

Chapter 7

Ab initio calculations

7.1 Small brains and big computers

The theoretical prediction of molecular structures and properties can be very valuable to (a) obtain an idea about stable molecular structures, (b) explore whether reactions are exothermic or endothermic and (c) predict molecular properties such as spectroscopic transitions or thermodynamic constants. There is an abundance of program packages to calculate molecular structures and properties, some of which are available for free via academic licenses. Some shall be discussed here to give a quick introduction in the use of semi-empirical and *ab initio* calculations.

Unfortunately, it is impossible to condense the theoretical methods behind common *ab initio* program packages into a small script or into a short lecture. At the same time, it can be dangerous to rely on black-box calculations because the quality of the output of the programs always relies on a meaningful input. Phil Bunker once summarized this problem in the words "small brain, big computer calculations", complaining that few people nowadays understand the theoretical underpinnings of the complex theoretical methods they use.

So, while the description here may be enough to generate nice pictures and some useful data on molecules of interest, please be aware that the programs don't distinguish between reasonable and unreasonable calculations. For a serious theoretical approach, a much deeper insight into the methods may be necessary. This chapter only serves for typical small brain, big computers calculations. The small brains, however, can gain a lot of insight and nice (and even publishable) structures from the approach described here. If a starting geometry is unreasonable, then the calculations won't converge or give unreasonable output. If the basis set is too small or the level of theory is not sufficient, then the results will off target in a more subtle fashion. If in doubt, ask a specialist to comment on your approach.

7.2 Programs

I suggest to install a graphical structure editor (e.g. ChemsSketch, Arguslab or WxMacMolPlt). ChemsSketch (<http://www.acdlabs.com/download/>) offers the possibility to draw a 2D chemical representation and automatically converts the molecule into a 3D structure. Arguslab (<http://www.arguslab.com/>) is a more cumbersome editor for the drawing of 3D structures and offers molecular-mechanics and semi-empirical structure optimization to optimize the structures. WxMacMolPlt (<http://www.scl.ameslab.gov/MacMolPlt/>) is rather a viewer for *ab initio* results than an editor, but also offers a molecule builder mode.

Mopac (<http://openmopac.net/>) is a very fast program for semi-empirical calculations and offers a wealth of options for the calculation of molecular properties. Semi-empirical methods reduce the time-consuming calculation of *ab initio* energies by partially replacing them with empirical factors which were scaled to reproduce known molecular properties. Be aware, that this approach may fail miserably if the problem/molecule at hand does not conform to the problems the programmers had in mind when creating their method. Due

to their speed, semi-empirical methods are still very useful to:

- Obtain a rough overview about molecular structures and properties.
- Quickly pre-optimize structures for higher-level ab-initio calculations.
- Generate reasonable molecular structures for presentations (cite the method).
- Obtain structures / properties for large systems which are inaccessible to higher-level theory.

PCGmess (<http://classic.chem.msu.su/gran/gmess/>) can be used for serious *ab initio* calculations. PCGmess is a version of Gmess which has been optimized to run on typical desktop computers. Due to the fast development of computers, a fairly cheap computer offers enough power for quite serious calculations. As most computations scale with some power of the number of atoms, it is also obvious that modern supercomputers only offer the possibility to moderately increase the size of a investigated system at the same level of theory as compared to a simple desktop system. Batch files can be very useful to start the PCGmess program, because it requires a command-line input. Two example batch files "PCGmess.bat" and "PCGmess2.bat" illustrate the calling of PCGmess. Edit the batch files to suit your needs and to point to the correct directories on your computer.

OpenBabel (<http://openbabel.org>) is a "translation" program for the quick conversion of molecular geometry data from one to another programs. Unfortunately, it may be necessary to copy/paste the relevant sections of an output file into Openbabel to ensure that the right coordinates are recognized.

WxMacMolPlt (<http://www.scl.ameslab.gov/MacMolPlt/>) is a very nice program to view the output of Gmess calculations, to follow the energy convergence of a calculation, or to visualize molecular orbitals or vibrations. If the Gmess output always has the same extension (e.g. ".out"), it may be useful to set corresponding file associations to open the output with WxMacMolPlt, or in a text editor like notepad (I use a friendlier equivalent called Metapad).

7.3 calculating molecular properties

7.3.1 Creating a molecular structure

I propose using ChemsSketch to draw a 2D structure and generate the 3D structure with "Tools → 3D Structure Optimization". Templates may be used to quickly draw common structures. Save the molecule as .mol file (e.g. "Adenine.mol", "Phenylalanine.mol") Convert .mol into .xyz using Openbabel. The example of phenylalanine in 2D and 3D representation is shown in Fig. 7.1.

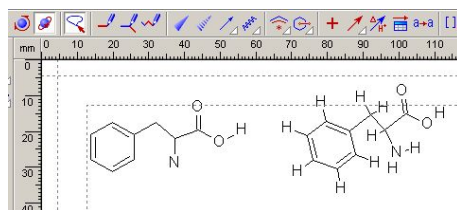


Figure 7.1: The amino acid phenylalanine drawn in ChemsSketch in 2D (left) and after "3D-optimization" (right). Note that the amino and carbonyl group don't seem to interact with each other or the aromatic ring but seem to have some random orientation relative to each other, the structure is not realistic.

Alternatively, use Arguslab (cumbersome editing, good viewer). The view in Arguslab is shown in Fig.. select Tools → Builder Tool, select element, and click the atoms onto the screen. Use Edit → Translate/Rotate (also menu button) to move/rotate molecule on screen. Select existing atoms or groups and use Edit → Attach Selection to Manipulator to rotate/translate selected atoms versus unselected atoms. Maybe

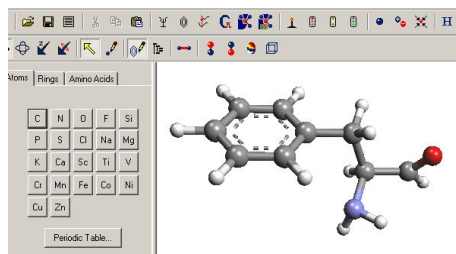


Figure 7.2: The amino acid phenylalanine in Arguslab after PM3 geometry optimization. There seems to be an electrostatic interaction between the hydrogens of the amino group and the carbonyl group which is reasonable.

optimize the geometry with a molecular mechanics or Semiempirical method (Calculation \rightarrow Geometry), but this is slow! Save the resulting structure as .XYZ file.

Now we have a file with the xyz coordinates of our molecule, e.g. "Adenine.xyz", "Phenylalanine.xyz".

7.3.2 Semi-empirical calculations

First we need to optimize our structure which is very fast in Mopac. Load the xyz file generated above into an editor (Notepad). Add 3 header lines, so the coordinates start in line 4. The first line contains all commands determining which method is used and what information the output should contain. The next two lines can be used to describe the calculation for future reference. Save the Mopac input file with the suffix .mop, e.g. "Adenine.mop", "Phenylalanine.mop". Important keywords are:

- AM1 - Use the AM1 hamiltonian
- PM3 - Use the MNDO-PM3 Hamiltonian
- PM6 - Use the MNDO-PM6 Hamiltonian
- MNDO - Use MNDO Hamiltonian
- CHARGE=n - give charge of system if $\text{---}=\text{0}$
- SINGLET - force singlet state
- TRIPLET - force triplet state
- AIGOUT - creates a Gaussian/Gamess compatible Z-matrix structure output (convenient alternative: use OpenBabel for XYZ style conversion unless you need internal coordinates)
- FORCE - calculate force matrix (vibrations, rotations)
- THERMO - calculate thermodynamic properties (vibrations, rotations, zero-point energy, heat of formation)
- POLAR - calculate polarizabilities

An example input file for a geometry optimization using the xyz coordinates from Phenylalanine.xyz is Phenylalanine.mop:

I created a file association of the .mop file with mopac (right-click, open with \rightarrow select program; now search the path to your MOPAC program and make it the default option). After the program ran, the output file (.out) gives the TOTAL ENERGY of formation, the IONIZATION POTENTIAL, the DIPOLE moments, and most importantly the CARTESIAN COORDINATES. But the method did not seem sufficient to identify a clear low-energy structure (phenylalanine seems to be quite floppy). in

```

PM6
Phenylalanine - Geometry optimization with PM6 semiempirical method
D:\TS\abinitio\mopac\Phenylalanine\
N      7.59330      -8.81560      -0.80340
C      7.41220      -7.76390      -1.81460
C      8.63880      -7.72410      -2.72860
C      9.86880      -7.40030      -1.91390
C      10.49450     -8.39260      -1.17550
C      11.62270     -8.00330      -0.12880

```

Figure 7.3: Input file for MOPAC geometry optimization with the PM6 semi-empirical method

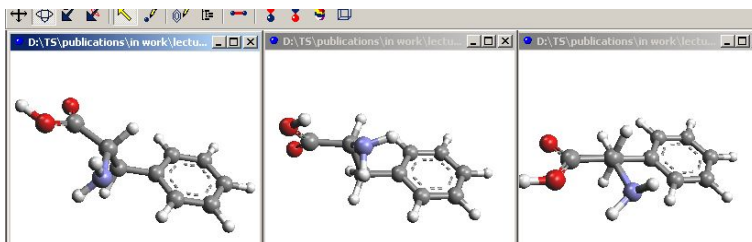


Figure 7.4: Two geometries optimized with the PM6 method in Mopac (left) and a starting geometry (right). A tight convergence criterion (SCFCRT=1.D-8) helped to find local minima due to the interaction of amino and carboxylic acid, but there are several minima!

I took the cartesian coordinates from an optimization, converted them in OpenBabel from 'mopout' to 'xyz' and used them to build a new input file "Phenylalanine2.mop" for a force calculation:

The new output file "Phenylalanine2.out" contains ROTATIONAL CONSTANTS and all information about the VIBRATIONAL modes and frequencies. The vibrations can be visualized in the visualization program Molden (somewhat awkward in Windows).

7.3.3 High-level ab initio methods

PCGauss offers reasonably fast and robust ab initio methods for desktop computers. Ensure that you have the correct program version for your processor and some free memory (more is better) and disk space.

Convert the xyz data from the drawing program or the semiempirical optimization into Gauss input format using OpenBabel. A [space]\$CONTROL group is needed to specify the type of calculation. The following is an example control group for a restricted Hartree-Fock calculation using the B3LYP DFT method to optimize a geometry which is given in cartesian coordinates: " \$CONTRL SCFTYP=RHF DFT-TYP=B3LYP5 runtyp=optimize COORD=CART EXETYP=run \$END". Please note that each Keyword group begins with [space] \$groupname and ends with [space] \$END. The Keyword group can span several lines, but each line may not exceed a certain length. We also need to specify a basis set, e.g. the small 3-21 Gaussian basis set: " \$BASIS GBASIS=n21 NGAUSS=3 \$END". The starting orbitals can be guessed: " \$GUESS GUESS=Huckel \$END". Now we need the coordinates which are in the group \$DATA, preceded by one free descriptive line and a symmetry label. For our Phenylalanine, the data group without symmetry (C1) is: " \$DATA Phenylalanine geometry optimization with DFT, 3-21G basis set C1 N 7.0 7.33240 -9.08680 -0.87340 ... \$END"

To run the calculation, it is necessary to rename the input file into "input" and call the PCGauss program from the command line. I created a batch file "PCGauss.bat" which automatically renames and runs an input file which be dragged/dropped onto the batch file (or an extension for input files can be associated with it). To use such a batch file, the script must of course point to the correct program location, so some editing is required. My batch file deletes all old working files after the run which may not be desirable, e.g. if a PUNCH file is needed to restart a calculation. During or after running the calculation, it is possible to look at the output file. The structures and many properties can be visualized with wxMacMolPlt. If the

```

PM6 FORCE]
Phenylalanine - Force calculation with PM6 semiempirical method
D:\TS\abinitio\mopac\Phenylalanine\
N      7.33240      -9.08680      -0.87340
C      7.41250      -7.90410      -1.76050
C      8.67630      -7.97210      -2.64630
r      q 888900      -7.506400      -1.884000

```

Figure 7.5: Input file for MOPAC force calculation at the optimized geometry from Fig. 7.5

calculation does not converge within a limited number of steps (specify in the \$STATPT group with the keyword NSTEP=*n*), the calculation must be restarted with the best unconverged geometry (search in output file from the bottom, or extract the data from the PUNCH file). Be aware that the output files are large and it is useful to search for the desired section as opposed to simply browsing through the file.

The program wxMacMolPlt can be used to inspect the results (see Fig. 7.6) and also includes an input builder which can help to find the right keywords for your calculation. A coordinate window can be used to edit the geometry and is convenient to get from XYZ coordinates to internal coordinates.

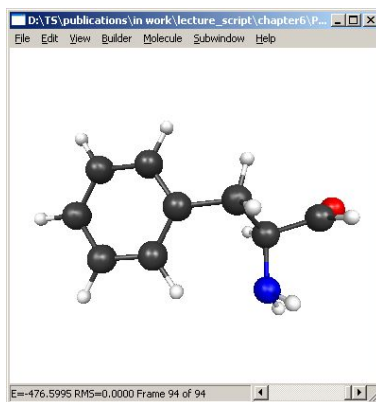


Figure 7.6: Optimized phenylalanine geometry from a PCGauss B3Lyp calculation as displayed in WxMacMolPlt.

For a detailed keyword description for PCGauss, look up the INPUT manual in the documents section. The following are some notes about the general use of PCGauss:

- It is advisable to optimize geometries first with a lower level method / smaller basis sets to save time.
- Find the unconverged/converged geometry below the label ****** THE GEOMETRY SEARCH IS NOT CONVERGED! ****** or ******* EQUILIBRIUM GEOMETRY LOCATED ******* (search for keyword Geometry from bottom up) to set up the next calculation.
- To restart an unconverged calculation, copy the last/best coordinates from the output file or punch file. The punch file also contains a \$VEC — \$END group which may be copied into the input file and read instead of guessing Huckel Orbitals (→ GUESS=MOREAD).
- The 6-31G basis set is often considered as the minimal reliable or standard basis.
- To account for 'weak' or long-range interactions such as van-der-Waals forces or induced dipoles, consider optional P polarization functions for H, D-polarization functions for N,C, (NPFUNC=1 NDFUNC=1) and diffuse L (sp) shells for N,C, and diffuse S shells for H. But be aware, that the larger basis set make the calculation more costly and can hamper the convergence. The inclusion of polarization/diffuse functions is usually indicated by adding */+ to the basis set name.

- The result from Hartree-Fock calculations can sometimes be improved by taking the electron correlation energy into account. PCGamess offers second-order Moller-Plesset perturbation theory (MPLEVEL=2 in \$CONTRL).
- Run a RUNTYP=HESSIAN job to get vibrational frequencies
- A TDDFT calculation gives fast excited state energies/properties at converged ground state geometries (CITYP=TDDFT in \$CONTROL and \$TDDFT NSTATE=5 ISTSYM=0 ISTATE=1 \$END for the 5 lowest states of symmetry 0).

The choice of theoretical methods is important to obtain a useful result. Unfortunately, most *ab initio* programs are quite limited in the number of methods they offer, so the choice of method is a compromise between the desirable and the possible. In PCGamess, the following methods may be useful:

- DFT (e.g. DFTTYP=B3LYP5) for fast and good structure / energy calculations of closed shell molecules. DFT will not give useful results for systems bound by van-der-Waals forces and is therefore not useful to investigate noncovalent cluster structures (unless you are certain that the structure must be hydrogen-bound). Because DFT is fairly cheap, it is often possible to explore the effect of larger basis sets.
- TDDFT for excited state calculations. All caveats from DFT apply.
- *Ab initio* HF methods for everything else. ROHF and UHF must be used for open shell or high-spin systems. Moller-Plesset perturbation theory (e.g. MP2) may improve the result.
- Excited states can be calculated with multiconfiguration methods (MCSCF), complete active space calculations (CASSCF) or configuration-interaction methods (CI), but the input must specify the relevant basis for the excitation which can be tricky.

What else to do:

- Test if electron correlation improves the result; MP2 (MPLEVEL=2 in the control group) allows to include some electron correlation (other options are MCSCF and CI).
- Start with a small basis set and increase it. If the basis is large enough, the energy will no longer decrease with a bigger basis. Special basis sets can be downloaded from the net (e.g. the basis set exchange at <https://bse.pnl.gov/bse/>) and specified for each atom.
- If the molecule contains floppy groups, you may want to look for different potential minima (different stable geometries) by generating different initial geometries. Fast (semiempirical) Methods can help you to quickly test many starting geometries.
- Always compare your results to available literature data! Look for published rotational / vibrational frequencies, ionization potentials, excited state energies. If you are unsure about the theoretical approach, use it to calculate a similar system with known properties first.
- It is possible to model the effect of a solvent environment by introducing a polarization continuum (e.g. for water: `COSGMSEPSI = 78.4END`) – but be aware that nobody quite knows how to do model solvation correctly.

7.3.4 Examples

I prepared some examples to give an idea about different types of input files. All example calculations are performed on phenylalanine.

First some example calculation in Mopac. All calculations ran without any problem on a P4 computer with 2.5 GB of Memory, only the reaction coordinate runs took more than a few seconds.

- Input "Phenylalanine.mop": Optimize the structure of phenylalanine with the PM6 method and starting from a structure drawn in Chemskech. The convergence criterium was increased (SCFCRT=1.D-8) to find a minimum geometry for the "floppy" amino acid residue. The optimized structure shows the hydrogens from the NH₂ pointing towards the carboxy OH ("isomer A"). The calculated final energy is -2014.437 eV. The calculated vertical ionization potential of 9.41 eV is much higher than the experimental value of 8.8 eV (7).
- Input "Phenylalanineo.mop": Optimize the structure of phenylalanine with the PM6 method and starting from the structure optimized above, but with the OH group inverted to form a OH..NH₂ hydrogen bond ("isomer B"). The optimized structure shows the expected hydrogen bond, but the final energy is slightly higher (-2014.372 eV). The calculated vertical ionization potential of 9.80 eV is far above the literature value of ~ 9 eV.
- Input "Phenylalanine2.mop": Force calculation at the optimized geometry from "Phenylalanine.mop". The calculation gives unrealistic low stretch frequencies of 2538 cm⁻¹ for O-H and 2830, 2790 cm⁻¹ for N-H (literature: 3581, 3420 and 3340 cm⁻¹ for a similar isomer(6)).
- Input "Phenylalanine_convert_coordinates.mop": Obtain internal coordinates for a reaction coordinate calculation. The resulting internal coordinates are used below to calculate the O-H dissociation coordinate
- Input "Phenylalanine_H-loss.mop": Calculate the potential energy surface for H-loss of the acidic hydrogen by optimizing the geometry for an increasingly large O-H distance. The hydrogen is atom 13 (H) bound to Atom 12 (O). The result is surprising, instead of a Morse-type binding potential we find numerous minima. Fig. 7.7 shows what happened: Instead of dissociating, the hydrogen atom migrates to the next heavy atom and finds many local minima before leading to molecular dissociation.
- Input "Phenylalanine_H-loss_COSMO.mop": Same as above, but using the COSMO solvent model to simulate water as a polarizable continuum (Fig- 7.8). The H-atom still hops to the neighboring O-atom, but then dissociates as expected. Note how the energy of the molecule is stabilized by the polarizability of the environment.

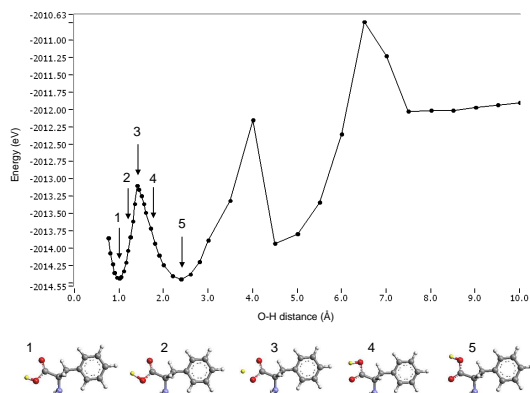


Figure 7.7: Potential energy surface as a function of the O-H distance. Instead of a simple H-loss, the Hydrogen migrates to the neighboring O-atom (structures 1-5), then to the phenyl ring (distance 4.5 Å) before leading to a somewhat unrealistic loss of a 2H-toluene isomer.

Some example calculations in PCGmess, using some pre-optimized geometry from the first Mopac run.

- Input "Phenylalanine_geometry_B3LYP_3-21G.inp": B3LYP geometry optimization for isomer B with a small 3-21G basis set. The calculation predicts a much stronger H-bond than the semi-empirical

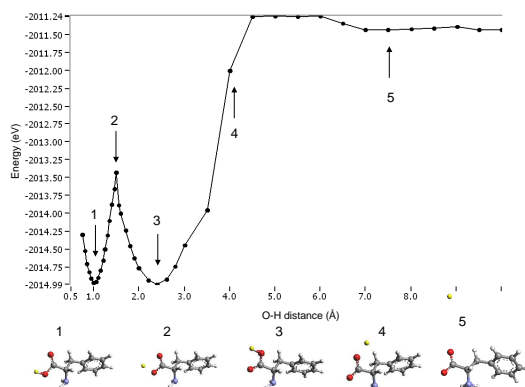


Figure 7.8: Phenylalanine potential energy surface as a function of the O-H distance using the COSMO solvent model to simulate water as a polarizable continuum. Instead of a simple H-loss, the Hydrogen migrates to the neighboring O-atom (structures 1-5) and then dissociates.

Mopac calculation. The calculated absolute energies are not comparable to those from the semi-empirical calculation.

- Input "Phenylalanine_geometry_B3LYP_6-31G.inp": B3LYP geometry optimization with 6-31G basis set. Compared to the calculation with small 3-21 basis set, the energy is lower by 79 eV! But the structure is similar to the cheap calculation.
- Input "Phenylalanine_geometry_B3LYP_6-31Gpd.inp". B3LYP geometry optimization with 6-31G** basis set (6-31G plus additional polarizable P and D orbitals). Compared to the calculation with 6-31G basis set, the energy is lower by 4.9 eV and the H-bond is slightly weaker.
- Input "Phenylalanine_geometry_B3LYP_6-31Gpd_diff.inp". B3LYP geometry optimization with 6-31G**++ basis set (6-31G** plus additional diffuse S and (sp) orbitals). Compared to the calculation with 6-31G** basis set, the energy is lower by 0.73 eV but the structure is almost identical.
- Input "Phenylalanine_TDDFT_excited_states_6-31Gpd.inp". B3LYP TDDFT calculation of excited states with the 6-31G** basis set. The vertical excitation energy of 5.38 eV and 5.50 for the first two excited states is in the range of the reported first absorption band (8).
- Input "Phenylalanine_HF_6-31Gpol.inp". HF geometry optimization with the 6-31G** basis set. The vertical excitation energy of xxx eV
- Input "Phenylalanine_HF_6-31Gpol.inp". HF geometry optimization with the 6-31G** basis set. The vertical excitation energy of xxx eV
- Input "Phenylalanine_HF_MP2_6-31Gpol.inp". MP2 geometry optimization with the 6-31G** basis set.

Bibliography

- [1] MOPAC2007, James J. P. Stewart, Stewart Computational Chemistry, Version 8.211W web: [HTTP://OpenMOPAC.net](http://OpenMOPAC.net).
- [2] Alex A. Granovsky, PC GAMESS/Firefly version 7.1.E, <http://classic.chem.msu.su/gran/gamess>.
- [3] Bode, B. M. and Gordon, M. S. J. Mol. Graphics Mod., 16, 1998, 133-138, <http://www.scl.ameslab.gov/MacMolPlt>.
- [4] ArgusLab 4.0.1, Mark A. Thompson, Planaria Software LLC, Seattle, WA, <http://www.arguslab.com>.
- [5] The Role of Databases in Support of Computational Chemistry Calculations Feller, D., J. Comp. Chem., 17(13), 1571-1586, 1996.
- [6] L. C. Snoeka, E. G. Robertsona, R. T. Kroemerb and J. P. Simons, Chem. Phys. Lett. **321**, 49 (2000).
- [7] Kang Taek Lee, Jiha Sung, Kwang Jun Lee, Young Dong Park, Seong Keun Kim, Angewandte Chemie Int. Ed., **41**, 4114 (2002).
- [8] Takayo Hashimotoa, Yuichi Takasua, Yuji Yamadab and Takayuki Ebata, Chem. Phys. Lett **421**, 227 (2006).