



JOINT INSTITUTE FOR NUCLEAR RESEARCH

Frank Laboratory of Neutron Physics

## **FINAL REPORT ON THE INTEREST PROGRAMME**

INTRODUCTORY COURSE: MD-SIMULATION RESEARCH (FROM  
ATOMIC FRAGMENTS TO MOLECULAR COMPOUND)

### **Supervisor**

Prof. Dr. Kholmirzo Kholmurodov

### **Student**

Ariel Alejandro Castro Sifonte

Institute of Technology and Applied Sciences

at University of Havana

Cuba

### **Participation period**

26 February - 14 April 2024, Wave 10

Dubna, 2024

## **Abstract**

Molecular dynamics (MD) simulates the collective behavior of atoms or molecules by following their classical equations of motion. It allows the study of the statistical and thermodynamic properties of physical, chemical or biological systems.

Despite its apparent simplicity, MD has proven itself to have great accuracy as well as cost-effectiveness, and is nowadays a key tool of the trade in multi-billion dollar industries, including, but not limited to, materials design and pharmaceuticals.

This report summarizes and iterates on what we learned in the course “MD-Simulation Research (From atomic fragments to molecular compounds)” at INTEREST program wave 10, as well as two other courses on nanoscale modelling and the physical foundations of MD I took at the University of Havana, imparted by Dr. Luis Alberto Montero and Dr. Roberto Mulet respectively. It also includes future perspectives to apply the learned knowledge.

## Introduction

Computer simulations in the physical sciences have been cemented in the last decades, not just a bridge between theory and experimentation, but as a pillar of scientific enquiry in its own right. Simulations help tell experimenters where to look, saving valuable time and resources in the process, as well as being able to offer insights not immediately accessible via theoretical or experimental methods.

In view of the different temporal and spatial scales involved in a physical system, a number of different simulation methodologies have been developed to assess the relevant phenomena in a cost-effective way, as shown in [Figure 1]. They range from the very accurate but computationally demanding Post-HF quantum mechanical methods, to continuum physics, where chemical information is ignored.

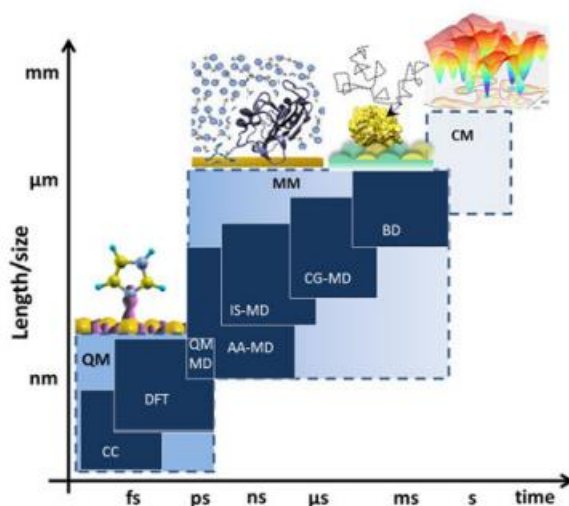


Figure 1. Typical time and length scales of different simulation techniques

Somewhat in between is Classical Molecular Dynamics (MD), which accounting for all-atom, coarse grained methods and Brownian dynamics, spans timescales from several picoseconds to just short of a second, and spatial scales from nanometers to a few micrometers. It is a simulation method consisting on the propagation of classical trajectories of particles (the definition of which is not always straightforward) on a semiempirical potential energy hypersurface (in configuration space) that allows to reproduce their interactions with sufficient accuracy.

## Basics of MD simulations

Most generally, (classical) MD is a computational technique for computing equilibrium and dynamical properties of a classical many-body system. The system moves in regular 3D-space along its physical trajectory as determined by Newton's equations of motion, which are integrated within MD and the simulated system develops over a period of time:

$$\mathbf{F}_i = m_i \mathbf{a}_i \Rightarrow \begin{cases} \dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \\ \dot{\mathbf{p}}_i = \mathbf{F}_i \end{cases}$$

where  $m_i$ ,  $\mathbf{p}_i$ ,  $\mathbf{a}_i$  are the mass, momentum and acceleration of the particle  $i$ . The forces acting on the particles explicitly relate to the inter-particle interactions modeled through the total potential energy  $V(\{\mathbf{r}_j(t)\})$  calculated from inter-particle potentials:

$$\mathbf{F}_i = -\nabla V(\{\mathbf{r}_j\})$$

the brackets  $\{\mathbf{r}_j\}$  denoting the set of all position vectors.

Conversely, in the equivalent Hamiltonian formalism we can define the (classical) Hamiltonian via:

$$H(\{r_i\}, \{p_i\}) = V(\{r_i(t)\}) + \sum_{i=1}^{3N} \frac{p_i^2(t)}{2m_i}$$

where  $i$  denotes both the particle in question and the generalized coordinate of choice.

The evolution of the system is given via the Hamilton's equations of motion:

$$\dot{r}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial r_i}$$

which is more well suited to evaluate thermodynamic properties of the system via ensemble theory, in which the microcanonical, canonical and isobaric/isothermal ensembles are the ones typically used.

In order to compare the computational and experimental results, one must invoke the ergodic hypothesis, which states that the ensemble average of an observable  $f$  is equal to its time average over infinite time:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_{t_0}^{t_0+t} f(\tau) d\tau = \int_{\Gamma} f(\xi) d\xi$$

where  $\Gamma$  is the set of all points in phase space. Thus, one has to ensure that the methods employed justify the assumption of ergodicity. The fact that the time averages are evaluated over finite time steps and at finite time hints to an optimum choice of computational speed and accuracy.

### *Periodic Boundary Conditions*

A first important issue is the fact that the system size in all simulations is smaller than in reality. This is not a big problem if the forces act over a length is much smaller than the system size. However, if the other case is true, we need to address, and the most common way is to apply periodic boundary conditions (PBC). In this way, the particles in the simulation box are virtually periodically repeated

in space to form an infinite system. In this way, the interactions between any two particles  $i, j$  is given by the equation:

$$\mathbf{F}_{PBC}(\mathbf{r}_i - \mathbf{r}_j) = \sum_{\mathbf{n}} \mathbf{F}(|\mathbf{r}_i - \mathbf{r}_j + \sum_{\mu=1}^3 L_{\mu} n_{\mu}|)$$

where,  $L_{\mu}$  are vectors along the edges of the volume of the rectangular system and the sum is taken over all vectors  $\mathbf{n}$  with integer coefficients  $n_{\mu}$ .

A sketch of the PBC scheme is shown in [Figure 2]. The application of boundary conditions such as PBC is quite time consuming, as it involves the calculation of terms of an infinite sum until convergence is reached. This though can be easily resolved through schemes in which not all the particle images are taken into account in the interactions, but a list of the nearest neighbors for each particle in the simulations is made.

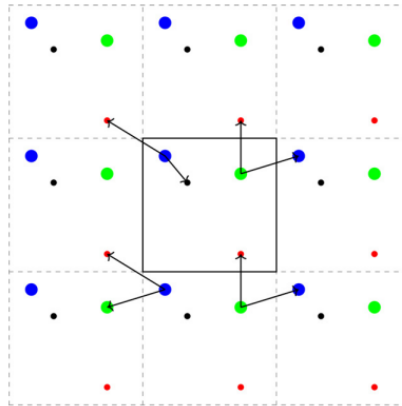


Figure 2. Sketch showing the distance vectors to the nearest neighbors of the particles in the simulation box at center, under periodic boundary conditions.

## Forces in MD

The key point in MD simulation is the computation of inter-particle forces, that is, the use of well adjusted, semiempirical classical potentials.

In simulations of physical systems, a key aspect is to model adequately the interactions between the particles of the system. Ongoing work is invested in developing classical inter-particle potentials. For this, three are the most important aspects:

- **Accuracy:** the potentials need to be accurate enough for the physical systems they have been developed for.
- **Transferability:** they need to be able to efficiently model additional systems.
- **Efficiency:** they should have simple mathematical forms in order to be easily implemented in a computational code and not increase the computational time beyond a reasonable threshold.

A potential should be able to model sufficiently the different types of bonding that can occur in a physical system. The particles of such a system can interact through bonds, ranging from strong

covalent bonds up to non-directed long-range interactions. Although methodologies have been devised, there is no hard rule for how to develop high quality classical potentials for the problem at hand.

### ***Simple pair potentials***

The simplest many-body systems to be captured by a simple pair potential are real, monoatomic gases. Simple model potentials can be used in order to model covalent and long-range Van der Waals interactions, such as the Morse potential:

$$V_{\text{Morse}}(r_{ij}) = D \left( e^{-2a(r_{ij}-r_0)} - 2e^{-a(r_{ij}-r_0)} \right)$$

and the Lennard-Jones potential, sketched in [Figure 3]:

$$V_{LJ}(r_{ij}) = 4\epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^6 \right]$$

In the last case, the  $r^{-6}$  term maps the dipole-dipole and London attractive forces at longer ranges, while the  $r^{-12}$  term traced back to the Pauli exclusion principle. Note that all the unknowns apart from particle-particle distances are parameters which need to be tuned for each particle type. Accordingly, these potentials have very limited transferability on their own.

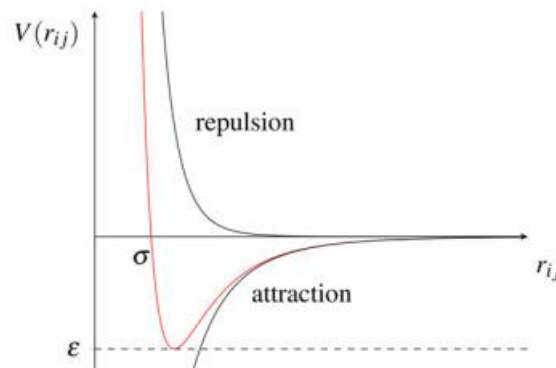


Figure 3. The Lennard-Jones potential, decomposed in its attractive and repulsive parts

### ***Force fields for biomolecules***

Biomolecular systems are typically very flexible and can access many degrees of freedom among bonding and non-bonding interaction. In a biomolecular force-field, covalent interactions between atoms (bond-stretching ( $V_{bond}$ ), bond-angle ( $V_{angl}$ ), bending, dihedral-angle torsion ( $V_{tors}$ ), and non-bonded interactions between atoms or units in different or the same molecule (separated by more than two or three covalent bonds) should be properly described. Typically, the most important non-bonded interactions are van der Waals ( $V_{vdW}$ ) and Coulomb ( $V_{Coul}$ ). The respective total potential energy would be expressed as:

$$V_{\text{Tot}} = V_{\text{bonded}} + V_{\text{non-bonded}}$$



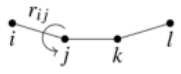

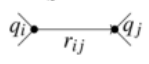
where:

$$V_{\text{bonded}} = V_{\text{bond}} + V_{\text{ang}} + V_{\text{tors}}$$

and:

$$V_{nonbonded} = V_{vdW} + V_{Coul}$$

Below we see common expressions for the different interactions:

$E_{bond} = \sum_{bonds} S_{ij}(r_{ij} - r_{ij}^0)$ 	<p>distance <math>r_{ij}</math> ; covalent bond with equilibrium-bond stretching term <math>r_{ij}^0</math></p>
$E_{ang} = \sum_{i,j,k} K_{ijk}(\cos \theta_{ijk} - \cos \theta_{ijk}^0)$ 	<p>covalent bond-angle ; equilibrium angle <math>\theta_{ijk}^0</math> bond bending term</p>
$E_{tor} = \sum_{i,j,k,l} \sum_n \frac{V_{nijkl}}{2} [1 \pm \cos(n\tau_{ijkl})]$ 	<p>torsional angle <math>\tau_{ijkl}</math>; dihedral angles-torsion term</p>
$E_{vdW} = \sum_{i,j} \left( -\frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \right)$ 	<p>van der Waals interaction through a Lennard-Jones potential</p>
$E_{Cb} = \sum_{i,j} \left( \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \right)$ 	<p>Coulomb interactions between partial charges <math>q_i ; q_j</math> ; <math>\epsilon(r_{ij})</math>: distance-dependent dielectric constant</p>

These expressions include distances and angles and a number of parameters which need to be tuned for a more ad-hoc category that is particle types. FF parametrization is a field of research in itself, and typically the parameters are chosen based on experimental (x-rays, Raman, neutron diffraction) as well as ab-initio QM simulations. Regarding the applicability and efficiency of an FF, often it cannot be applied to a variety of molecules it was not developed for. In addition, the FF parameters are consistent with specific solute molecules and solvent and cannot be used universally.

### **Long range Coulomb interactions. Ewald summations**

In order to improve computational efficiency, one needs to dispense with terms that do not contribute significantly to the total potential energy of the system. This is typically done by establishing a cutoff range for the different interactions. This methodology fails for the long-range Coulomb interactions, which decay with order  $r^{-2}$ , and thus the effective cutoff range is useless since it engulfs the entire system many times over, as well as being computationally inefficient to implement directly. Thus this case calls for a new approach, which is typically done via Ewald summations.

The method replaces the summation of interaction energies in real space with an equivalent summation in Fourier space and the long-range interaction potentials are divided into two parts: a short-range contribution calculated in real space, and a long-range contribution which does not have a singularity and is calculated using a Fourier transform to wavenumber space:

$$\frac{1}{4\pi\epsilon_0} \sum_{i<j}^N \frac{q_i q_j}{r_{ij}} = \frac{1}{2V_0\epsilon_0} \sum_{k \neq 0}^{\infty} \frac{e^{-k^2 4\alpha^2}}{k^2} \left| \sum_j^N q_j e^{-ikr_j} \right|^2 + \frac{1}{4\pi\epsilon_0} \sum_{i<j}^{N^*} \frac{q_i q_j}{r_{ij}} \operatorname{erfc}(\alpha r_{ij})$$

As a result, the Ewald summation method converges faster than a direct summation and is very accurate.

Considering that the calculation of long range Coulomb forces is very often the limiting factor for rapid force-field recalculation, potent hardware with architectures optimized to carry on this algorithm has been developed and currently continues to be developed [Figure 4], which is a testament to the importance of MD simulations in current applied research.



Figure 4. Special purpose MDGRAPE hardware for, among others, the correct estimation of Ewald sums

### ***On system modelling***

As previously stated, the particles in the model do not need to map one-to-one to the atoms present in the real system. Typical examples include coarse-graining or considering polymer chains as single particles. Other example, more close to all-atom molecular dynamics, is the modelling of aromatic rings for DL\_POLY, which include negatively charged particles at the center of the ring, on top and bottom, to account for the effect of delocalized p-orbital electrons.

### ***Examples of state-of-the-art force fields***

Widely employed, fine tuned force fields, some of them proprietary, include CHARMM, GRONINGEN MOLECULAR SIMULATIONS (GROMOS), AMBER and DL\_POLY. The last two we discussed in some detail during this interest program. In general, DL\_POLY is very efficient and suited to simple atomic, molecular and condensed matter systems, while AMBER is specifically very well suited to biochemical systems and thus not generalizable beyond this domain.

## **Environment conditions**

Given that most chemical processes of interest occur at constant temperature and pressure, it's important to ensure these conditions are met when performing simulations if needed.

In classical MD simulations, setting the environmental conditions can be explicitly done using thermostats (for applying a finite temperature) or barostats (for applying a pressure). If none of these is used, the MD simulations sample the microcanonical (NVE) ensemble, as this is the ensemble arising from the Newtonian dynamics which conserve the total energy and volume of the system

### ***Thermostats***

In principle, a thermostat is based on the modification of the Newtonian dynamics in order to generate a statistical ensemble at a constant temperature and modulate the temperature of a system. It makes



use of the fact that the temperature of the system is essentially the average kinetic energy of its constituent particles:  $\langle K \rangle \propto T$ .

Note that the goal of a thermostat is not to keep the temperature constant (fix the kinetic energy), but to ensure that the average temperature is the desired one. In order to achieve that goal, some clever mechanisms have been devised. Some examples include imposing an appropriate friction coefficient in the particle dynamics via the Langevin equation (Langevin thermostat), and adding an artificial particle associated with an artificial mass (Nosé thermostat).

### **Barostats**

Conversely, the idea behind barostats is that the pressure in a classical N-body system is calculated via the "Clausius virial theorem":

$$P = \frac{2}{3V}(K - \Xi)$$

Where  $K$  is the kinetic energy, and  $\Xi$  is the inner virial for pairwise additive interactions, which is a tensor quantity dependent of each distance and force vectors. In that way, a correction of the pressure in simulations can be achieved via modifying this inner virial  $\Xi$ . To that end, several ways have also been devised, included but not limited to the Berendsen and Parinello-Rahman barostats.

## **Numerical integration**

Newton's equations of motion can be solved analytically only for  $N < 3$ , where  $N$  is the number of particles in the system. Otherwise, the equations can be solved only by numerical integration. Typically, the integration of the equations of motion is obtained by a Taylor expansion (the so-called Verlet type-algorithms).

One has to take present the following criteria before choosing an integration scheme:

- **Accuracy** denoting to which power of the time step the numerical trajectory will deviate from the exact trajectory after one integration step.
- **Energy conservation** along the exact trajectory.
- **Time-reversibility**, which is a result of energy conservation and assures that moving back in time would lead to the same trajectories as when moving forward in time.
- **Symplecticity**, which is related to the preservation of the phase-space in the simulations. It is physically impossible to remove or add regions in the phase-space without any external reason. Accordingly, the integrator should sample the phase-space, but keep it intact.

Most state-of-the-art integrators are based on the Verlet algorithm, since simpler ones like Euler and Runge-Kutta integrators famously fail energy conservation in systems as simple as the harmonic oscillator.

In it, the equations of motion are solved based on the current positions  $\mathbf{r}_i(t)$  and forces  $\mathbf{F}_i(t)$ , and previous positions  $\mathbf{r}_i(t - \Delta t)$ . A Taylor expansion is taken up to order-4 for positions and velocities. By subtracting and adding these expansions, the equations for the integration scheme are obtained:

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \frac{(\Delta t)^2}{2m_i}\mathbf{F}_i(t) + \mathcal{O}(\Delta t^4)$$

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^3)$$

The integration starts by specifying the  $\mathbf{r}_i(t)$  and  $\mathbf{v}_i(t)$  for the initial time.

The Verlet algorithm is fast and time reversible. It is symplectic as it conserves the volume of the phase space as imposed by the Hamiltonian dynamics. It shows a good short-term energy conservation and a small long-term energy drift, while the trajectories are very close to the constant energy hyper-surface in phase space. A strong disadvantage of the original Verlet algorithm is that velocities do not enter directly, which makes the simulations at constant temperature difficult. To that end, schemes like Velocity-Verlet and Beeman have been devised.

### ***Convergence***

MD simulations span a number of spatial and temporal time scales. Based on the continuous methodological development and the increasing computational power, these scales are constantly being extended. For example, in biomolecular simulations, typically involving the modeling of biomolecules in a solvent, in 2001 it was possible to simulate about 106 atoms for 1 ns. In 2010, a 1 ms simulation of a small protein in water was reported. Today, the simulation of biomolecular systems for 1 ms is possible. What hinders the use of very large systems and their simulations for very long times are mainly the computational resources, which are relevant to how fast the simulations can converge.

Every simulation of a physical system needs to be converged, either this means that its minimum energy is found or the forces acting on the system are below a tolerance. In view of the before mentioned time and size scales for MD simulations, the question arises whether these time scales are long enough to generate reliable trajectories, properties, etc.

Let  $Q$  be a property of a system, which needs to be determined,  $\tau_{equil}$  the equilibration time of the simulation,  $\tau_{relax}$  the relaxation time of property  $Q$ , and  $\tau_{sample}$  the sampling time. Reliable results are expected when:

$$\tau_{equil} > \tau_{relax} \gg \tau_{sample}$$

Accordingly, the choice of  $\tau_{relax}(Q)$  is of high importance when targeting the property  $Q$ . The time  $\tau_{relax}(Q)$  is generally better defined from the decay time of the autocorrelation function  $\langle Q(t')Q(t' + \Delta t) \rangle$ .

Now, when the simulation starts from a non-equilibrium initial state,  $\tau_{relax}(Q)$  is given from the average rate of relaxation of  $Q$  towards the equilibrium and is measured over many non-equilibrium trajectories. In cases for which MD simulations starting from a different initial condition do not converge to the same averages for  $Q$ , then  $\tau_{relax}(Q)$  is longer than the simulation time and the simulations should run over a longer time in order to obtain reliable results for the property  $Q$ .

## Applications and Future Goals

MD simulation methods are well suited to describe the dynamic behavior of a host of systems spanning from nanoscale to mesoscale and in ranges of time spanning from nanosecond to several seconds. They may appropriately describe and predict thermodynamical and mechanical properties, as well as vibrational and rotational spectra and sample conformation space of molecules, macromolecules, liquid, solid-state and supramolecular systems. One of the greatest engines driving the continuous research and improvement of MD methodologies is the needs of the biomedical research sector, ranging from the understanding of biochemical and physiological pathways to drug research and nanobiology.

As for me, I'm a graduate student in Nuclear Physics at the Institute of Technology and Applied Sciences of the University of Havana (InSTEC, in Spanish). Until very recently, my research focused on investigating properties of Doped Graphene Quantum Dots (GQD) relevant to biomedical applications, complementing experimental research at the Cuban Centre for Advanced Studies (CEA, in Spanish).

GQDs are zero-dimensional nanomaterials consisting of one or several stacked flakes of graphene with linear dimensions in the order of nanometers. They thus exhibit quantum confining effects which induce size-dependent photoluminescence, as well as low toxicity and relatively facile synthesis. They thus exhibit a host of potential applications, ranging from thermoelectricity to cancer theranostics.

While Classical MD simulations are unable to model the photoluminescence properties of GQD, as they depend on the electronic structure, it is extremely useful in modelling its interactions and interfaces with macromolecules and supramolecular structures such as lipid bilayers (important for describing the absorption of these particles in the cellular medium) [Figure 5], proteins (showing, for example, inhibition for amyloids such as those responsible for Alzheimer and Parkinson disease), and Nucleic Acids (to study the genotoxicity of these compounds and possibly hint at applications in cancer therapy). Moreover, the synthesis process and mechanism of these compounds, as well as ways to fine tune it to achieve desired properties, is not yet fully elucidated, and Classical MD is the tool of choice to emulate chemical synthesis and explain its results under a range of conditions.

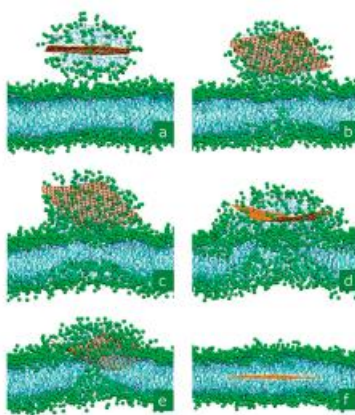


Figure 5. MD simulation snapshots showing stages of the self-insertion of a GQD inside a phosphated lipid bilayer.

While this project was ongoing, I decided to switch research directions, concretely to astrophysics. That is not to say that the knowledge gained during this project is not relevant in this field too, as MD simulations have found a host of applications in this field, ranging from the study of the microscopic behavior of gas and dust clouds in the interstellar medium (in which molecules as complex as ethanol and benzene have been found via spectroscopic analysis, in conditions very different to those found on Earth), to assessing the conditions under which protocells may originate in other astronomical bodies, such as Jupiter's moon Europa or exoplanets.

## **Acknowledgements**

I would like to thank the organizers of the INTEREST program for devising such a great format for projects that enable the learning and application of advanced topics directly linked to research, as well as fostering the development of abilities for remote collaboration. Especially, I would like to express my gratitude to the project supervisor, Prof. Dr. Kholmurzo Kholmurodov, for allowing me to join this project, for destining its time and effort into it, and for its patient guidance. His teaching approach and provision of bibliography contributed decisively to the quality of the learning process. I would also like to thank Dr. Luis Alberto Montero and Dr. Roberto Mulet at UH, as their teaching approach indirectly influenced the style and scope of this report, as well as my theoretical and practical knowledge of MD simulations.

## Bibliography

1. Scopatz A and Huff K 2015 *Effective Computation in Physics* (O'Reilly Media. Sebastopol, CA)
2. Frenkel D and Smit B 2001 *Understanding Molecular Simulation: From Algorithms to Applications vol 1* (Academic. San Diego, CA)
3. Allen M P and Tildesley D J 1989 *Computer Simulation of Liquids* (Oxford University Press. Oxford)
4. Ozboyaci M et al. 2016 *Modeling and simulation of protein– surface interactions: achievements and challenges* Quarterly Reviews of Biophysics, 49, e4, page 1 of 45
5. Nosé S 1984 *A unified formulation of the constant temperature molecular dynamics methods* J. Chem. Phys. 81 511–9
6. Martyna G J, Tobias D J and Klein M L 1994 *Constant pressure molecular dynamics algorithms* J. Appl. Phys. 101 4177–89
7. Kholmurodov K 2011 *MD-simulation in chemical research: from atomic fragments to molecular compound* (Dubná, course notes)
8. Tomobe, K et al 2017 *Water permeation through the internal water pathway in activated GPCR rhodopsin* PloS one, 12(5), e0176876.
9. Mocchi F et al. (2022) *Carbon Nanodots from an In Silico Perspective* Chem. Rev. 2022, 122, 13709–13799
10. García-Vázquez R et al 2019 *Relaxation of ArH<sup>+</sup> by collision with He: Isotopic effects* A&A 631, A86
11. Pohorille A et al. 2001 *Molecular Simulations in Astrobiology*. <https://ntrs.nasa.gov/citations/20010095454>