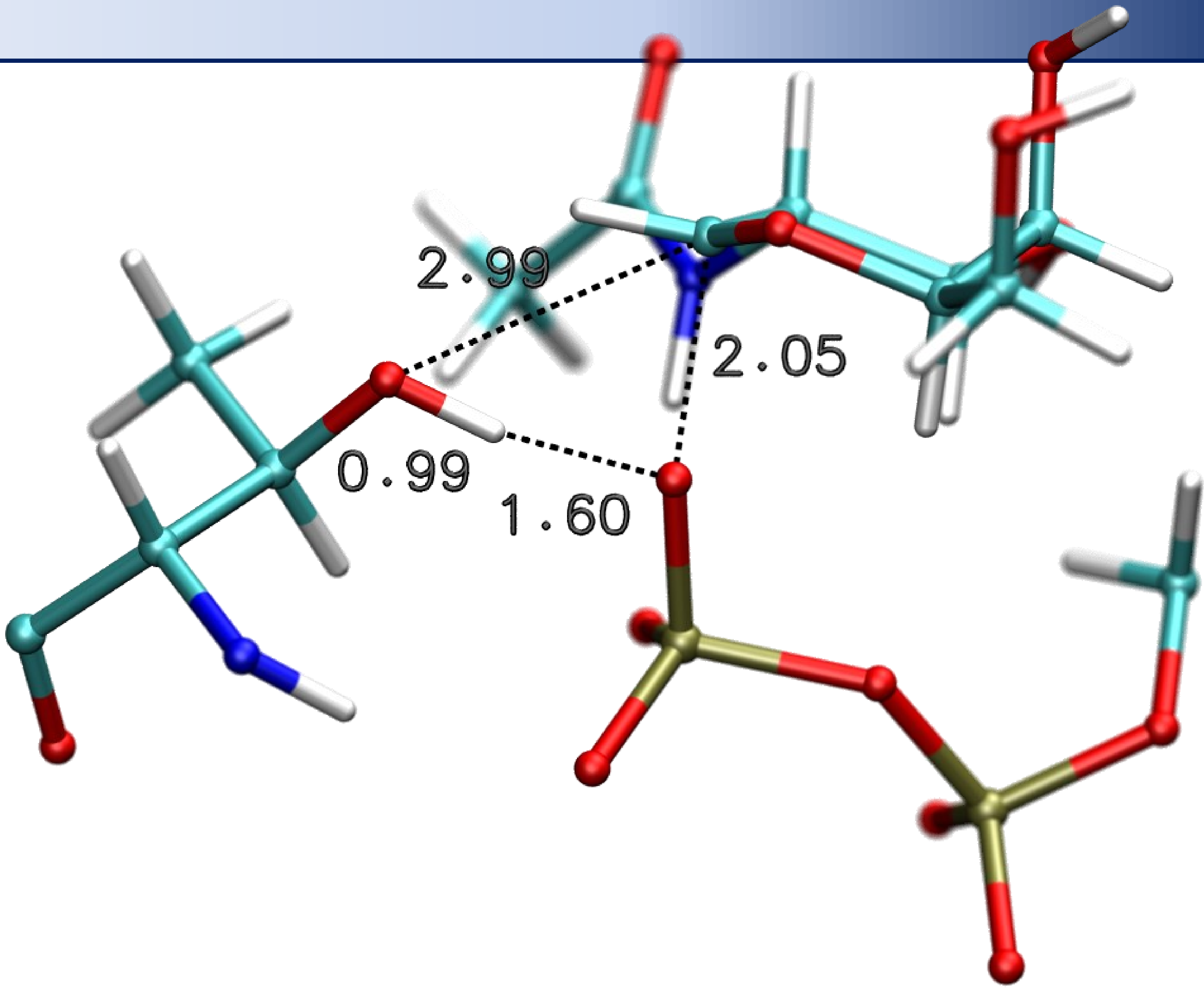


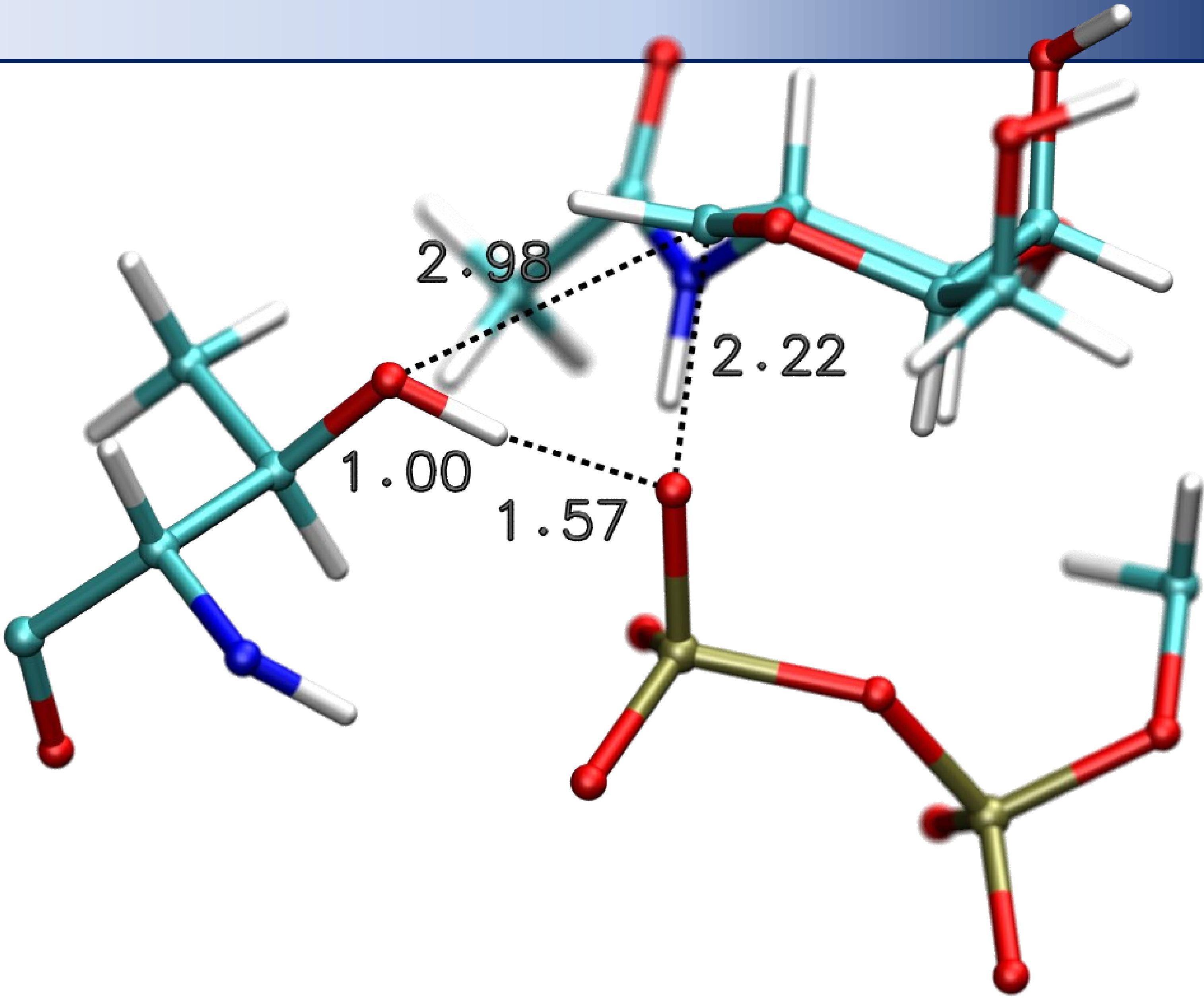
počáteční stav

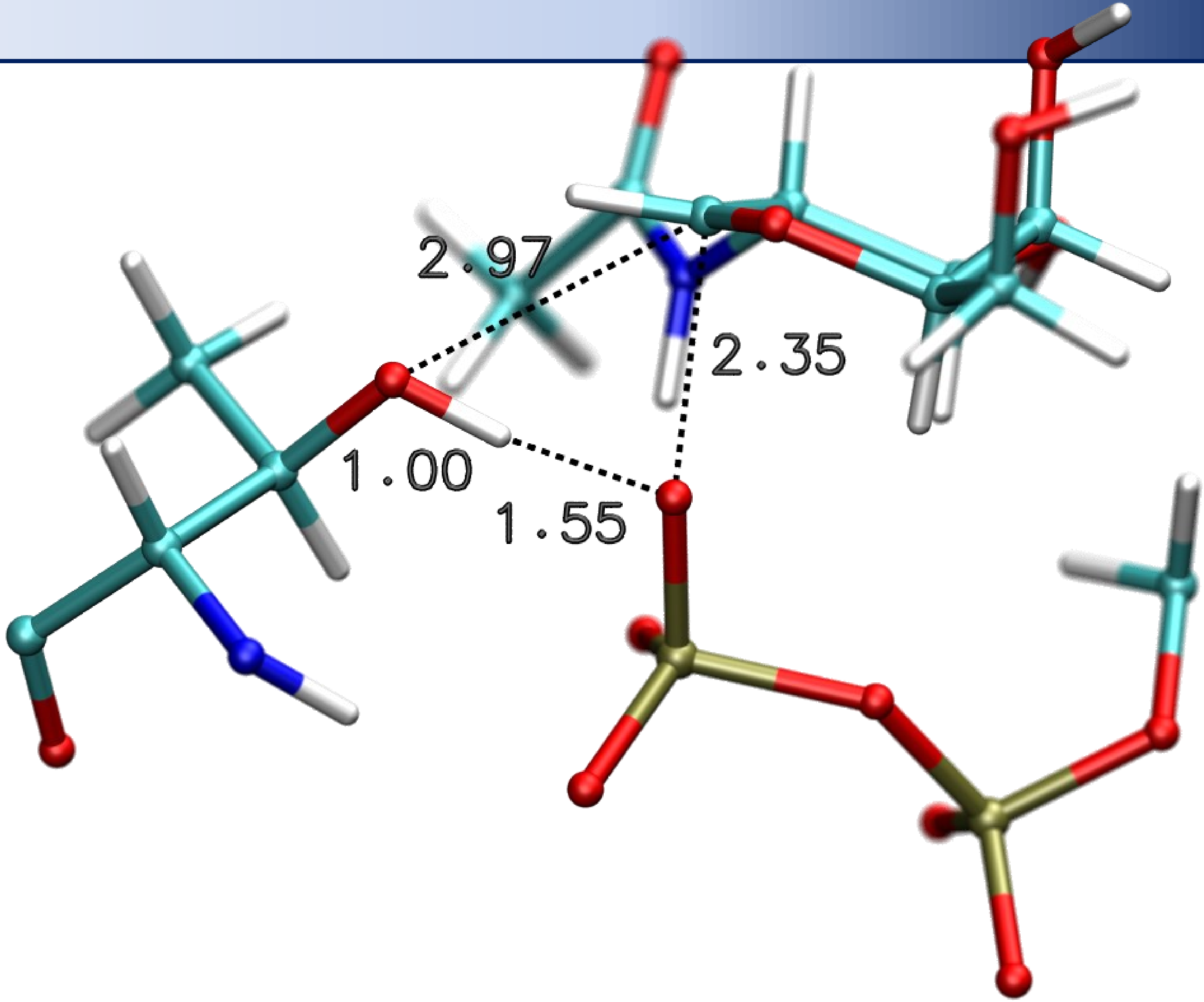

3.02
1.51
0.97
1.77

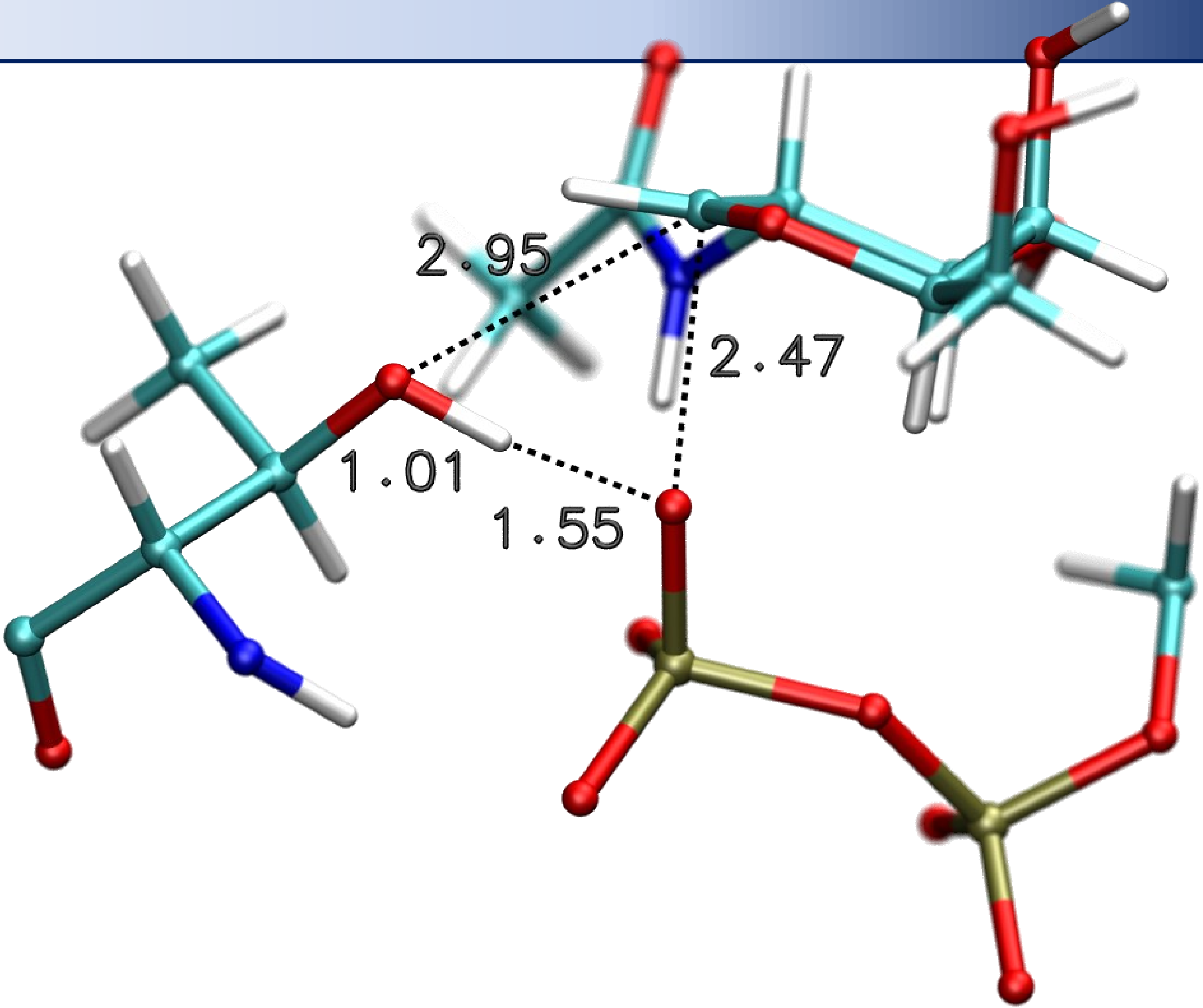

2.98
1.53
0.97
1.74

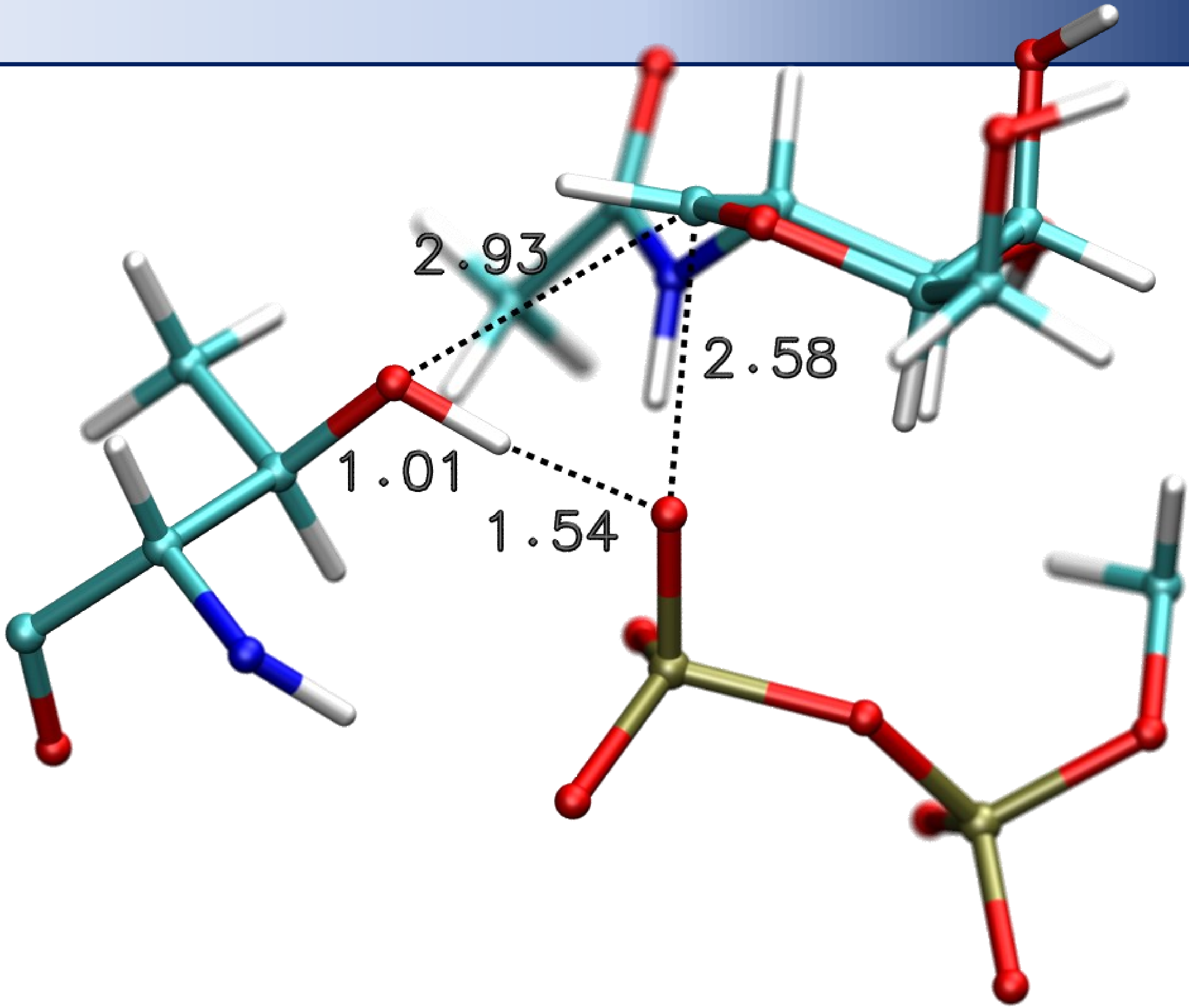

2.95
1.68
0.98
1.67

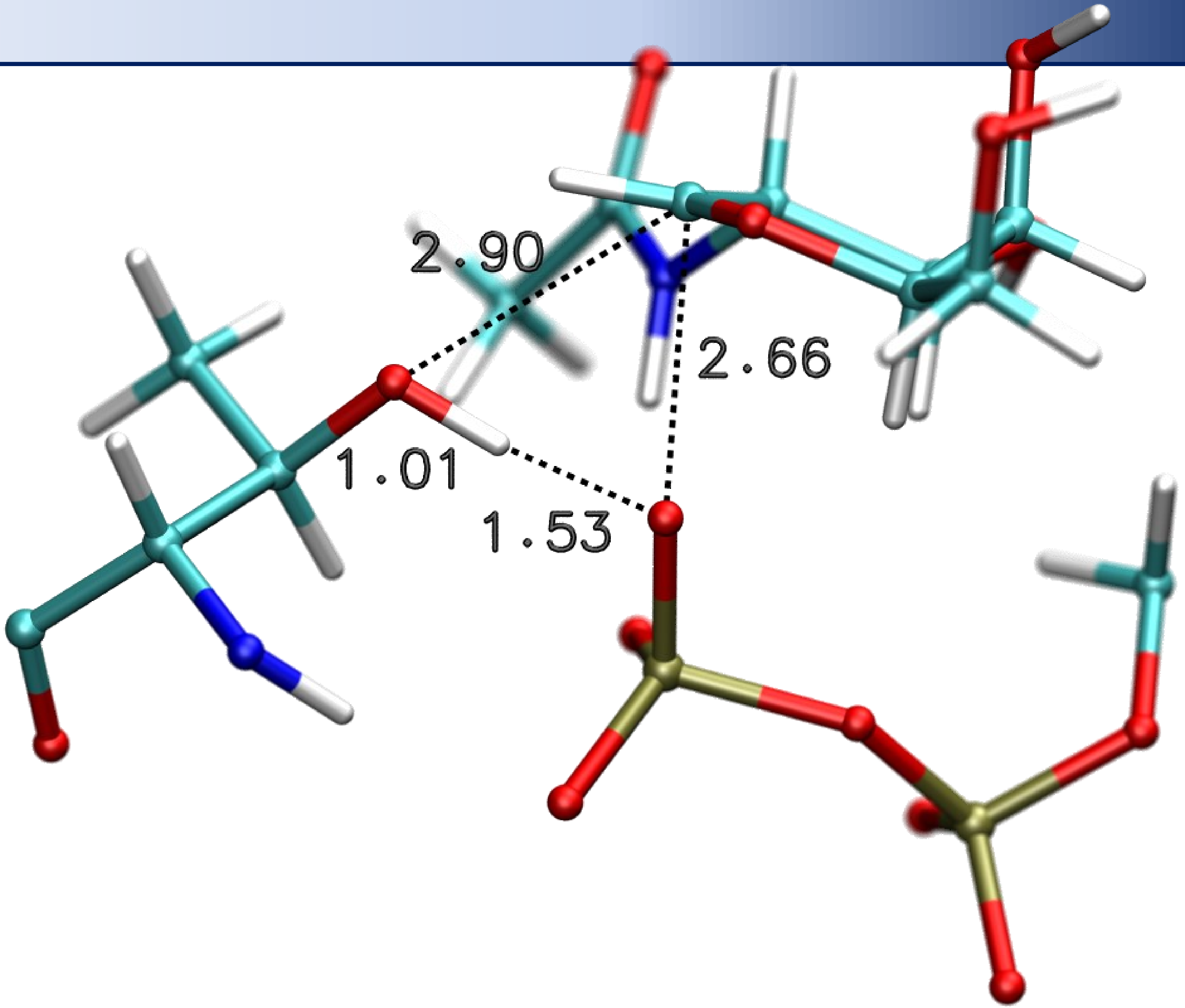

2.99
2.05
0.99
1.60


2.98
2.22
1.00
1.57

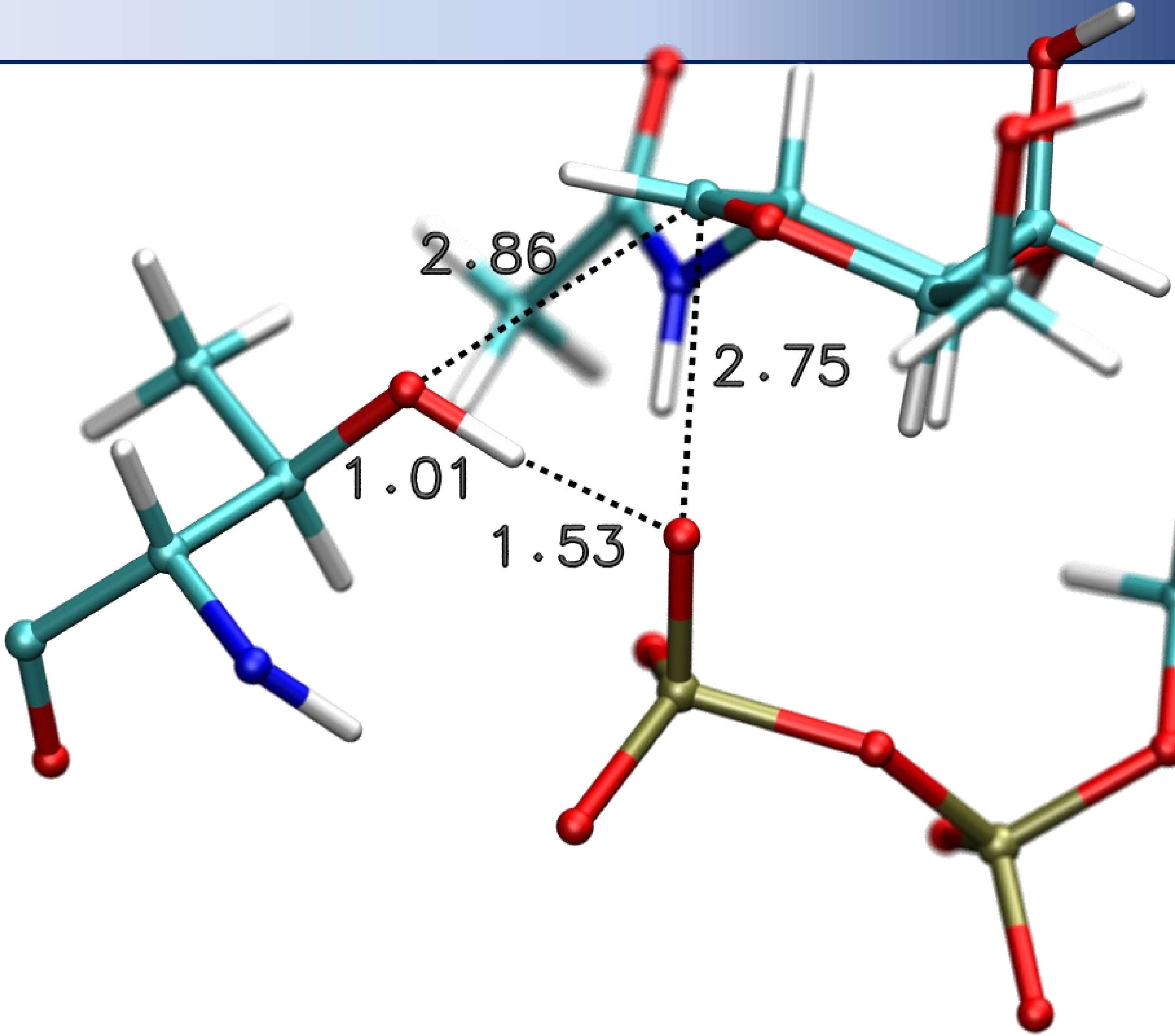

2.97
2.35
1.00
1.55

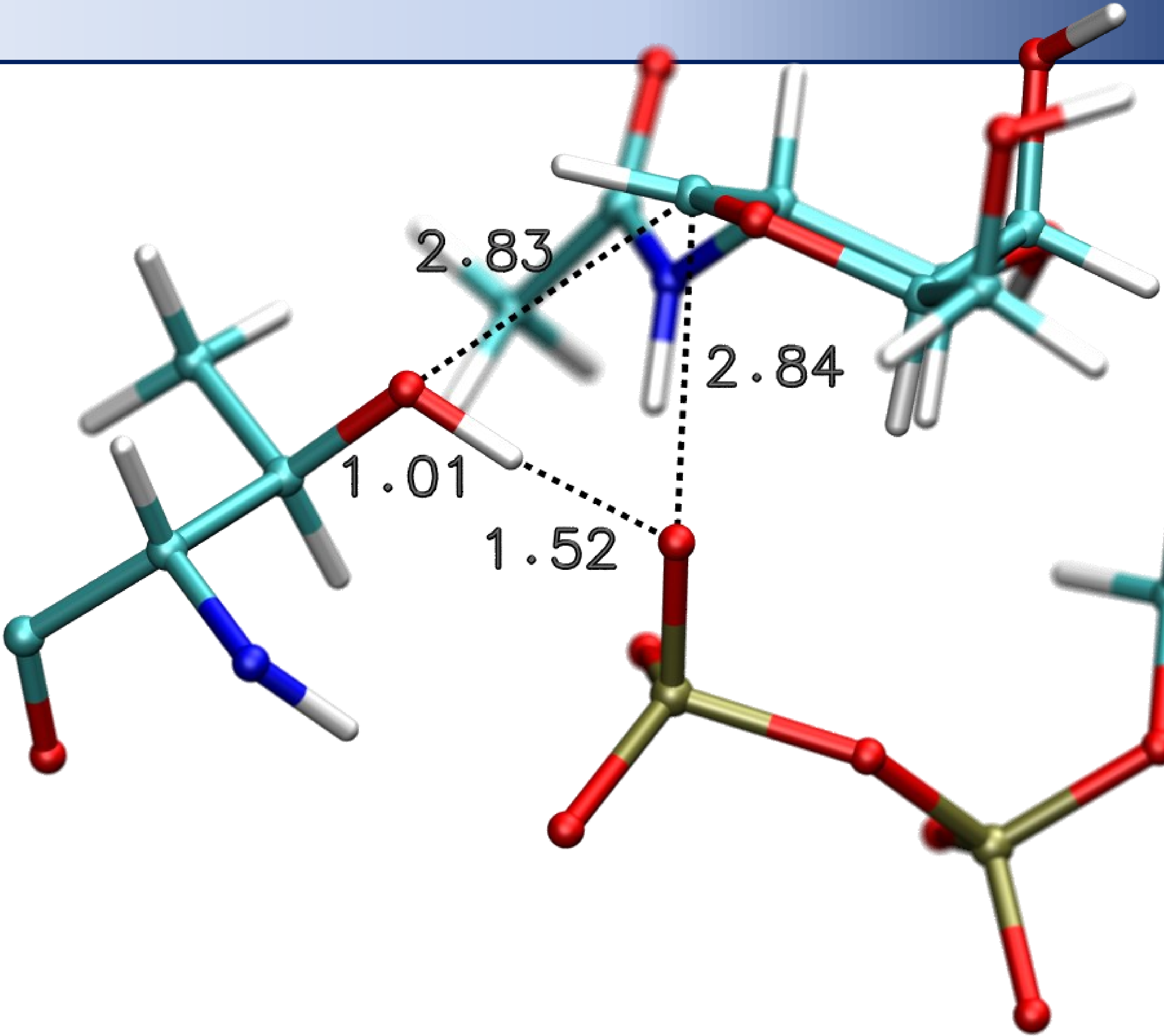

2.95
2.47
1.01
1.55

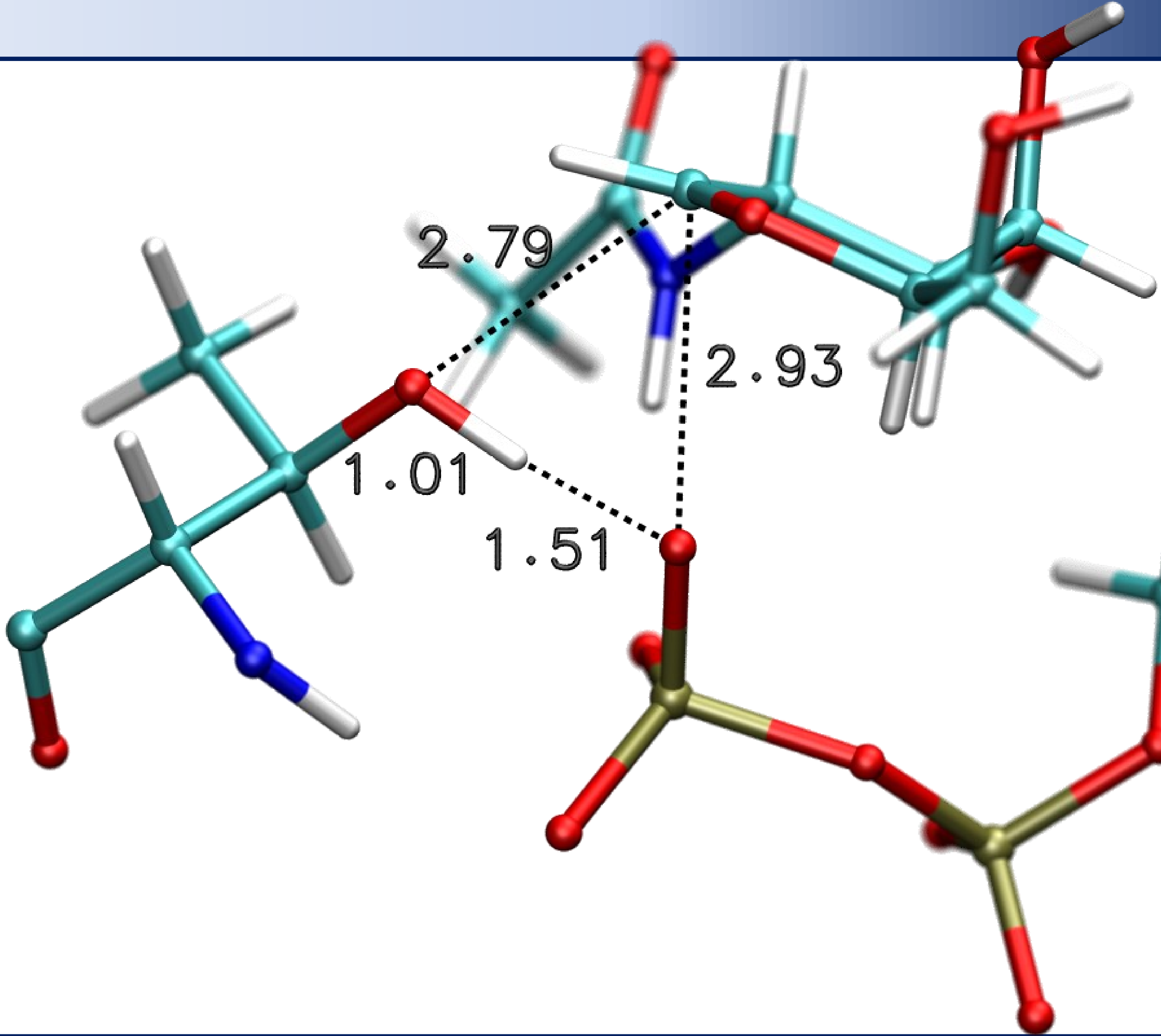

2.93
2.58
1.01
1.54

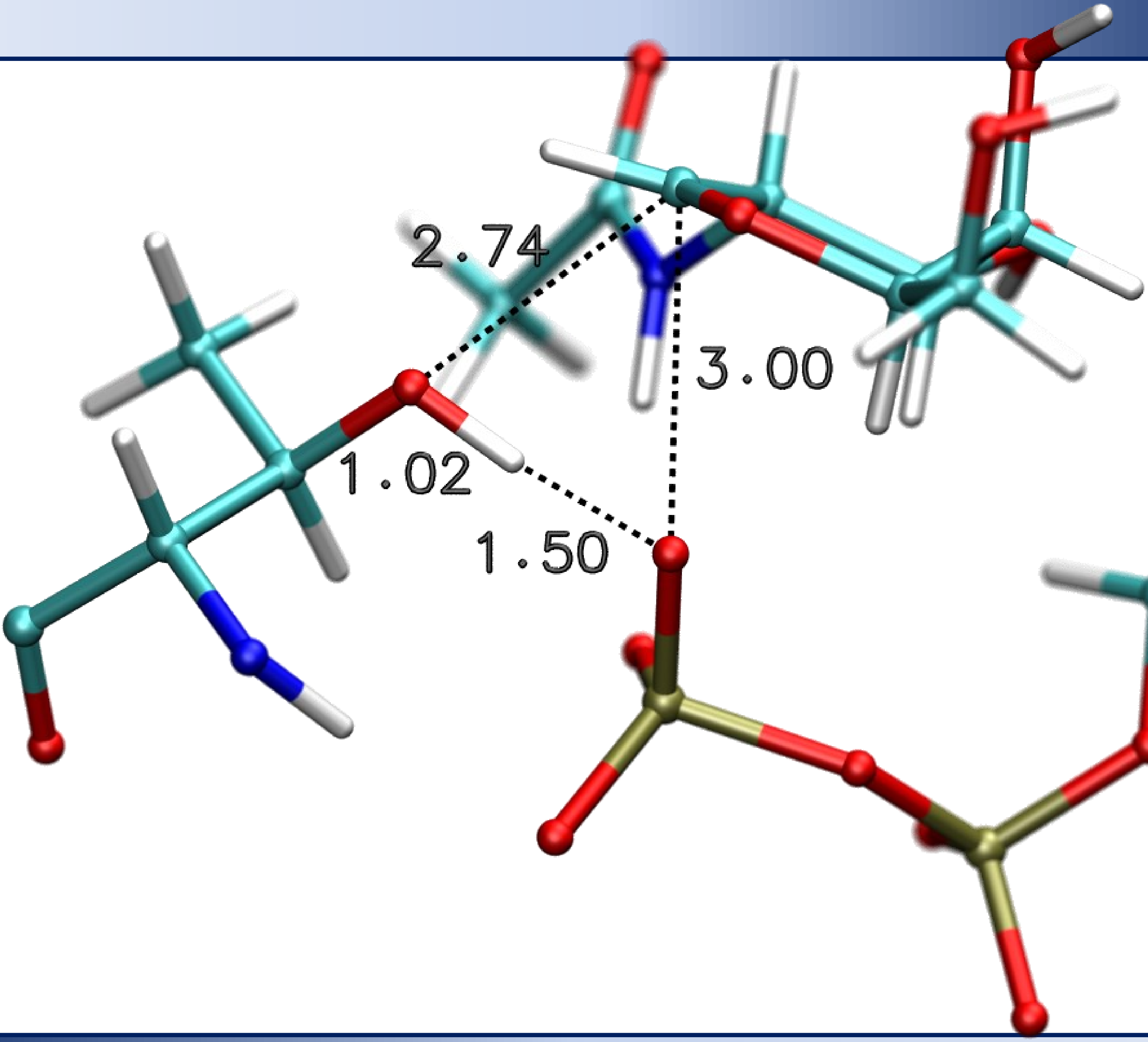


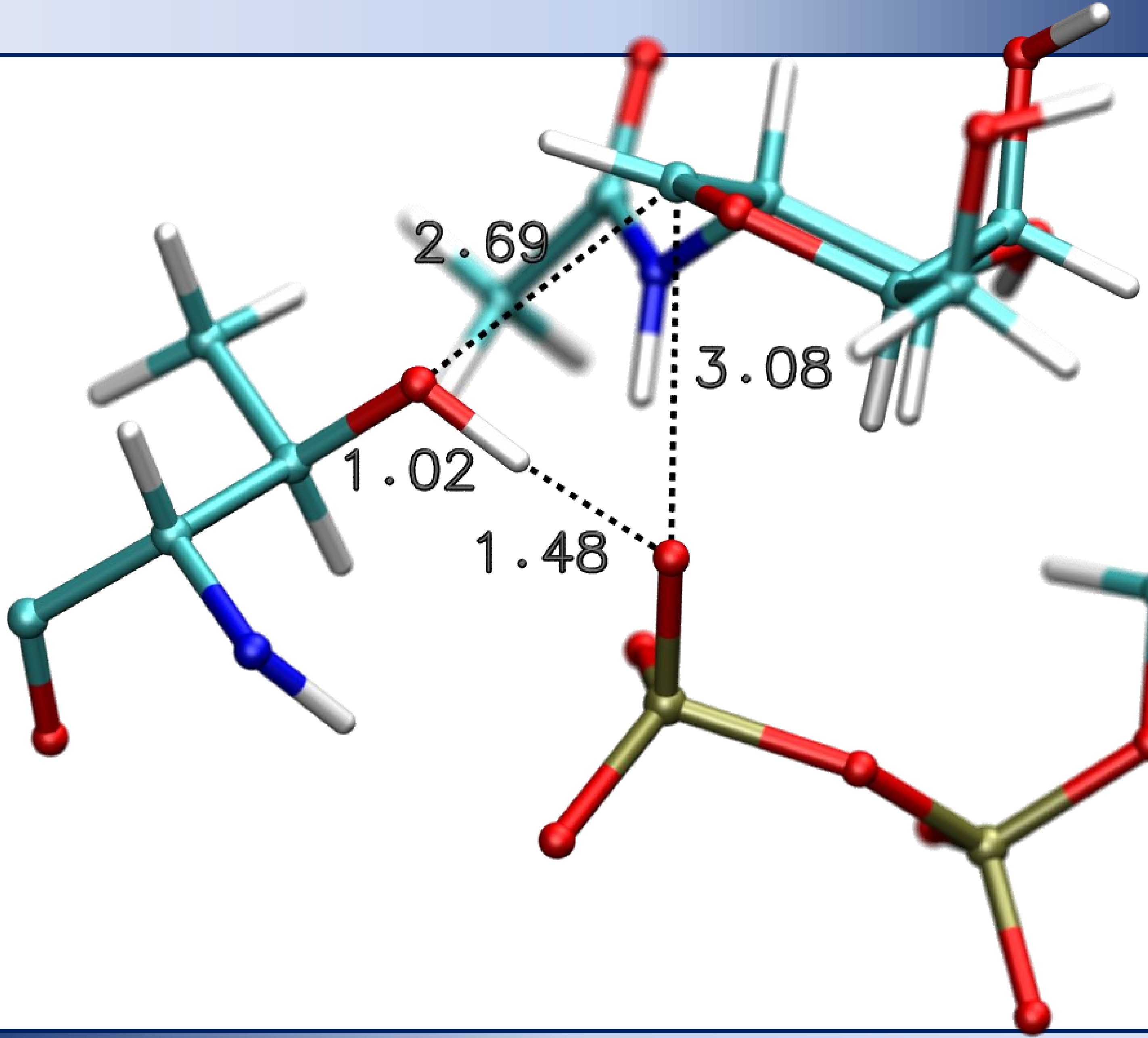
2.90
2.66
1.01
1.53

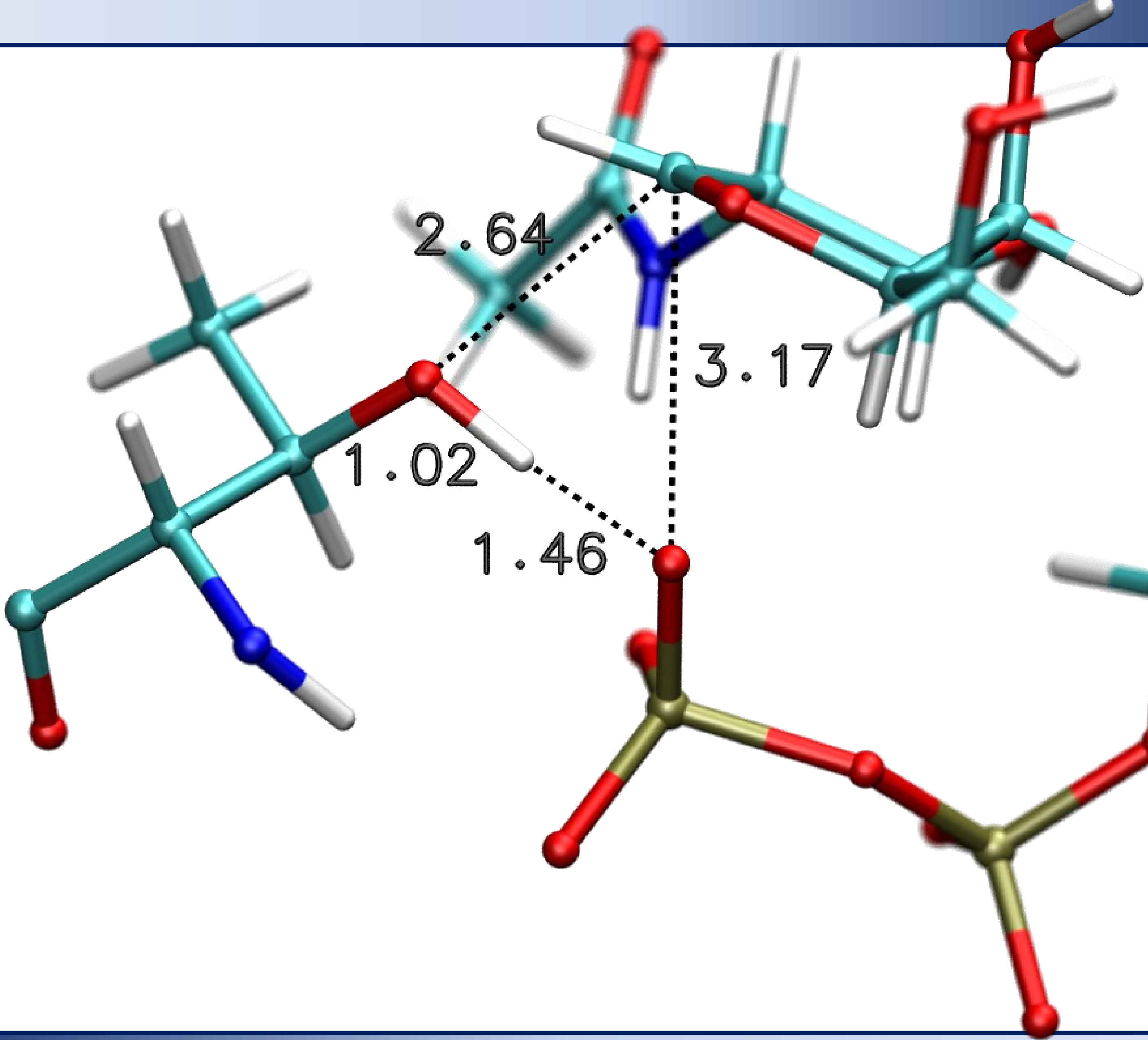

2.86
2.75
1.01
1.53

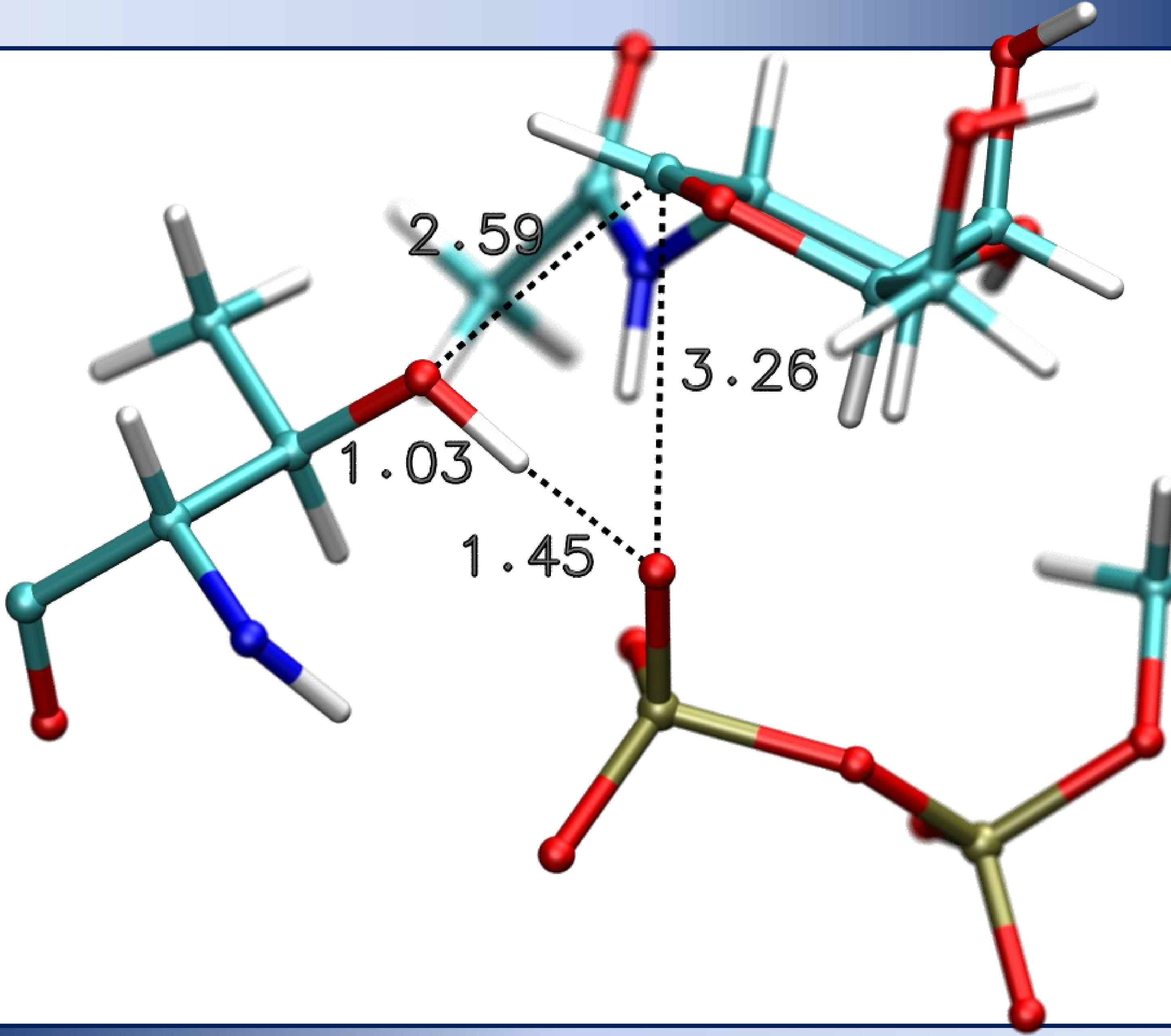

2.83
2.84
1.01
1.52

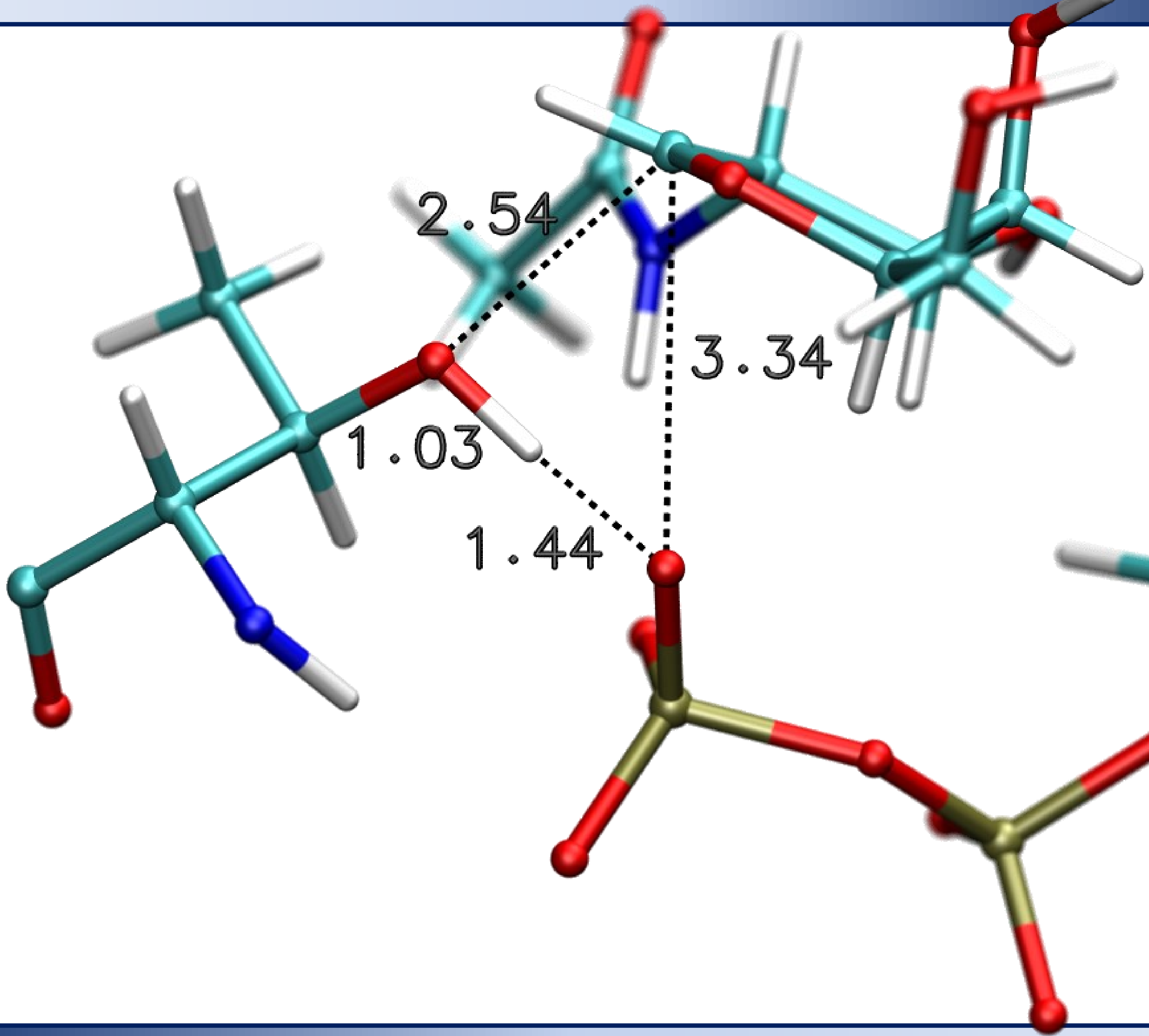

2.79
2.93
1.01
1.51

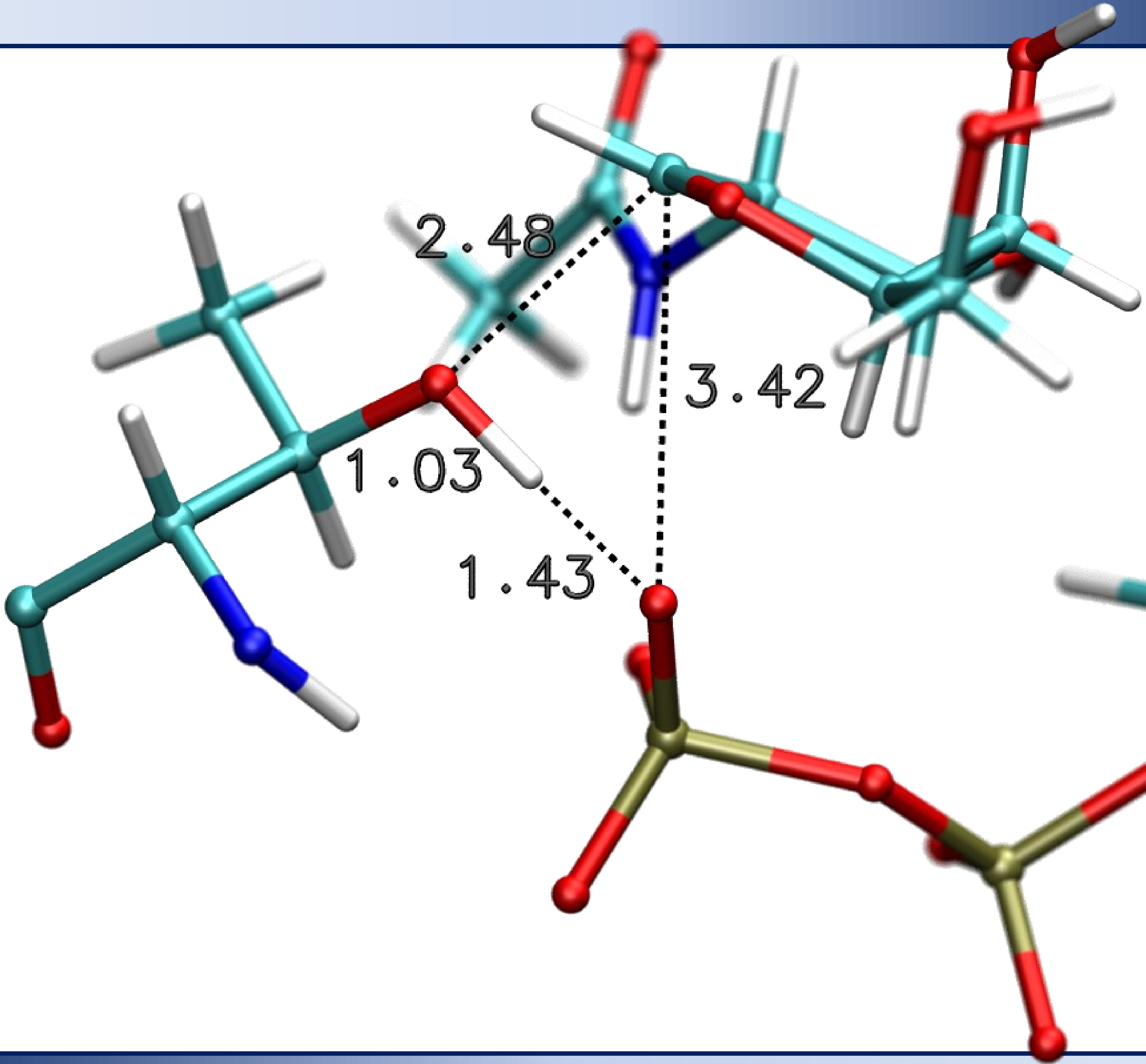

2.74
3.00
1.02
1.50

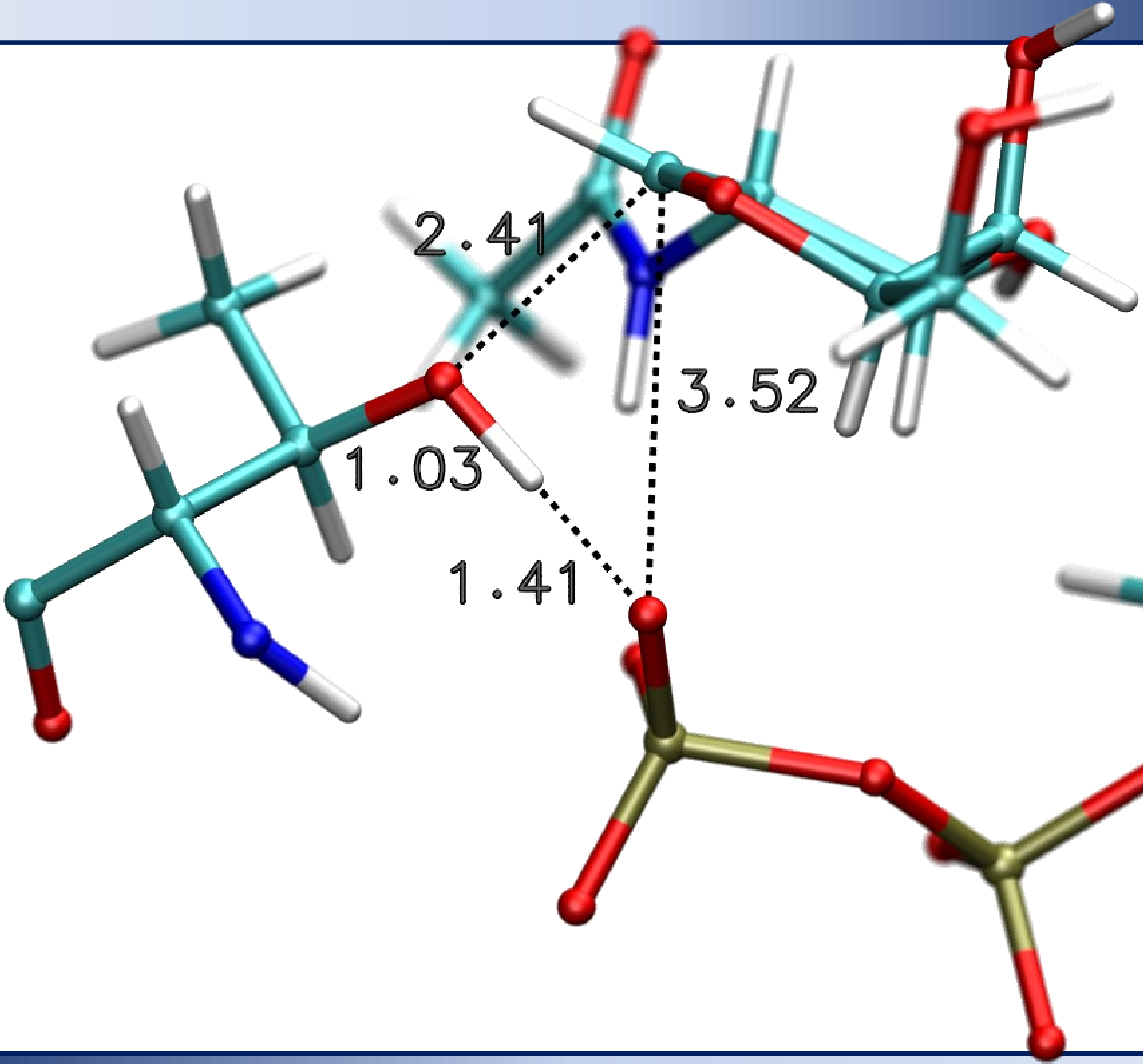

2.69
3.08
1.02
1.48

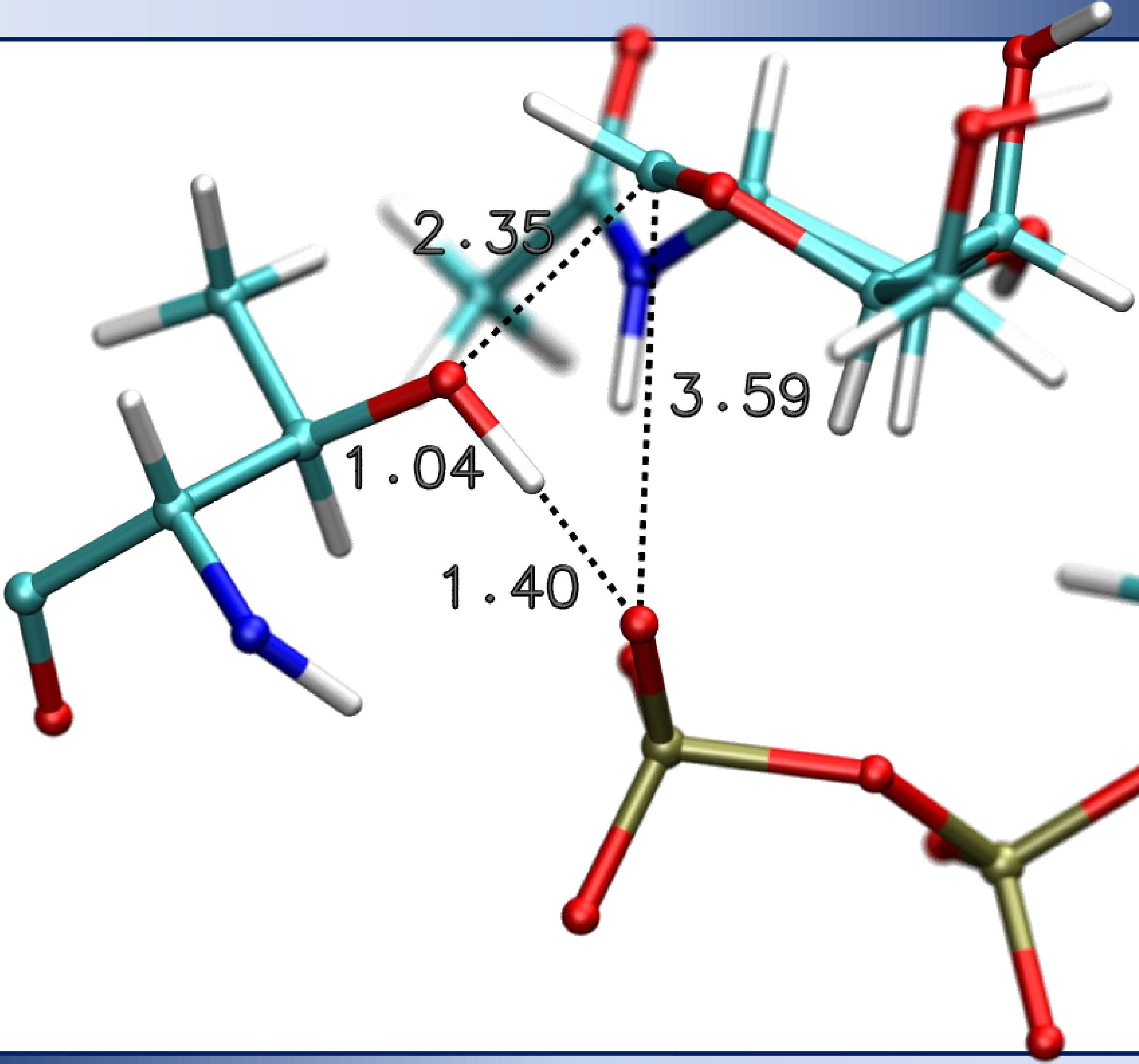

2.64
3.17
1.02
1.46

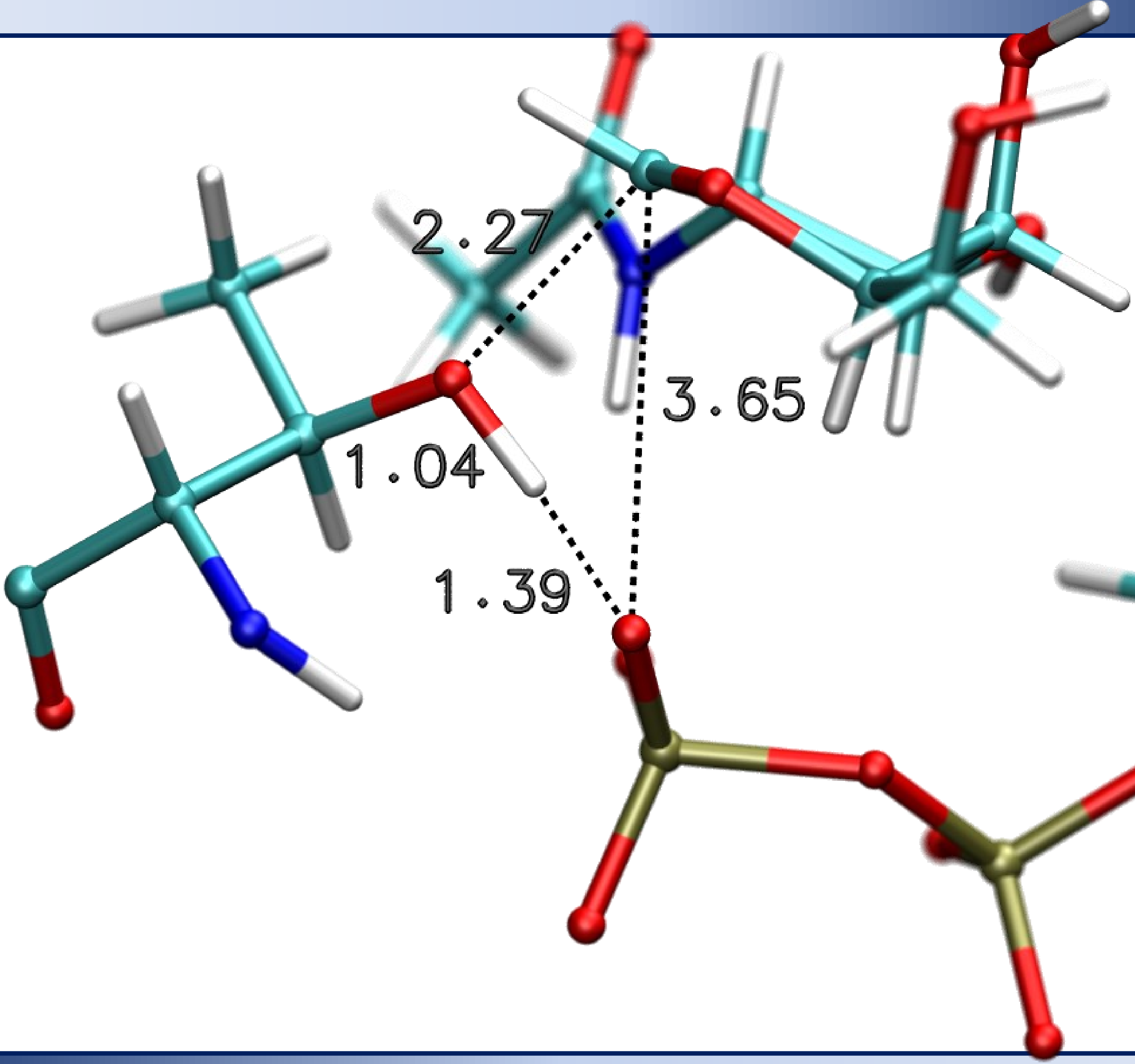

2.59
3.26
1.03
1.45

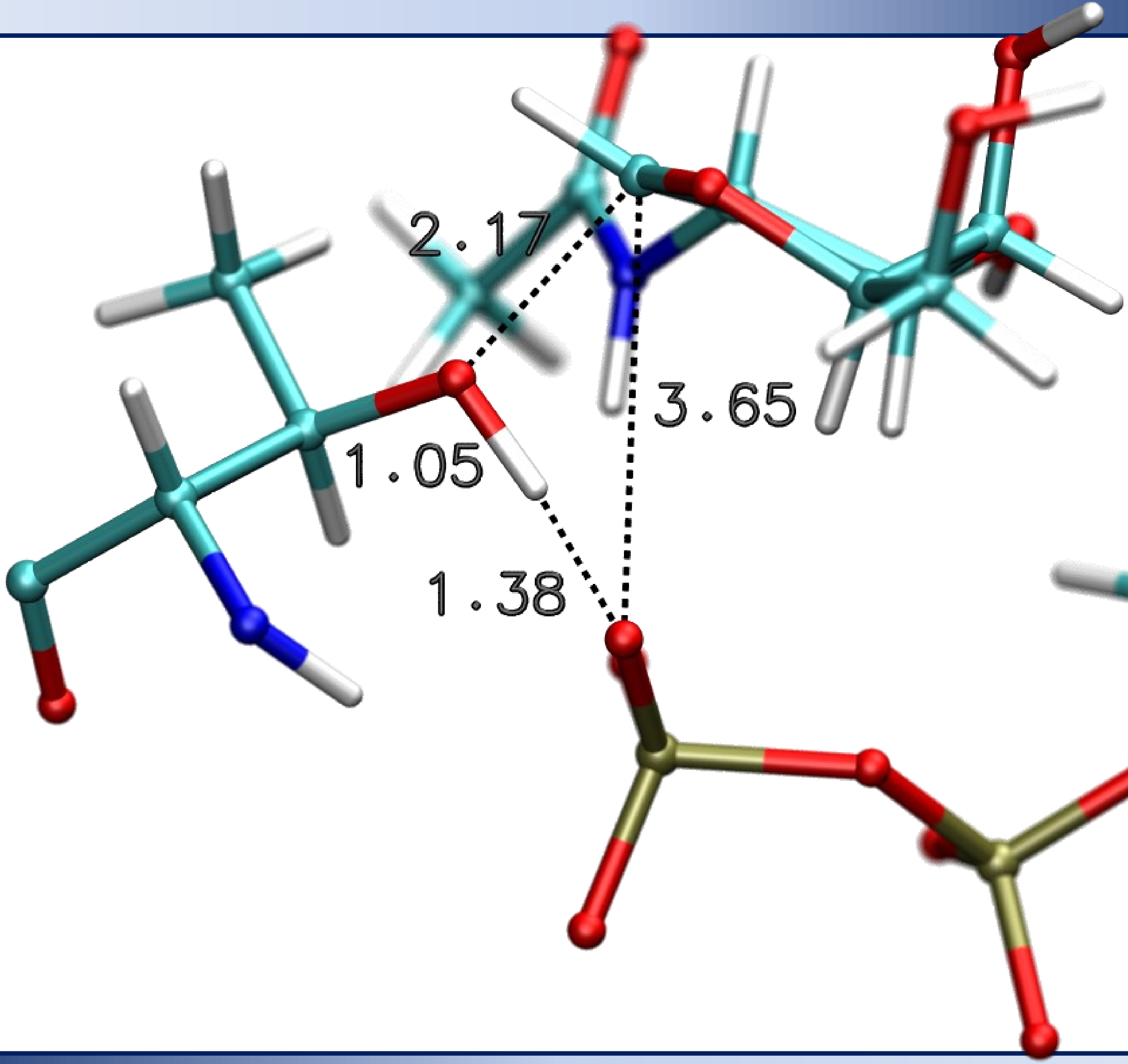

2.54
3.34
1.03
1.44

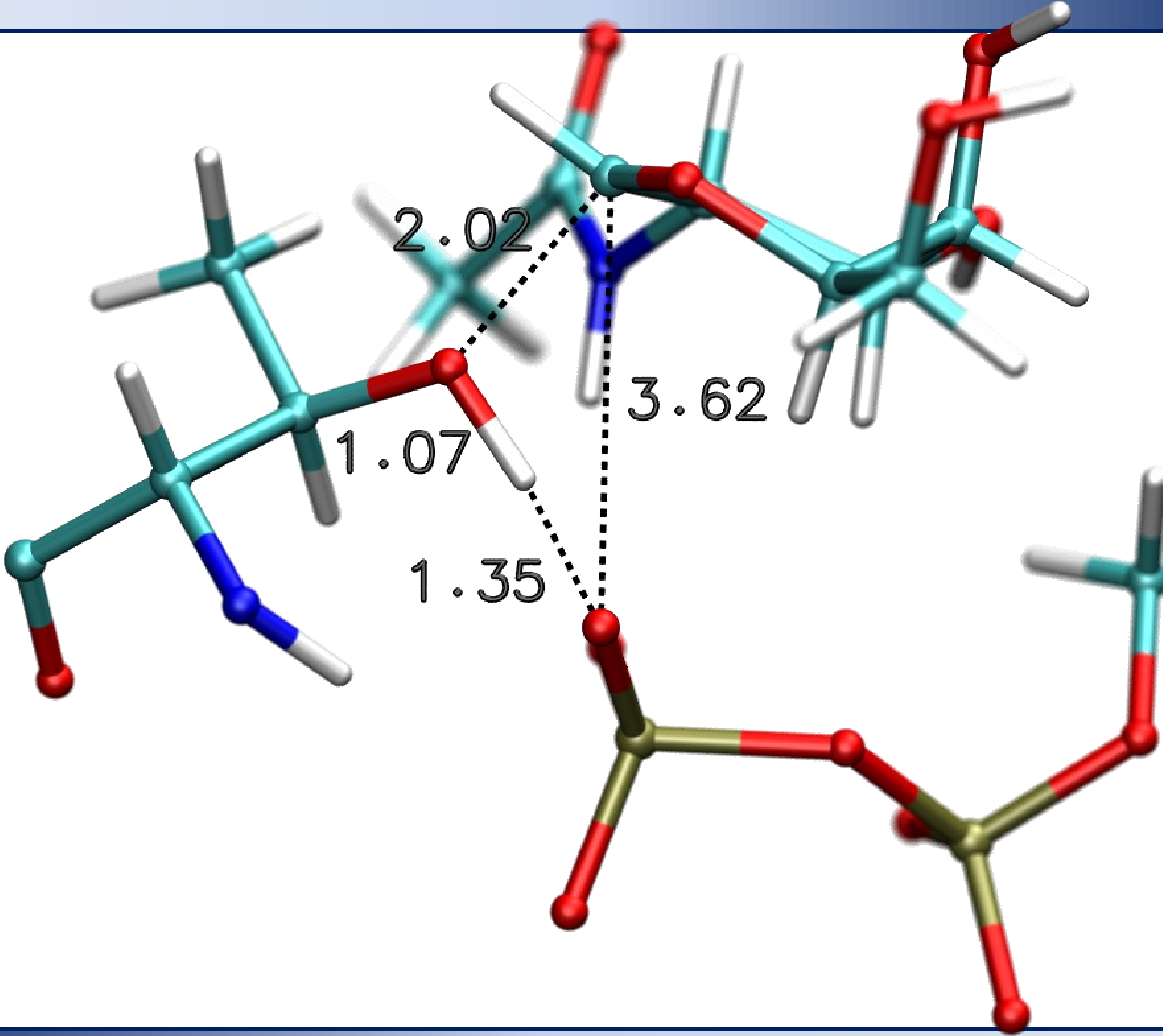

2.48
3.42
1.03
1.43

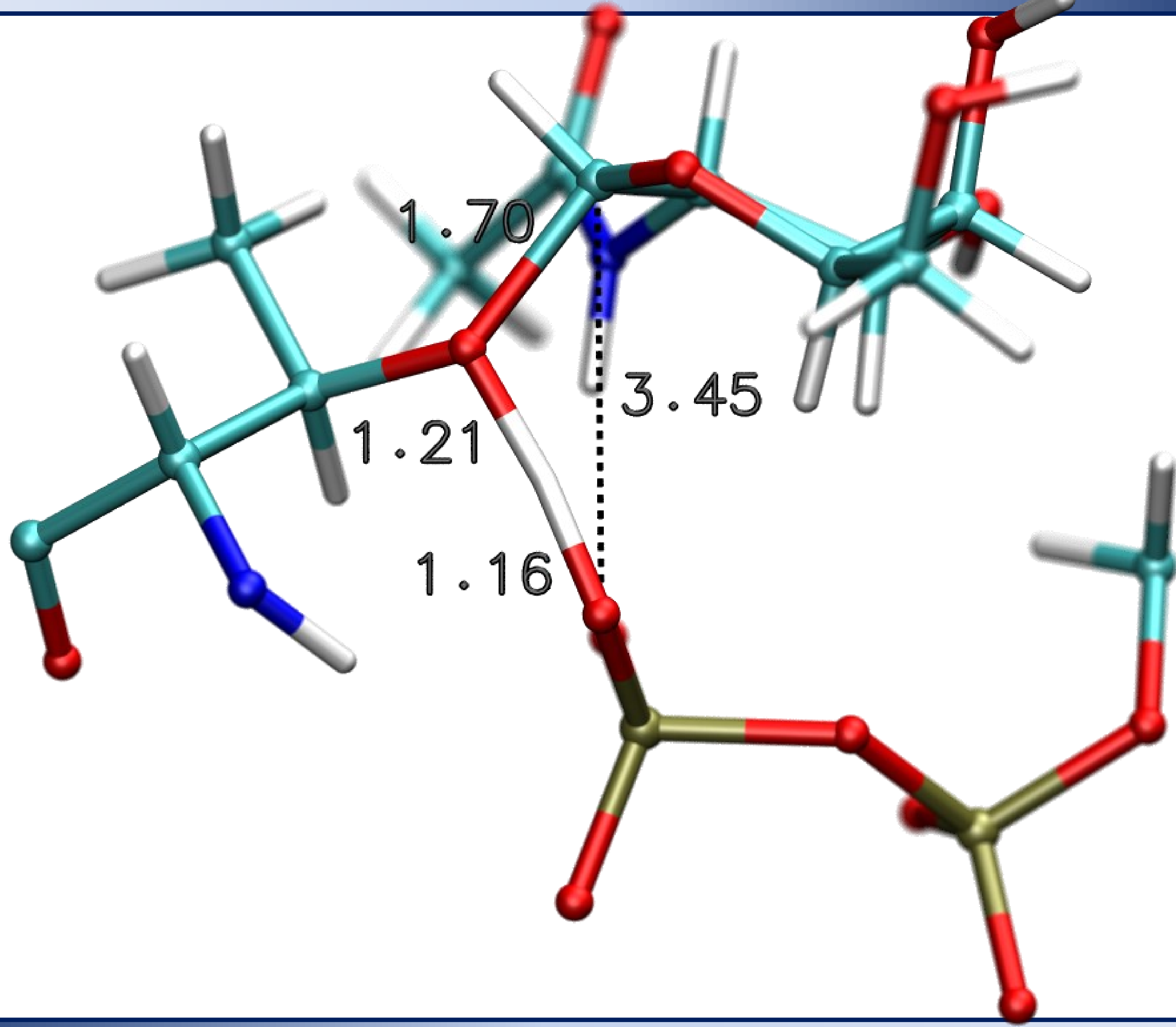

2.41
3.52
1.03
1.41

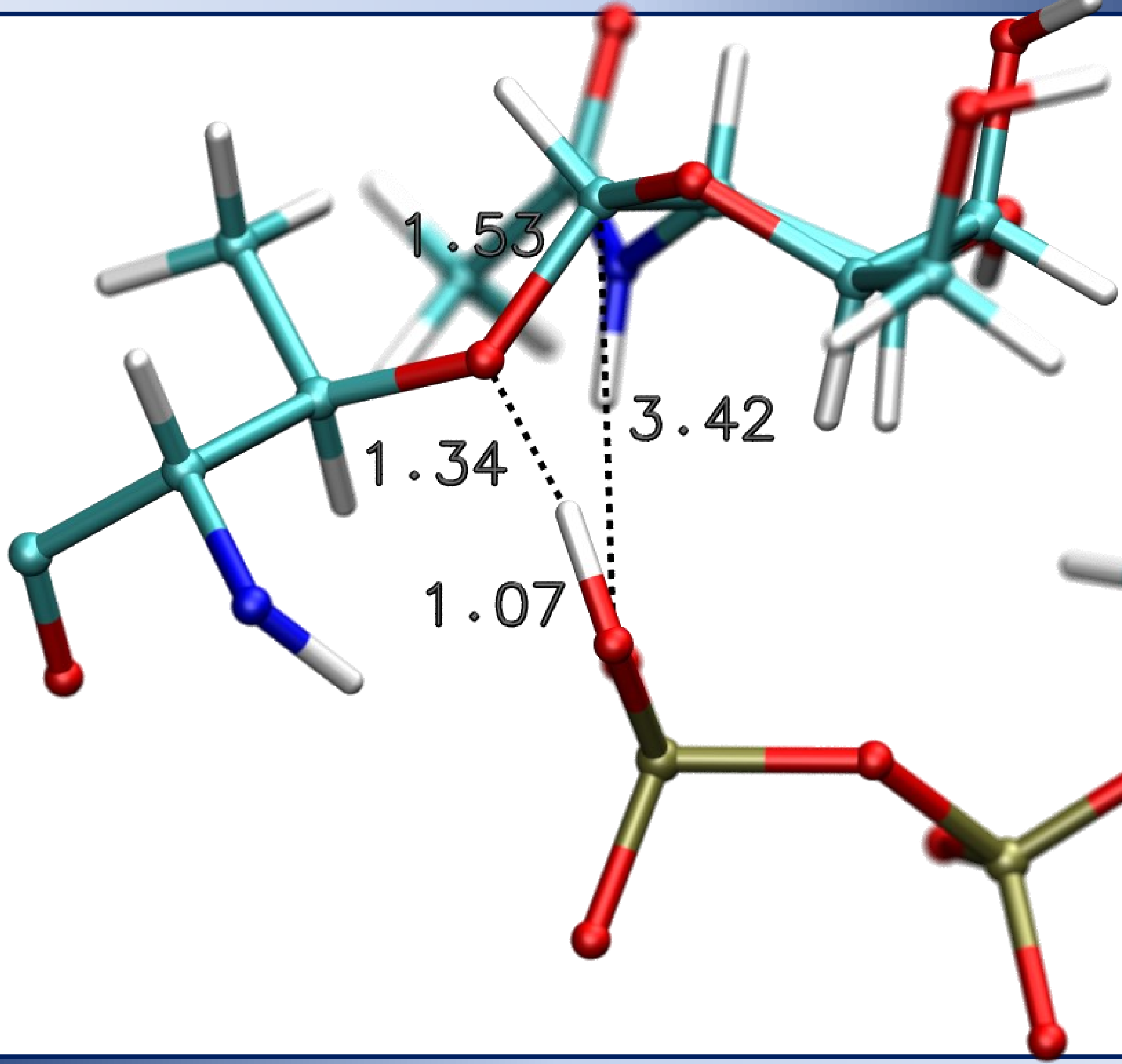

2.35
3.59
1.04
1.40

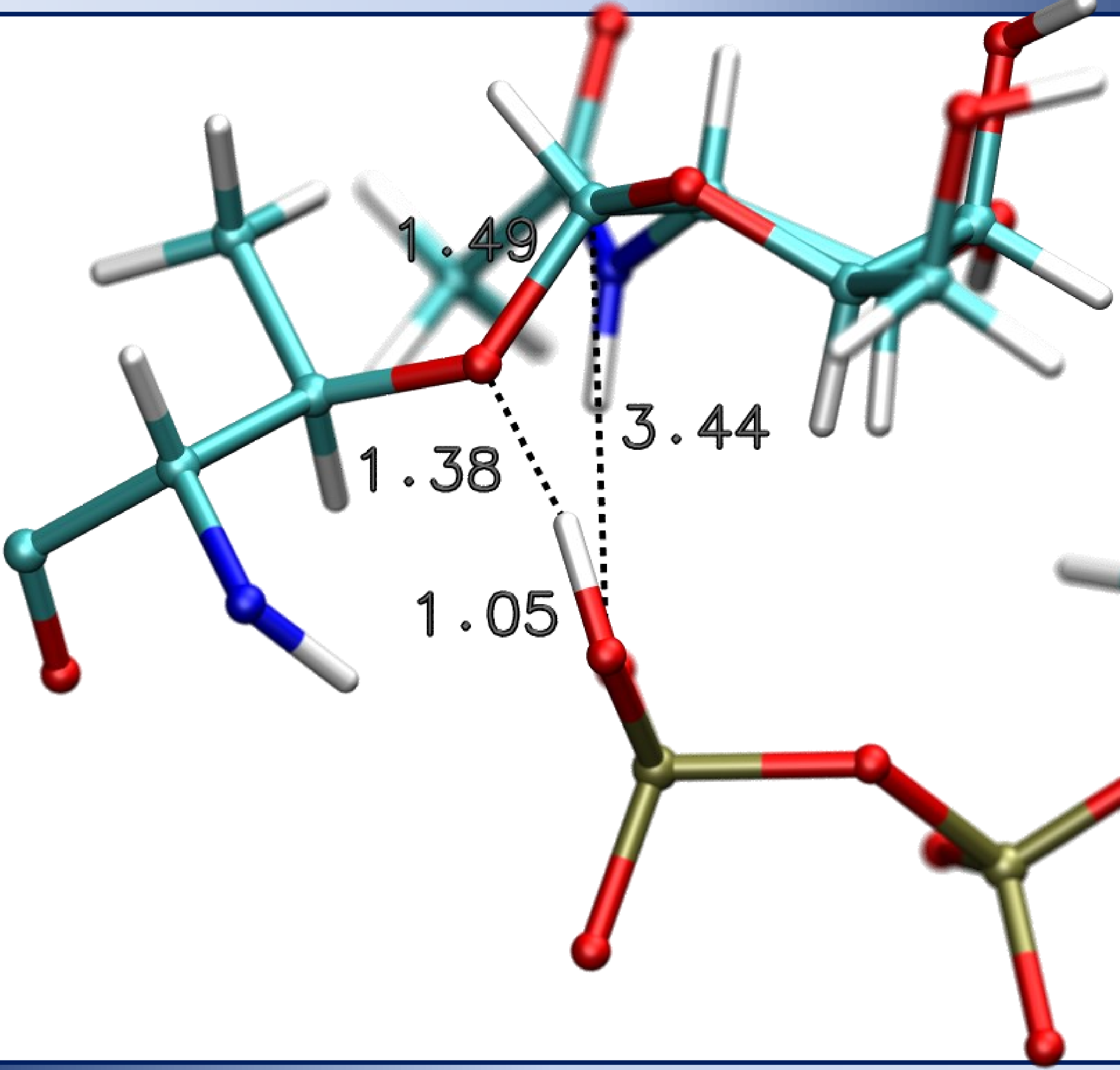

2.27
3.65
1.04
1.39


2.17
3.65
1.05
1.38

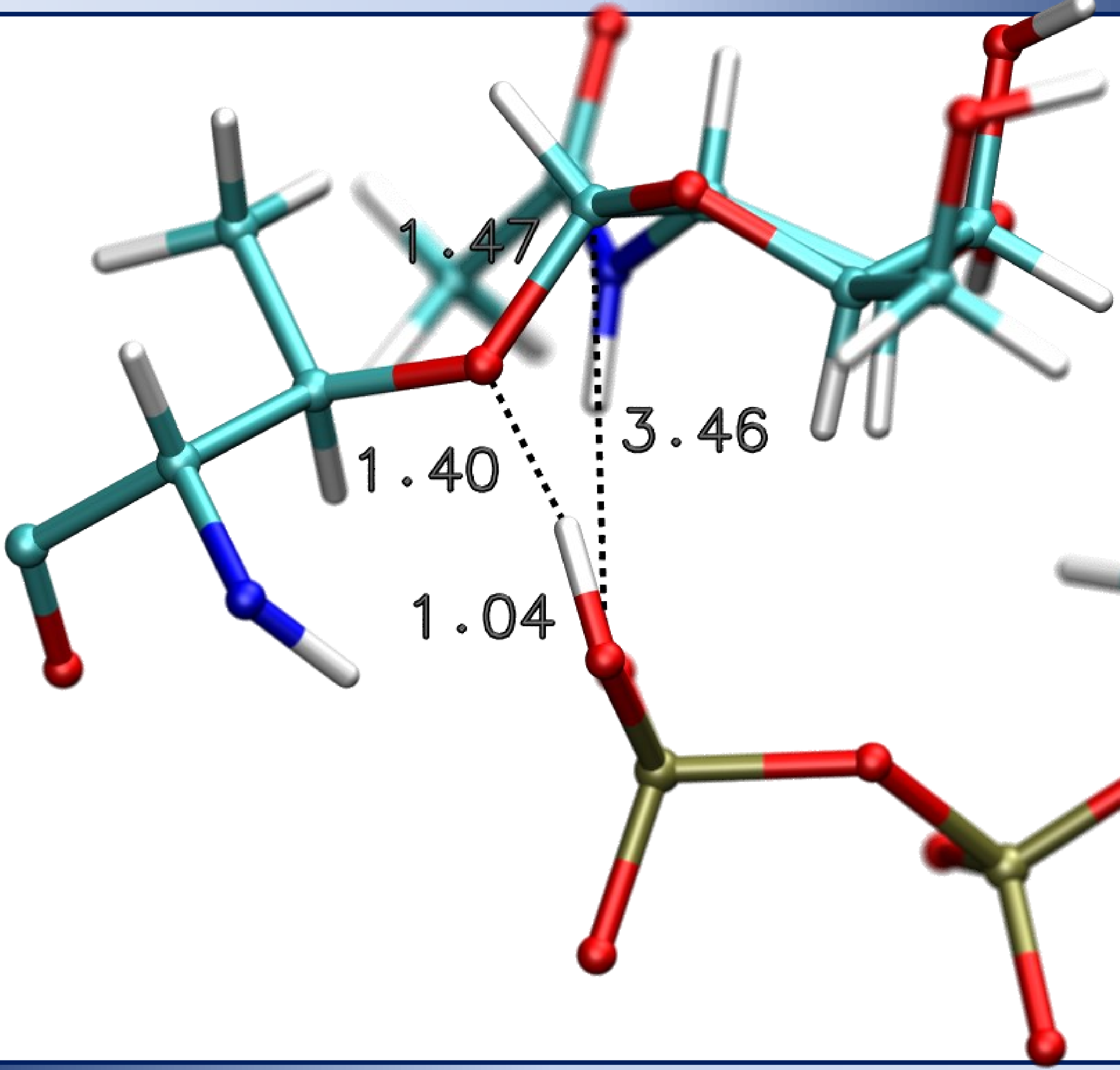

2.02
3.62
1.07
1.35

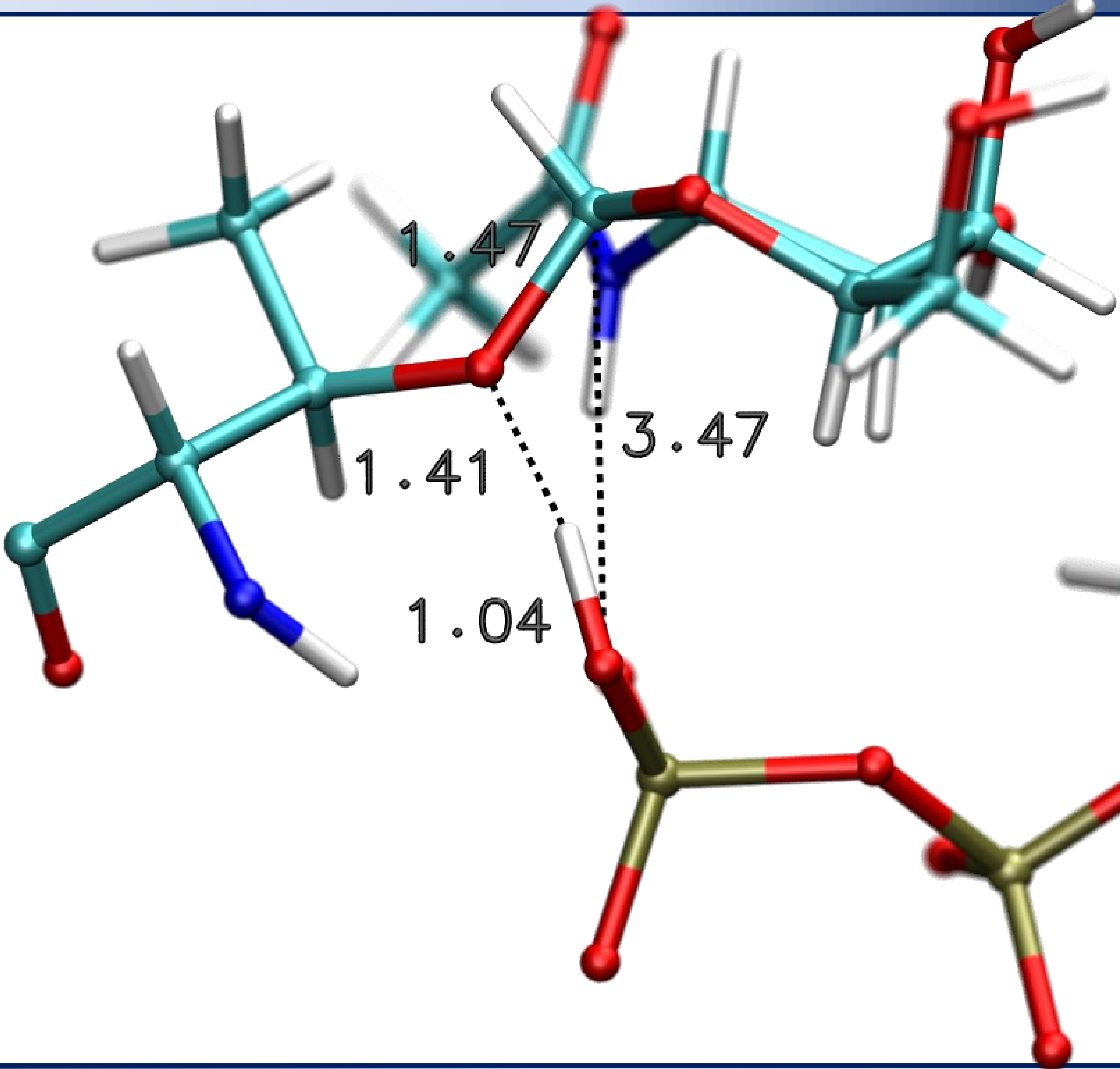

1.70
3.45
1.21
1.16

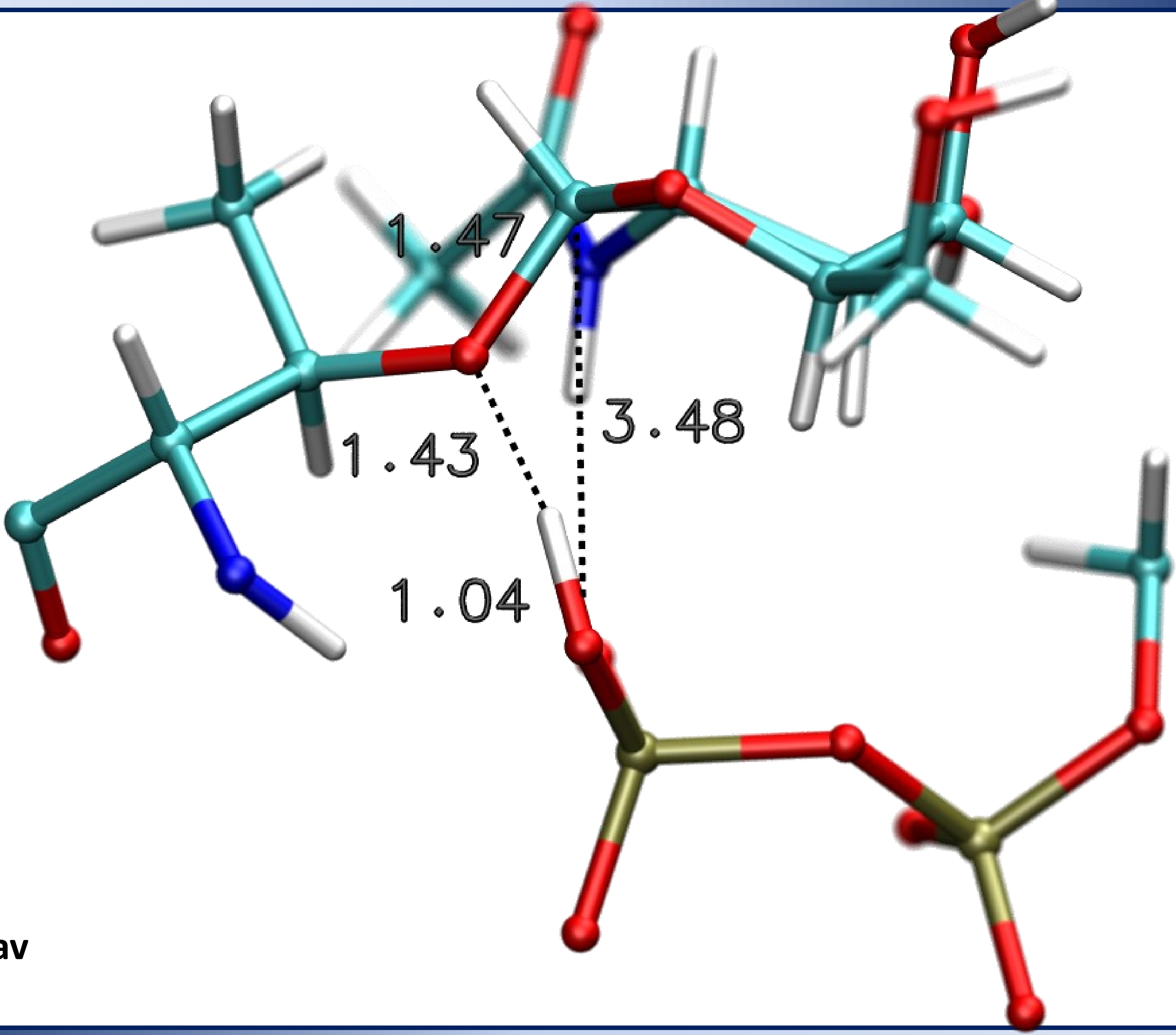

1.53
3.42
1.34
1.07



1.49
3.44
1.38
1.05

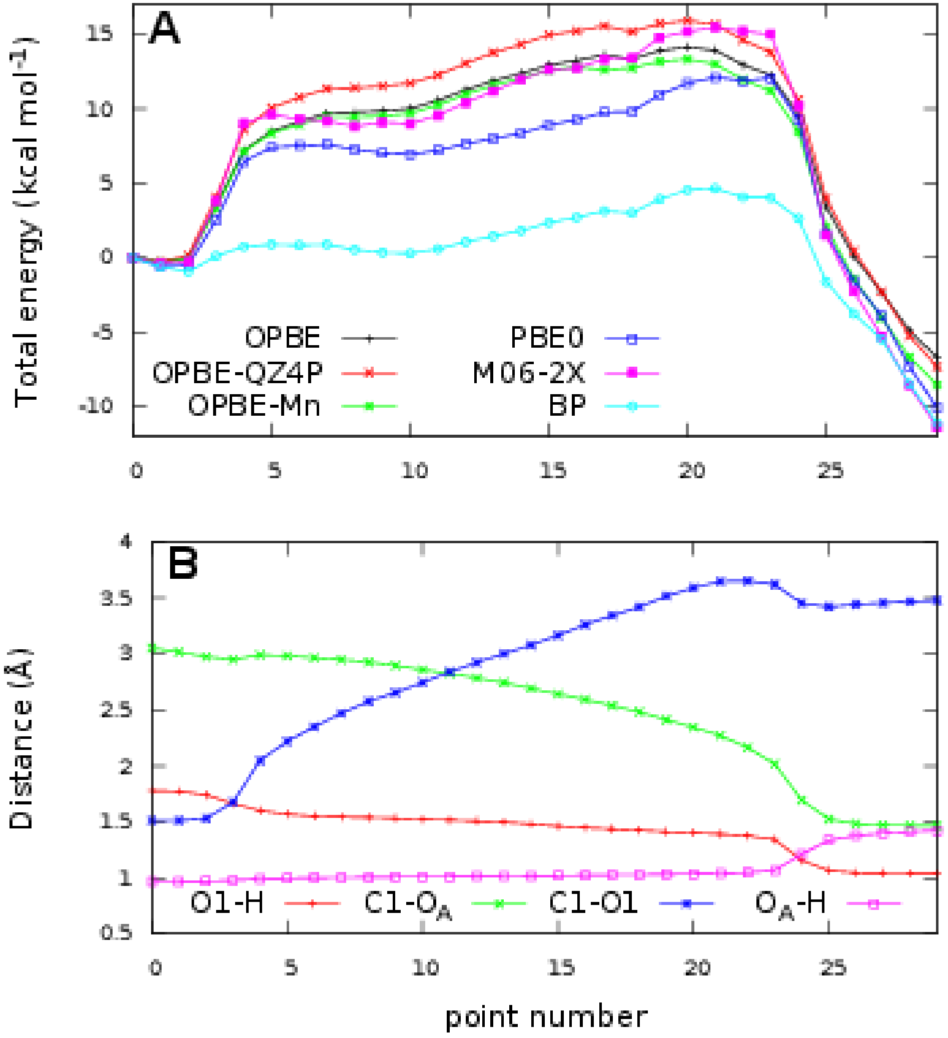

1.47
3.46
1.40
1.04


1.47
1.41
3.47
1.04



koncový stav

Result


A
Total energy (kcal mol^{-1})
15
10
5
0
-5
-10
0
5
10
15
20
25
OPBE
OPBE-QZ4P
OPBE-Mn
PBE0
M06-2X
BP
B
Distance (Å)
4
3.5
3
2.5
2
1.5
1
0.5
0
5
10
15
20
25
O1-H
C1-O$_A$
C1-O1
O$_A$-H
point number

Specifics of methods

Quantum mechanical methods:

- computational complexity increases with the required accuracy of the calculation and the size of the studied model
- these are computationally (CPU) as well as data (RAM) demanding calculations
- acceleration using parallelization is possible, but usually does not scale well (scaling is not linear for very precise methods)
- parallel run is more suitable on SMP nodes, it requires fast data connection of computing nodes, when run on clusters

Molecular dynamic simulations (using molecular mechanics):

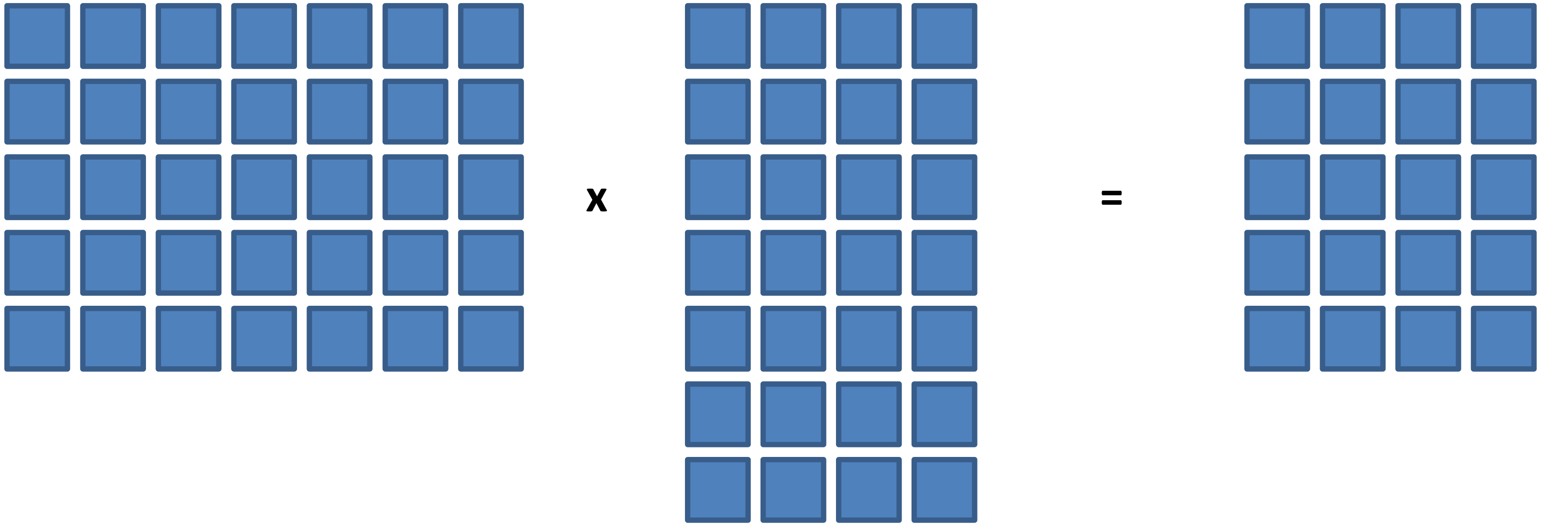
- computational complexity increases with the size of the model and length of required sampling
- Due to the low algorithmic complexity, calculations can be performed using GPGPU
- creates a large amount of data (trajectories)
- speeding up the calculation using parallel execution is easy
- parallelization can be performed on several levels (calculation of forces, more walkers or replicas), for the last two cases it is possible to achieve linear scaling

Exercise 1

1. What does the time complexity O(N) determine?
2. How many times is calculation of potential energy of benzene molecule by quantum chemical method CCSD(T), if we change used base from aug-cc-pVDZ to aug-cc-pVTZ? The number of base functions is 192 for aug-cc-pVDZ and 414 for aug-cc-pVTZ.
3. If the potential energy calculation time using the CCSD(T)/aug-cc-pVDZ takes 5 hours, how long will be calculation using the CCSD(T)/aug-cc-pVTZ?
4. The enzyme-catalyzed first order reaction has a single rate determining step with activation Gibbs energy of 18 kcal/mol. What is the reaction half-life at 300 K?
5. How long would a molecular dynamic simulation of one enzyme-substrate complex from the previous task have to take to observe substrate transformation with 50% probability?
6. Determine the number of integration steps that need to be performed in simulation from task 5, assuming that integration step is 0.125 fs (QM / MM dynamics in CPMD).
7. Determine the machine time that would be required to perform simulation, assuming one integration step takes 5 seconds. Discuss the value.
8. Determine the machine time required to perform 1 μs long molecular dynamic simulation of a cellulose fragment within a water box with a total number of 408609 atoms on one GTX 1080 graphics card under NPT conditions? Use the data provided here for a solution: https://ambermd.org/GPUPerformance.php

Model problems and systems

Matrix multiplication



A(n, m) **B**(m, k) **C**(n, k)

Use:

- finding eigenvalues and vectors of square matrices (quantum chemistry)
- solution of a system of linear equations (QSAR, QSPR)
- transformation (displacement, rotation, scaling - display and graphics)

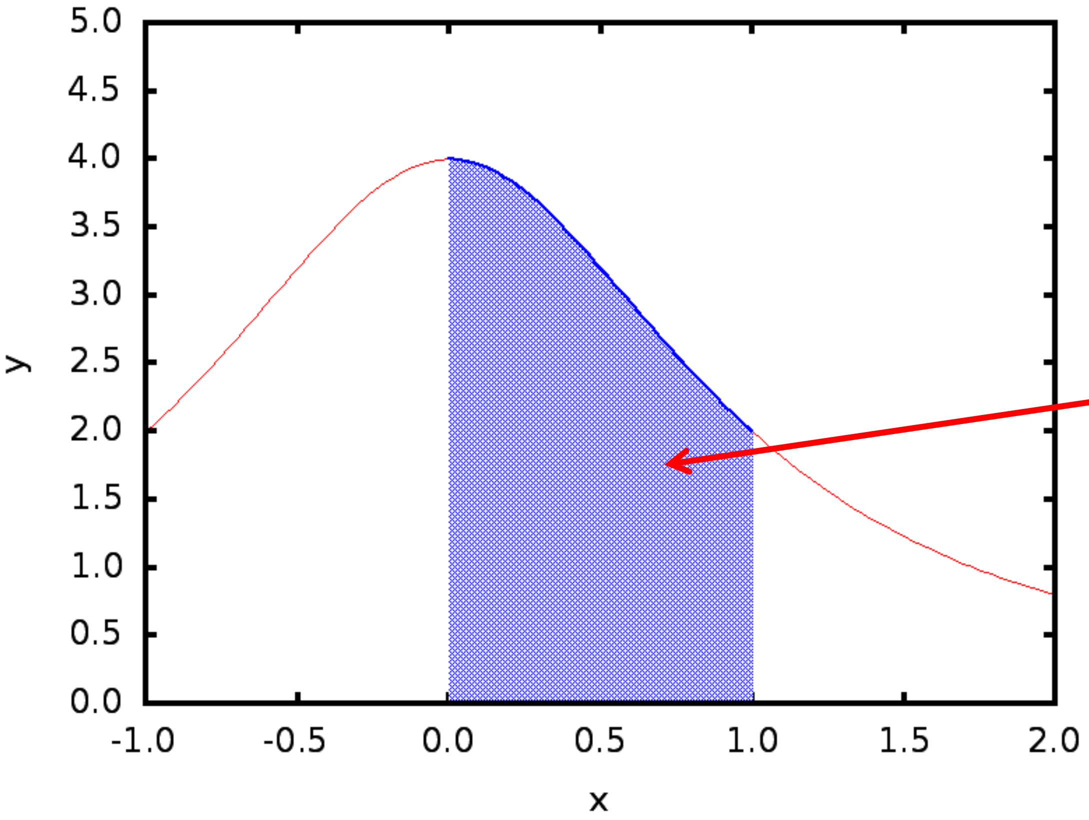
Revision/self-study:

- How is matrix multiplication done?
- How many operations need to be performed?

Numerical integration

The calculation of certain integrals can be performed by numerical methods, which are used if:

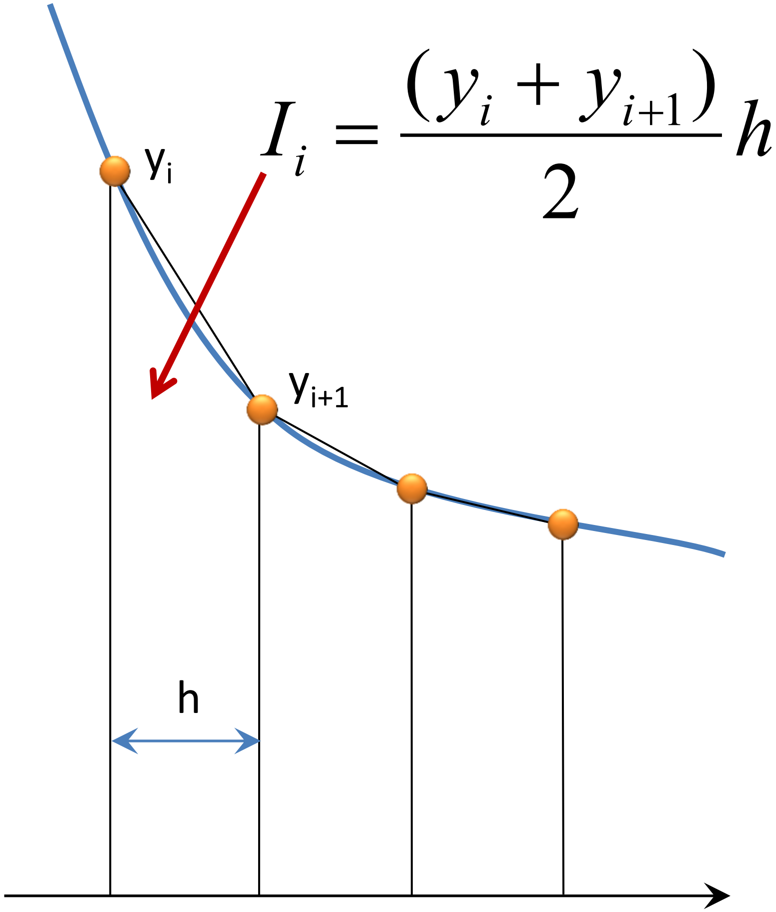
- the function cannot be integrated analytically
- analytical integration is practically impossible (accuracy *vs* computational complexity)



$$I = \int_0^1 \frac{4}{1+x^2} dx$$

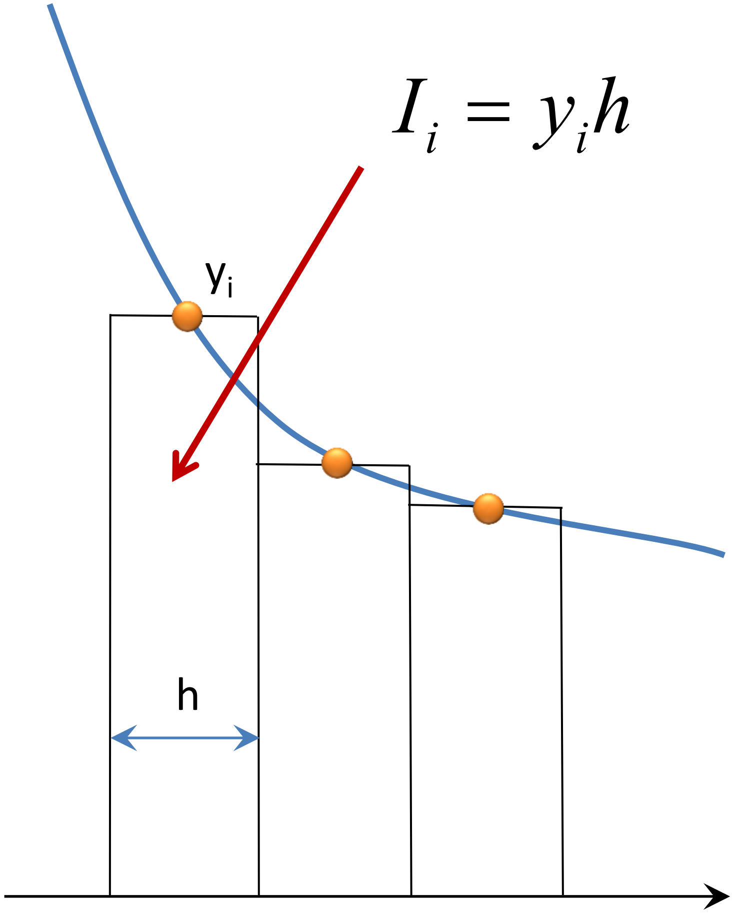
a certain integral is the area under the curve in the range of integration limits

Numerical integration methods



$$I_i = \frac{(y_i + y_{i+1})}{2} h$$

trapezoidal method

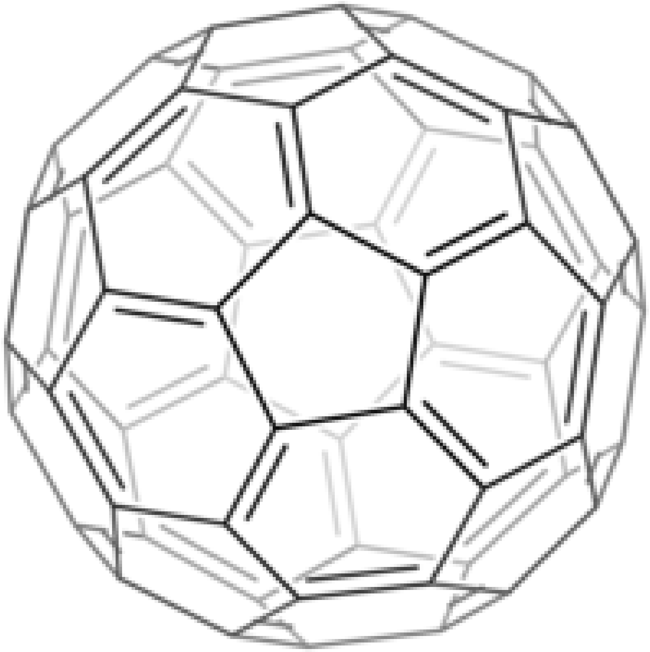


$$I_i = y_i h$$

rectangular method

Fulleren C_{60}

https://en.wikipedia.org/wiki/Buckminsterfullerene



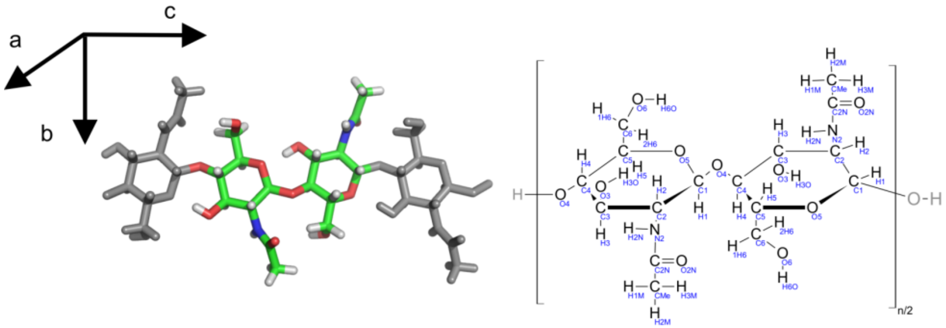
Tasks:

- creating a model of C_{60} molecule
- geometry optimization
- calculation of molecular vibrations

Methods:

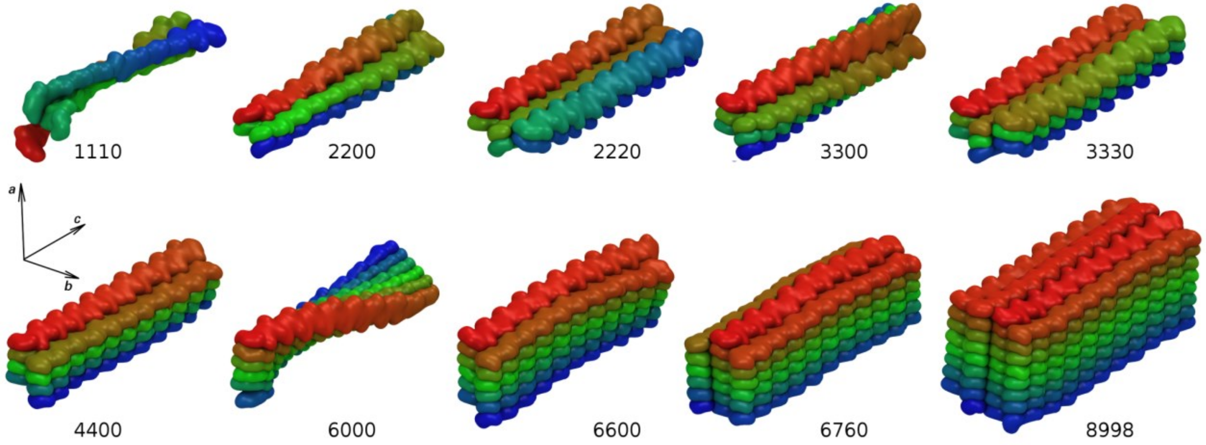
- semiempirical quantum-chemical method PM6

Chitin fibers



building unit

mechanical properties of chitin nanofibers



Tasks:

- MD simulation of 6000 fiber

Strelcova, Z.; **Kulhanek, P.; Friak, M.;** Fabritius, H.-O.; Petrov, M.; Neugebauer, J.; Koca, J. The structure and dynamics of chitin nanofibrils in an aqueous environment revealed by molecular dynamics simulations. *RSC Adv.* **2016**, *6 (36)*, 30710–30721
DOI: 10.1039/c6ra00107f

Relationship with course C2115

Matrix multiplication:

- limiting factors related to computer architecture (memory throughput)
- optimized libraries for numerical calculations (BLAS, LAPACK, Intel MKL, AMD MCL)

Numerical integration:

- limiting factors related to computer architecture (rounding errors and their impact on the integration result)
- parallelization of the calculation (OpenMP versus MPI)

Fulleren C_{60}:

- running calculations in the program Gaussian
  - in MetaCentrum (PBSPro)
  - in the WOLF cluster (PBSPro and Infinity)

Chitin fiber:

- molecular dynamics simulations in pmemd
  - scaling CPU parallel implementation
  - CPU and GPU runtime comparison

Exercise 2

Fulleren C_{60}:

1. Build a 3D model of a fullerene C_{60} molecule and optimize it using the force field MMFF94. To build a 3D model, use a structure in SMILES format (wikipedia for C_{60}). Save the resulting model in the format xyz. Use either avogadro or Nemesis program to build the model.

Chitin fiber:

Equilibrated the chitin fiber model can be found in the directory:
/home/kulhanek/Documents/C2115/Lesson02/chitin

system topology is 6000.parm7

coordinates, velocities and size of the box is in 6000.rst7

2. Display model in VMD.
3. How many atoms does the model contain?
4. How many fibers of chitin does the model contain?
5. What is the shape of the simulation box?

Self-study

1. How is matrix multiplication done?
2. How many operations need to be performed when multiplying matrices?
3. What is the computational complexity of matrix multiplication?
4. Which numerical method integration is more accurate, rectangular or trapezoidal?
5. Find other methods of numerical integration.
6. Is it possible to calculate the indefinite integral by numerical integration?