

# First part

## Slide 1

Good morning.

My name is Emiliano Ippoliti. I work in the Jülich Research Center in Germany and this morning my task is to give you a short introduction on QM/MM approaches.

So far, during this week, you have seen and used methods based on classical physics.

However, if we want to talk about QM/MM, that is a short for Quantum Mechanics/Molecular Mechanics, we need to talk about quantum physics. But I also know that some of you is not so much familiar with quantum physics and quantum mechanics yet. Therefore, in the first part of this lesson I am going to revise some basic elements about quantum mechanics and quantum chemistry, essential to understand, let us say, the big picture. But I would like that you keep in mind from the beginning that if you want to use QM/MM approaches to investigate your systems, what you will see and learn today is absolutely not enough and you will need to do the effort to go more in depth in the theory of computational quantum chemistry if you really plan to employ it in your investigations.

After my theoretical lesson, my colleague Modi Vaibhav has prepared for you some tutorials where you will have the chance to watch QM/MM at work, with practical examples and real calculations, by using the CP2K code.

## Slide 2

This is a rather popular quantum code with many features, some of which I will mention in this lesson, and others will be introduced by Modi in the tutorial.

Summarizing, my intent today, is to provide you with a quick overview of the essential concepts of quantum mechanics and computational quantum chemistry and then shortly introduce QM/MM in order to make easier to understand what Modi is going to show you in the tutorial.

## Slide 3

In particular, here is the outline. In the first part we will discuss about

- when and why quantum mechanics is useful in biology;
- moreover, as just said, I will give you a short recap of the crucial concepts in quantum mechanics and computational quantum chemistry;
- then, we will talk about why we need to introduce hybrid QM/MM approaches;
- and in the second part of this lesson, after the break, we will go more in depth about how to build a QM/MM method, and how quantum and

classical regions can be coupled in a QM/MM scheme, focusing in particular to the approaches implemented in the CP2K code.

## Slide 4

Let us start by answering the question: Why do we need Quantum Mechanics in describing biological systems?

In fact, in the previous days you used approaches, such as force field based molecular mechanics or docking, where the finest resolution is the atom, that is the atoms are described like points in the space, moving according to the classical Newton equations, if the dynamics was of interest.

However, there are many phenomena in nature, including in biology, where such level of detail is not sufficient. Here are some examples:

- First, when chemical reactions are involved (for example if you want to study enzymatic reactions);
- secondly when you want to work with systems containing metal atoms, for which no universal classical parametrization is in general available and ad hoc force field parameters have to be tuned in each specific case;
- third example is when you want to study a phenomenon that involves proton transport, such as in the aerobic generation of ATP or in photosynthesis: in fact, in hydrogen-bonded solvents like water, protons do not diffuse as the other common cations, that is as a random Brownian mass motion due to thermal fluctuations. Instead, the excess proton diffuses via the so-called Grotthuss mechanism, sketched in the picture here at right, that is through the formation and concomitant cleavage of covalent bonds involving neighboring molecules;
- as last example, quantum mechanics is necessary when we need to perform first principle-based predictions of spectroscopic data, such as absorption and fluorescence spectra or even NMR, because empirical parameterizations are unavailable or unreliable.

What do all these examples have in common? The fact that the dynamical behaviour of the electrons inside the atoms is fundamental for a correct description of the phenomenon and cannot be neglected as done for example when we use the force-field approach. Unfortunately, electrons, as well as the lightest nuclei such as protons, cannot be dynamical described through the classical Newton equation and a more complex theory is required, that is quantum mechanics.

## Slide 5

The fundamental equation of quantum mechanics is the so-called Schrödinger equation, here at the top of the slide. Calculate the quantum properties of a system, implies to solve the corresponding Schrodinger equation, which you can consider it the equivalent of the Newton equations in the classical world.

The unknown variable of the Schrödinger equation is  $\Psi$ , the so-called wave function of the entire quantum system. It is a function of the coordinates of the quantum elements in the systems, that is electrons and nuclei for a molecular system, and the time.

Knowing the wave function of the system at a certain time allows us to compute the properties of that system at that time by solving in principle an integral like this.

The squared modulus of  $\Psi$  is proportional to the charge density distribution of the system.

$H$  in the Schrödinger equation is instead the so-called Hamiltonian and it represents the physics of the system. It is the equivalent of the force field at classical level, that is it contains the interaction energy terms of electrons and nuclei.

For example, here I reported a typical Hamiltonian used to describe molecular systems in quantum chemistry, where now small and capital "r" represent electronic and nuclear coordinates, respectively. From left to right you can recognise

- the kinetic term for the nuclei,
- the kinetic term for the electrons,
- the Coulomb interaction terms between electrons,
- between electrons and nuclei,
- and finally, the Coulomb interaction terms between the nuclei.

The Schrödinger equation is mathematically extremely complex to solve. Just to give you an example, with even only one particle, so  $N$  equals to 1, and a non-so-complex Hamiltonian, exact analytical solutions are not available. Therefore, for the many-body systems we usually deal with, the Schrödinger equation can only be solved approximately by numerical solutions generated via computers and many approximate approaches have been devised in the years for this aim.

## Slide 6

In particular, if we are interested only in the properties of a system and not to its dynamical behaviour we can prove that it is enough to solve a slightly simpler equation, called time-independent Schrödinger equation, where now the wavefunction  $\Psi$  is independent from the time. Here at the right I listed some of the many schemes developed to approximately solve this simpler equation: Hartree-Fock theory, Couple Cluster, Density Functional Theory, etc

When instead we are also interested to the dynamics of the quantum system, the full time-dependent Schrödinger equation needs to be solved and some different approaches are nowadays available to approximately find the solutions to this equation. Are these the so-called *ab initio* molecular dynamics schemes, such as the Ehrenfest, the Bohr-Oppenheimer or the Car-Parrinello molecular dynamics schemes.

All those schemes share the assumption or better the approximation that the motion of atomic nuclei and the one of the electrons in a molecule can be treated separately, but also in addition that the nuclear motion can be considered as a classical motion.

This mostly because the masses of the nuclei are at least 3 orders of magnitude larger than the electronic mass, and thinking classically the nuclei are much slower than the electrons.

This assumption is often collected under the name Born-Oppenheimer approximation, not to be confused with the molecular dynamics scheme just mentioned, even if nowadays this name refers to a more technical aspect of the assumption that I cannot describe here. In the large majority of the molecular systems, including the typically large biological systems, this approximation is well verified and can be safely employed.

I do not have time to enter in much more details about the different ab initio molecular dynamics schemes, but I want to briefly mention how the Bohr-Oppenheimer dynamical scheme works, because that is implemented in the CP2K code and you will use it in the tutorials.

## Slide 7

In this slide I wrote down the two equations that describe the Bohr-Oppenheimer molecular dynamics scheme.

The first thing to note is the separation between the electronic degrees of freedom, first equation, and the nuclear degrees of freedom, second equation.

The second thing to note is that the electronic problem does not evolve in time, it is a time-independent Schrödinger equation, while the nuclear degrees of freedom evolve in time as classical entities, that is according to a Newton-like equation: mass times acceleration (the two dots over capital R mean second derivative with respect time) equals to minus the gradient of quantity that represents the potential felt by the nuclei.

Schematically, the algorithm associated to this molecular dynamics approach proceeds this way: at each time step a time independent Schrödinger equation involving only the electronic degrees of freedom is solved via some electronic structure method like the ones mentioned before (Hartree-Fock, Density Functional Theory, etc). Note that in the  $H_e$  Hamiltonian in this equation, the nuclear coordinates capital R are not dynamical variables but just parameters: in this approximation, the electrons move within a static electric field due to the presence of the nuclei.

The electronic wavefunction  $\Psi_0$  found solving this time-independent Schrödinger equation, is used in the next step of the algorithm, to calculate the forces on the nuclei via the right-hand side of the second equation. In fact, the forces are obtained as minus the gradient of a potential that depends on the electronic wavefunction  $\Psi_0$ .

Finally, having obtained the forces, the nuclei are moved according to a Newton-like equation, and the cycle can start again for the new step by using the new nuclear coordinates capital R just found as the new parameter of the time independent Schrödinger equation.

Note that the electronic problem in the first equation has an infinite number of solutions, each one corresponding to a different energy state. However, among those solutions here we are interested to the wavefunction corresponding to the state with the smallest energy, i.e. the ground state, as indicated with the subscript "0".

Instead the "min" symbol in the second equation refers to the fact that the electronic problem consisting of solving the time-independent Schrödinger equation, i.e. of finding  $\Psi_0$ , can be recast in a variational problem, i.e. a problem of finding a minimum. This is the so-called wavefunction minimization or optimization procedure, and on a computer, it is computationally more convenient to implement it than any other algorithm that tries to solve directly the time-independent Schrödinger equation. We will see an example of this minimization algorithm in a few slides.

Let's now focus on the methods to solve the first equation.

## Slide 8

As I mentioned before, many methods have been devised in the years to approximately solve the time-independent Schrödinger equation. Some are very accurate and computational expensive, others are computationally less demanding but also limited in accuracy.

However, almost every quantum code used by computational biophysicists and biochemists implements the Density Functional Theory, including CP2K, the code you will use in the tutorial. In fact, this relatively recent approach represents probably the best compromise between accuracy and computational cost and it is therefore currently one of the very few approaches that offers the possibility to deal with systems of order of hundreds of atoms with sufficient accuracy.

## Slide 9

Let's briefly describe the Density Functional Theory. It is based on the following two theorems proved by the physicists Hohenberg and Kohn:

- First, the ground state energy (and therefore the ground state properties) of a many-electron system is a unique functional of the electronic density  $\rho$ . Here functional means a function of another function. In fact, the density  $\rho$  is a function of the three space coordinates. In each point of the 3D-space you have a value of the electronic density.
- Second, the functional for the ground state energy  $E_\rho$  is variational, in the sense we have mentioned before.

The benefit to use this method is that instead to calculate the wavefunction  $\psi_0$ , which depends on all the electronic coordinates, the properties of the system depend only on  $\rho$ , which in turn depends only on 3 coordinates, the spatial coordinates.

The drawback is that the functional  $E_\rho$  is not known, and therefore, as it is, the theory cannot be used in practice.

## Slide 10

Luckily, the physicists Kohn and Sham in 1965 had the idea to recast the problem in order to make Density Functional Theory a practical method. Their idea is simple: they hypothesized a fictitious system of non-interacting electrons with a local potential that generates by construction, the same electronic density as the one of the real fully-interacting system. In this way the problem to find the density  $\rho$  of the real fully-interacting system that minimizes the functional  $E$  of  $\rho$  is recast to the problem of solving  $N$  single-electron equations, much easier to solve than an equation of  $N$ -electrons. And above all, all the terms in the single electron equations are known apart from one, the so-called exchange-correlation functional.

In these Kohn-Sham equations the  $\phi_i$ 's represent the single electron wavefunctions, not to be confused with the  $\psi_0$  in the previous slide, that is the wavefunction of the entire system formed by  $N$  electrons. And  $\rho$  can be obtained from the  $\phi_i$ 's by using the first relation in the slide.

In practice, quantum chemists have proposed many recipes to approximate the unknown exchange-correlation functional  $E_{xc}$ , for example by calculating it for the simplest cases, such as the homogeneous electron gas, or by fitting experimental data.

When you have to specify the level of theory you are going to use to solve the electronic problem with Density Functional Theory you need also to state explicitly the exchange-correlation functional to be employed.

Having decided which exchange-correlation functional to use, how do we get the electronic density  $\rho$ ?

## Slide 11

In other words, how does the algorithm that solves the system of  $N$  single-electron Kohn-Sham equations work?

This is done through an iterative procedure because the Kohn-Sham equations are nonlinear, which means that some terms in the equations depend on the electronic density itself, that is on the solutions.

The iterative procedure to solve these equations can be summarized this way:

- First, we start with an arbitrary electronic density in order to define completely the equations to be solved;
- then, we find the  $\phi_i$ 's, i.e. the single electron wavefunctions that can be used to get the new electronic density by using this equality
- and we measure somehow the difference between the new and the previous density. If the difference is below some predetermined threshold we considered the new density already converged and we stop the iterations, otherwise we take the new density and we go back to the first step.

This self-consistent approach is very commonly used in quantum chemistry and in general when you have to solve nonlinear equations like the Kohn-Sham equations.

## Slide 12

All the equations we have seen so far, including the KS equations are continuous equations. Their solutions, as for example the  $\phi_i$ 's in the Kohn-Sham equations are functions defined on the space. However, to put such a problem on a computer, we need to discretize it. How can we discretize the problem of solving the Kohn-Sham equations?

This is done by expanding the wavefunctions (or the density) over a finite set of known functions, and we refer usually to this set of functions as the basis set. This way the problem to solve a continuous differential equation is recast in the problem to diagonalize a matrix and find eigenvalues and eigenvectors.

When one wants to specify the type of quantum chemistry calculation is going to perform on a computer, they need to specify both the level of theory (e.g. DFT together with the chosen exchange correlation functional) and also the employed basis set. This defines completely the employed level of theory.

Commonly, two classes of basis sets can be identified:

- the localized basis sets, such as the atom-centered Gaussian functions, very suitable to describe the wavefunction of localized objects like molecules,
- and nonlocal basis sets, such as the plane waves, which were originally employed to describe the wavefunctions of condensed matter/solid state systems.

Both types of basis sets have advantages and disadvantages. In the slide I listed some of them but I do not want to enter now in more details.

Really, the code that we will use in the tutorial, CP2K, implements also a more sophisticated approach, which combines both classes:

## Slide 13

this is the so-called hybrid or dual Gaussian and Plane Wave method, in short GPW.

The method uses an atom-centered Gaussian-type basis to describe the wave functions, but also an auxiliary plane wave basis to describe the density.

Using a plane wave basis set for charge density means using grids in real space to represent the charge density. In fact, by a mathematical operation called Fourier transform, that computationally can be performed in a very efficient way on a grid with an algorithm called Fast Fourier Transforms (FFT), one can pass from the representation on the real space grid to the representation on the reciprocal space, that is the G space of the plane waves. Finer grids, that is with smaller cells, correspond to larger cutoffs in the reciprocal space.

What is the advantage of this dual representation, that is localised basis set for the wavefunctions and nonlocal basis set for the density? The advantage is mainly on performance or better scaling: with a density represented as a sum of plane waves or, which is the same, on a regular grid, the efficiency of the Fast Fourier Transforms algorithm can be exploited to obtain the long-range energy terms in a time that scales linearly with the system size, thus circumventing one of the major bottlenecks of standard Gaussian-based calculations.

We will come back on this in more details later today.

As you will see at the tutorial, in order to set up a calculation in CP2K that uses the GPW approach, you will have to provide information on both the Gaussian Type Orbitals and the plane waves basis set to be used.

## **Slide 14**

The largest systems investigated so far via full quantum mechanical approaches, that is by describing the entire molecular system through quantum mechanics, include less than 10,000 atoms.

In contrast, typical sizes of biological systems are much larger than 10,000 atoms.

Therefore, investigating interesting biological systems at full quantum mechanical level is beyond the current state-of-the-art hardware and software technologies. But as we have seen at the beginning, there are cases where a quantum mechanical resolution is required, also for large biological systems. This implies that for these cases at present the only viable way is to resort to multiscale approaches, as for example the hybrid quantum mechanical/molecular mechanics one.

## **Slide 15**

In fact, in the biological systems the region where the electronic description is necessary is usually a spatially limited area of the system (e.g. the region where

the chemical reaction takes place), and this feature makes a QM/MM approach very suitable for these systems.

This because, in the QM/MM approach the system is fictitiously separated in two parts that are described at different levels of theory:

- A small part, the QM or quantum part, usually the chemically active region or in general the region where the electronic degrees of freedom are important, that is treated at quantum level by computationally demanding electronic structure methods, as for example density functional theory.
- And the rest of the system, which contains atoms that for example do not directly participate in the reaction, that is instead described efficiently at a lower level of theory, usually by classical force fields. This part is usually referred as the MM or classical part.
- A QM/MM interface is the part of the code, or a standalone code that couples in a coherent way the two different resolutions.

OK, we have reached the end of the first part of the lesson. In the second part, we will go more in depth on how the coupling between the quantum and the classical regions can be done, and we will describe different QM/MM approaches, including the one implemented in the CP2K code that you will see in the tutorial.