Ab initio electronic structure calculations on the Grid

***M. Sterzel* – ACC "Cyfronet" AGH**

COST Training School on Molecular and Material Science Grid Applications

Trieste, 15-18 September 2008

www.eu-egee.org

Outlook

- **Aim of this talk**
  - To demonstrate benefits of Grid computing in chemistry
- **Parts of the talk topics**
  - Work on the Grid vs. cluster
  - Chemical Software packages available on the Grid
    - Parallel execution
  - Typical Chemical Computations on the Grid
    - Geometry optimizations
    - Numerical Frequencies
    - Chemical reactions
  - Final Remarks

Authorization

- **On a cluster - password**

```
okrucyusz:~ sterzel$ ssh ymsterze@ui
Last login: Wed Mar  5 18:40:24 2008 from ha2rtr.agh.edu.pl
[ymsterze@ui ymsterze]$
```

- **On the grid - certificate (Grid identity card)**
  - User has to be a member of Virtual Organization (at least one)
  - Virtual Organization determines the resources user can use
  - To access grid resources user needs to obtain proxy:

```
[ymsterze@ui ymsterze]$ voms-proxy-init -voms gaussian
Your identity: /C=PL/O=GRID/O=Cyfronet/CN=Mariusz Sterzel
Enter GRID pass phrase:
Creating temporary proxy ............................... Done
Contacting  voms.cyf-kr.edu.pl:15001[/C=PL/O=GRID/O=Cyfronet/
   CN=voms.cyf-kr.edu.pl] "gaussian" ...................Done
Creating proxy ......................................... Done
Your proxy is valid until Thu Mar  6 07:20:50 2008
```

Job file

- Grid JDL file

```
Executable    = "/bin/bash";
Arguments     = $VO_GAUSSIAN_SW_DIR/g03/gaussian.x water.com";
JobType       = "MPICH";
NodeNumber    = 4;
StdOutput     = "water.out";
StdError      = "water.err";
InputSandbox  = {"water.com"};
OutputSandbox= {"water.out",
                 "water.log",
                "water.err"
                };
```

- PBS file

```
#!/bin/bash
#PBS -l ncpus=4
#PBS -q long
#PBS -o water.out
#PBS -e water.err
#PBS -N my_job_name
#PBS -M my@email
#PBS -m e
export g03root=/somewhere
. $g03root/g03/bsd/g03.profile
$g03root/g03 water.com
```

Job management

- **PBS**
  - `qsub` – submit job to a queue
  - `qdel` – delete job from a queue
  - `qstat` – show job status in a queue

- **EGEE Grid**
  - `glite-wms-job-submit` – submit job to the Grid
  - `glite-wms-job-delete` – remove job from the Grid
  - `glite-wms-job-status` – show status of the job on the Grid
  - `glite-wms-job-output` – retrieve job files from the Grid

… just a few new commands…

Chemical software

- **Freely available packages on EGEE Grid:**
  - GAMESS
  - DALTON
  - CPMD
  - Newton X
  - DL_POLY
  - NAMD
  - RWAVEP
  - GROMACS
  - Autodock
  - Tinker
  - Solvate
  - PIC-DMSC
  - MCGBgrid
  - QMC
  - ABCtraj
  - VENUS
  - CRBS
  - LM
  - COLUMBUS
  - DINX
  - Abinit
- **Commercial packages on EGEE Grid:**
  - Gaussian
  - Turbomole
  - Wien2k

Gaussian VO

- **Why Gausian?**
  - Large number of computational methods implemented
  - One of the first *ab initio* codes
  - The most popular among communities
  - User friendly
  - Available for many platforms along with GUI
- **Gaussian VO**
  - Invented and operated by ACC CYFRONET
  - All license issues confirmed with Gaussian Inc,
  - Open for *every* EGEE user
  - Any computing centre with site Gaussian license may support it (4 supporting centres, another 3 in the line)
  - 30+ users since the start in September 2006
  - VO manager – Mariusz Sterzel (`m.sterzel@cyfronet.pl`)
  - Enabled for parallel execution up to 8 processors

Turbomole – an alternative

Advantages:

- Probably the fastest B3LYP implementation
- Analytical gradients for excited states at DFT and CC2 levels
- Variety of fitting approaches speeding up calculations
- Very well scalability during parallel execution
- Extremely fast and very well parallelised (ri)CC2 and (ri)MP2

Disadvantages:

- Limited number of DFT functionals (only “good” ones available)
- Lack of parallel version of analytical second derivatives
- Lack of parallel version of TDDFT
- Only NMR chemical shifts implemented, no spin-spin couplings

Gaussian VO – participation

As a user:

- Register at:
  `https://voms.cyf-kr.edu.pl:8443/voms/gaussian`
- Wait for VOMRS admin acceptance
- `voms-proxy-init --vo gaussian` and you are ready to use the program...

As a participating centre:

- Just sent an e-mail concerning participation to VO manager
- After confirmation of the license status at your centre with Gaussian Inc, detailed information concerning set-up will be sent back to you

More details at:

`http://egee.grid.cyfronet.pl/Applications/gaussian-vo/`

Grid and parallel execution

Past:

- Serial jobs only
- Job of MPICH type always enforced execution of `mpirun`

Present

- `mpirun` no longer enforced
- Instead a wrapper script can be executed which will automatically set up environment for required MPI flavour
- No possibility to request desired # of processors on a WN

… Unfortunately not all sites are set up…

Chemical software

- “Old codes” – mostly written in FORTRAN
- Serial – parallel execution added later
  (with exceptions)
- Different parallelization models used
- Low scalability in many cases
- Only selected computational methods parallelized

… all that makes parallel grid ports of chemical software even more complicated

Selected cases

- **Gaussian**
  - Parallelization via OpenMP or Linda
  - OpenMP – SMP machines or multiprocessor/core clusters up to # of processors/cores on WN
  - Linda – allows the parallel execution between nodes. Requires equal # of processors for each WN
  - For Linda additional expenses required (commercial package), available only for specific platforms
- **Turbomole**
  - Uses MPI – currently HPMPI
  - No specific requirements
- **GAMESS**
  - Uses sockets but MPI execution possible (slower)
- **ADF**
  - Uses MPI (MPICH, OpenMPI, HPMPI, …)
  - One of the best parallelized QC codes // to my knowledge ;-)

Grid solution

- **Gaussian**
  - Parallel execution via OpenMP on a single WN up to # of processors/cores available on that worker node
  - Necessarily queue system set-up requires a Site admin help
  - Torque set-up:
    - Modification of `/var/spool/pbs/torque.cfg`

      to: SUBMITFILTER /var/spool/pbs/submit_filter.pl
  - Other settings -- typical
    - Job has to be of MPICH type
    - # of processors controlled via `NodeNumber` variable
    - Gaussian `%Nproc` route is automatically set-up by script executing Gaussian
  - Execution with 8 processors per job possible.

Sample script

```
Executable    = “/bin/sh”;
Arguments     = “$GAUSSIAN_SW_DIR/gaussian.run myfile.com”;
JobType       = “MPICH”;
NodeNumber    = 8;
InputSandbox  = {“myfile.com”};
StdOut        = “myfile.out”;
StdErr        = “myfile.err”;
OutputSandbox = {“myfile.log”, “myfile.chk”,
                 “myfile.out”, “myfile.err”};
Requirements  = other.GlueCEUniqueID==“ce.cyf-kr.edu.pl”
```

Grid Solution cont.

- **Turbomole**
  - No special set up except shared directory needed, # of processors automatically discovered by Turbomole scripts
- **NAMD**
  - Similar to Turbomole. If the NAMD executing script was set-up properly during installation the necessarily “node file” is created every time program is executed
- **GAMESS**
  - Depends on Grid port
  - In case of MPI no additional input needed
  - DDI case – may require WN reconfiguration especially if large DDI memory is requested by a job

Execution benchmarks

Scheduling time

- MPI jobs
  - 4 proc./job -- usually less than hour
  - 8 proc./job -- waiting time even 3-4 hour
- OpenMP jobs
  - Job waiting time much longer, heavily depends on site overload
    - *4 proc./job -- from less than hour up to 6 hours*
    - *8 proc./job -- in some cases job waiting time exceeds 12 hours*

Parallel job execution can be inefficient in case of short (less than 24 h) jobs

Grid application in chemistry

- **Tasks to which Grid can be applied directly:**
  - Conformational analysis
  - Numerical frequency computations
  - Zero Point Vibrational Averaging
  - Property computations for series of geometries from Molecular Dynamics simulation
  - Determination of chemical reaction paths
  - Determination of potential energy surfaces (PES)

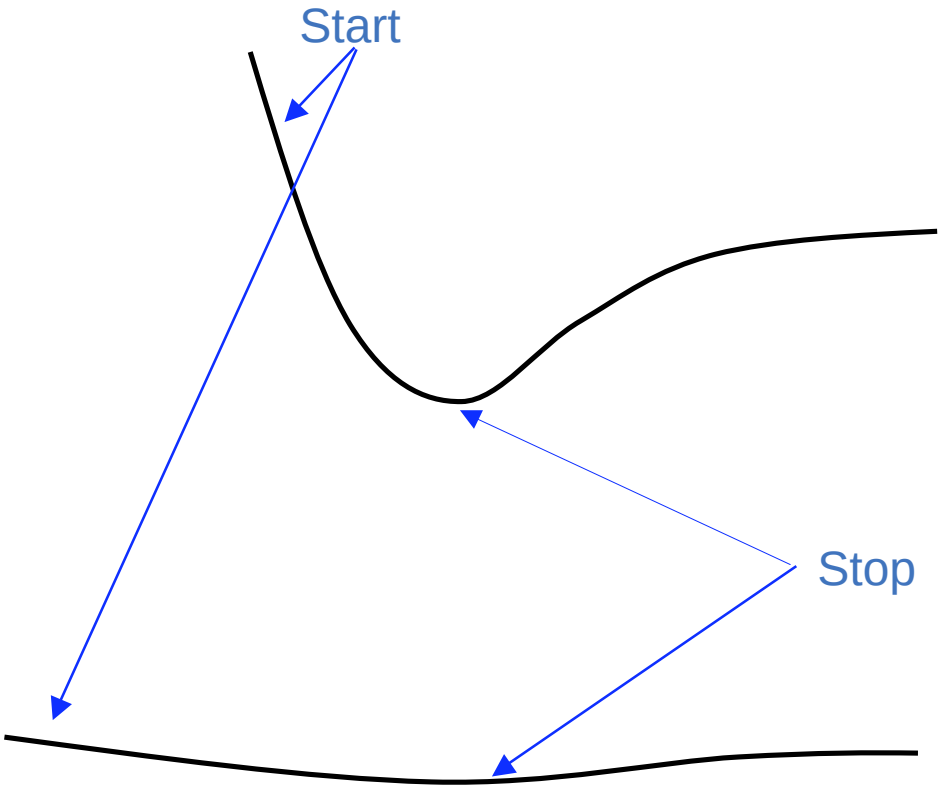
  … all kind of “brute force” tasks, or tasks which operate on huge data sets

- **Other tasks**
  - Computations need to to be planned in order to maximize benefits from the grid computing

An example

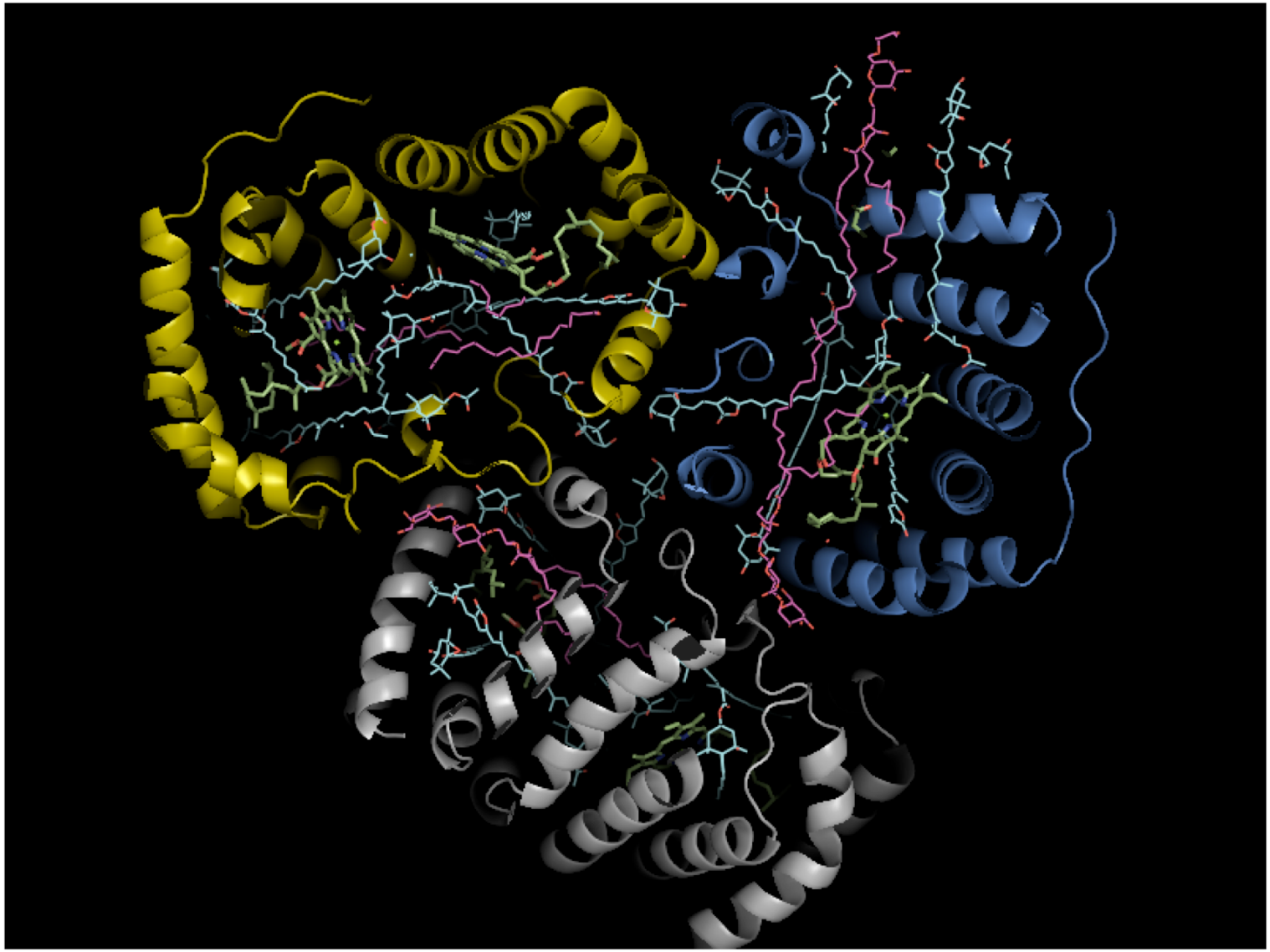
Geometry optimization:

- Steep potential
  - Few steps needed

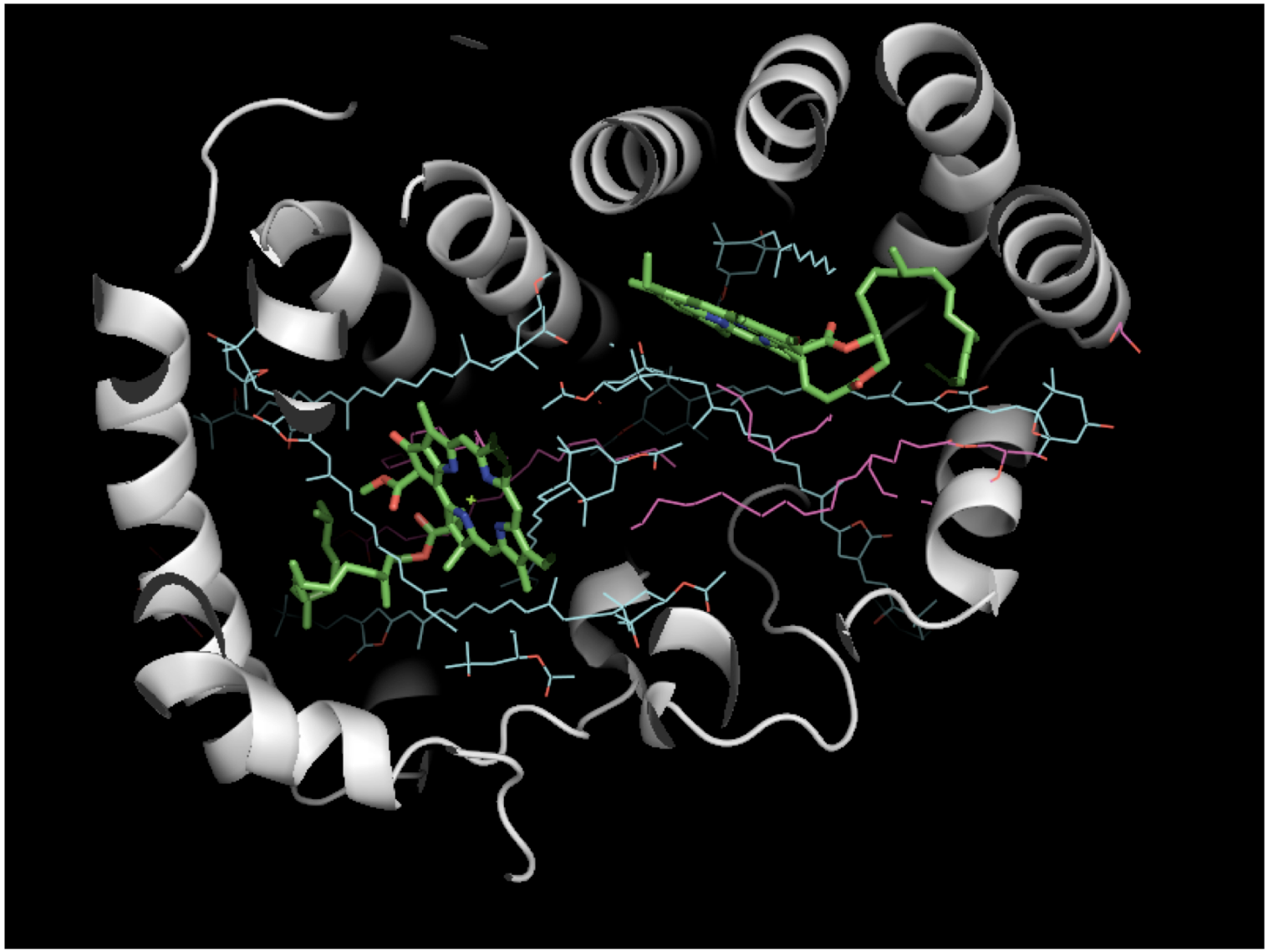


- Flat potential
  - Many steps needed
  - Energy and gradient convergence have to be increased to high values

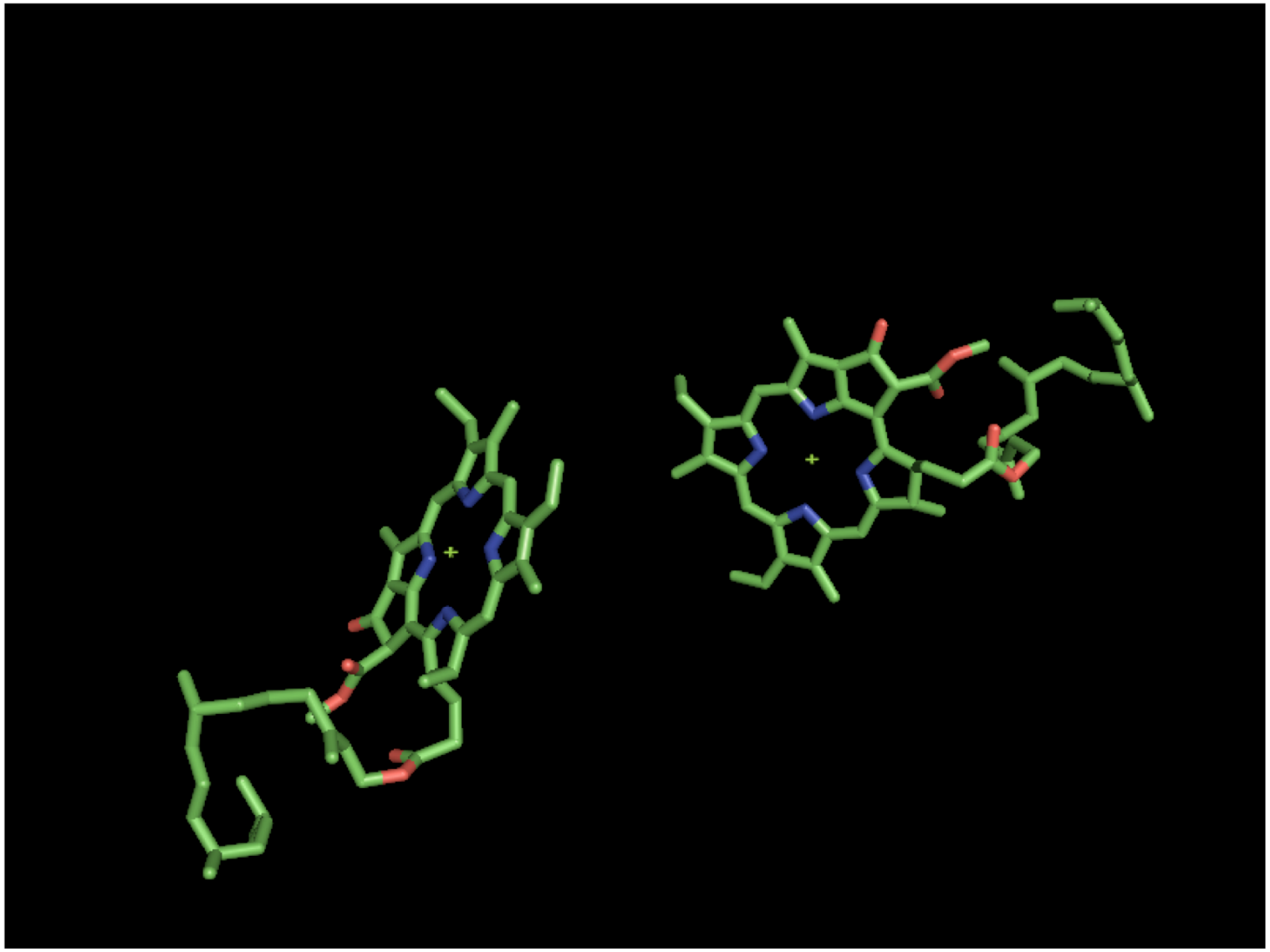
PCP complex



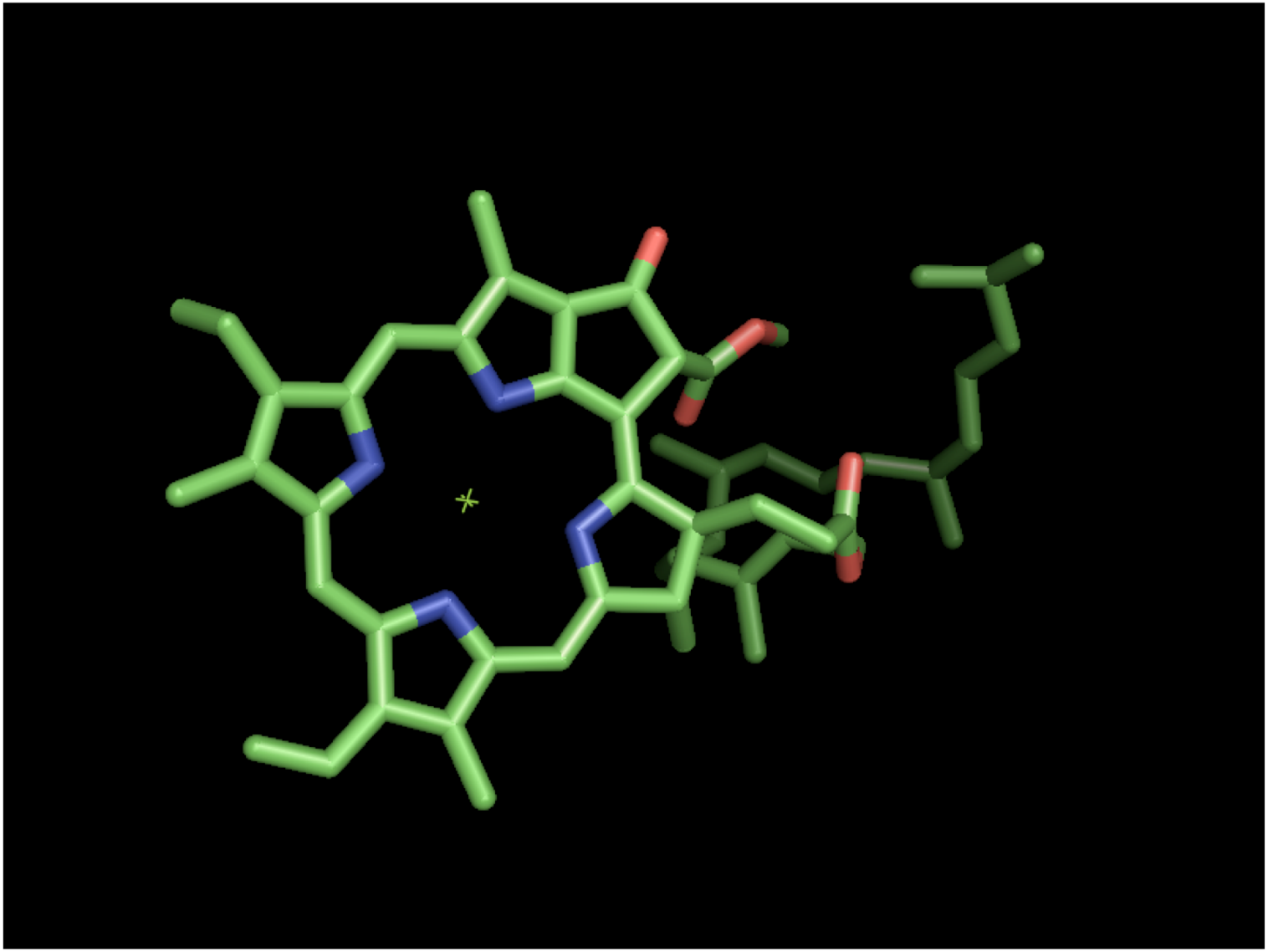
PCP cont.



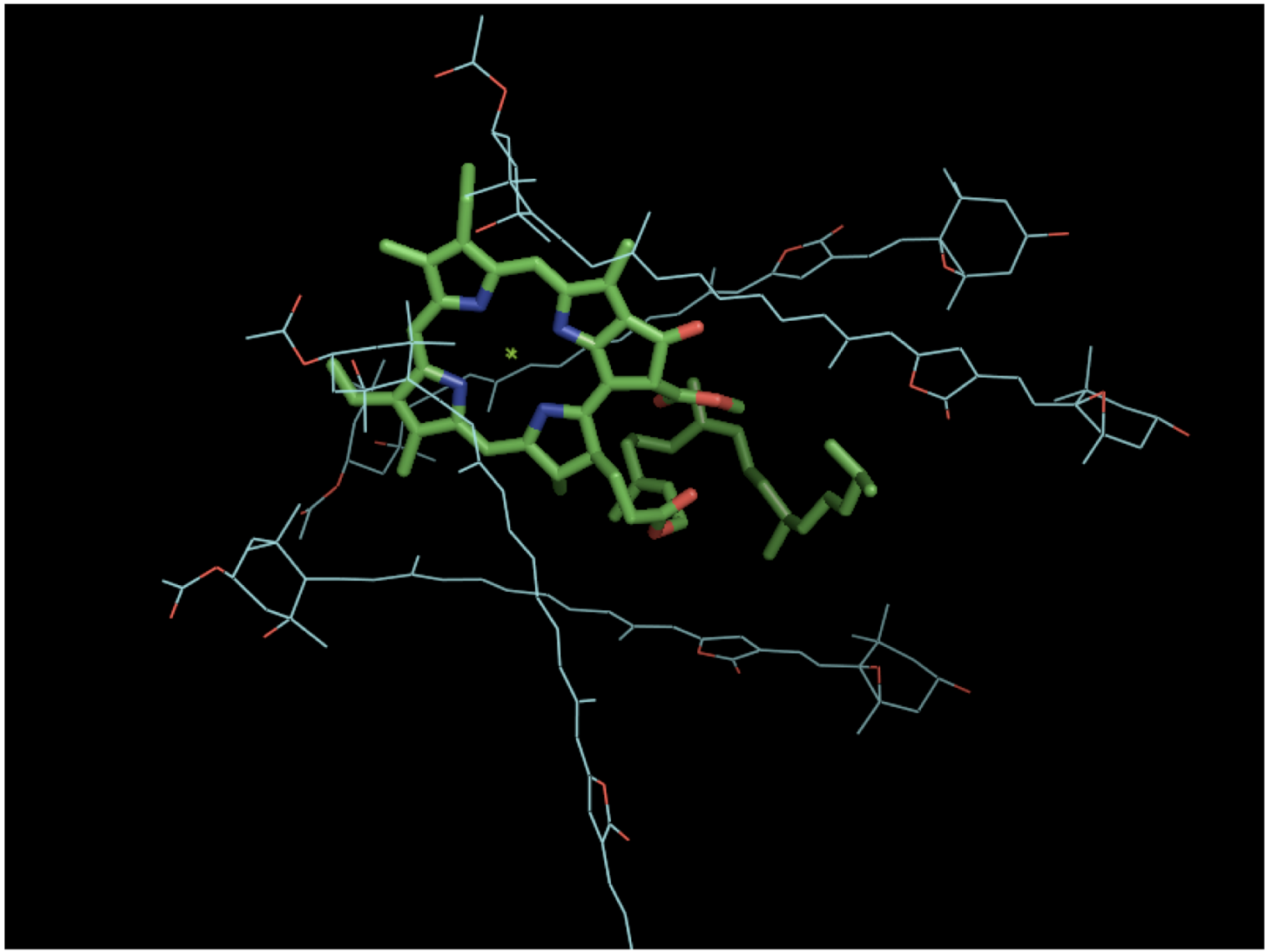
PCP cont.



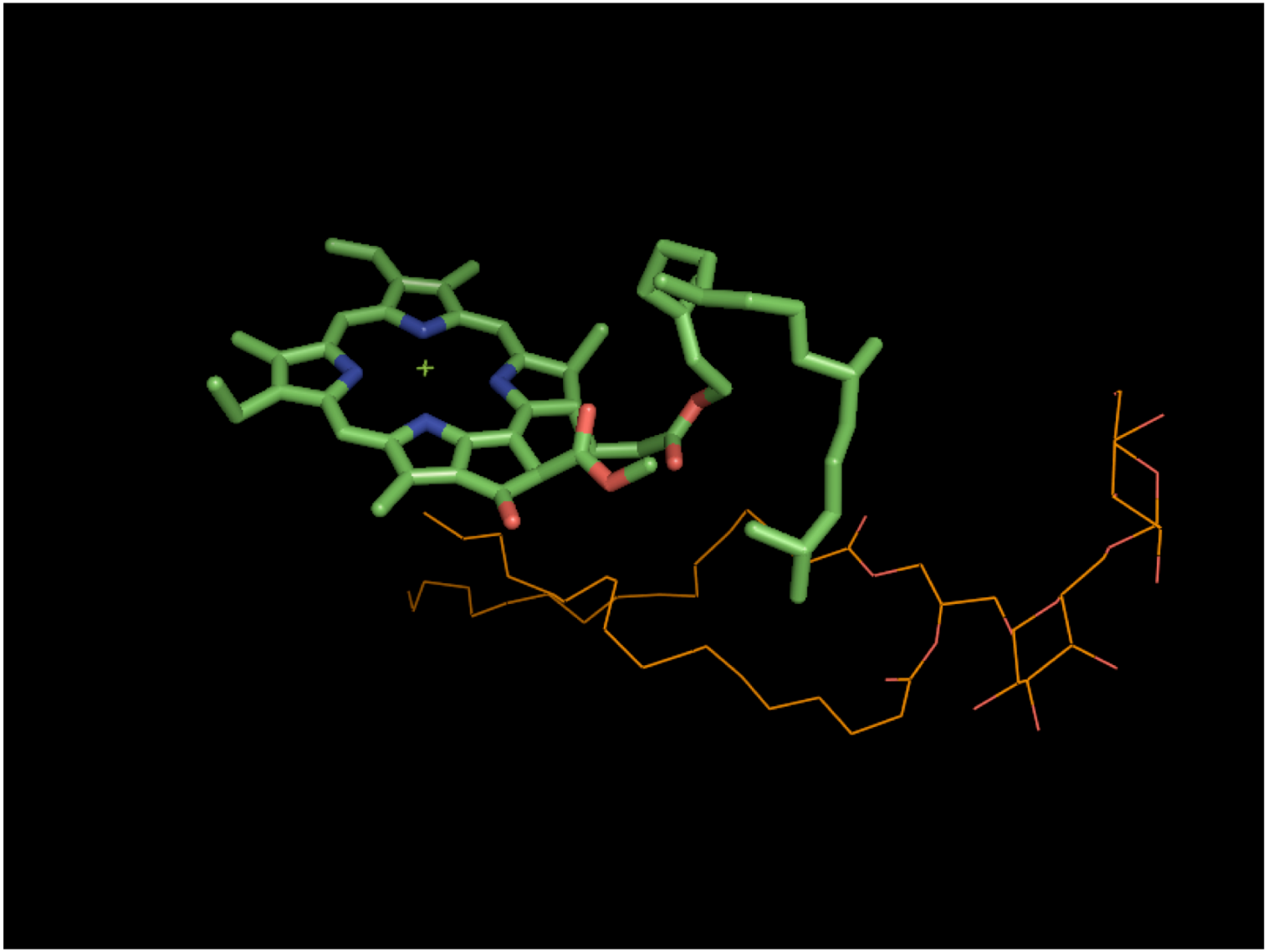
PCP cont.



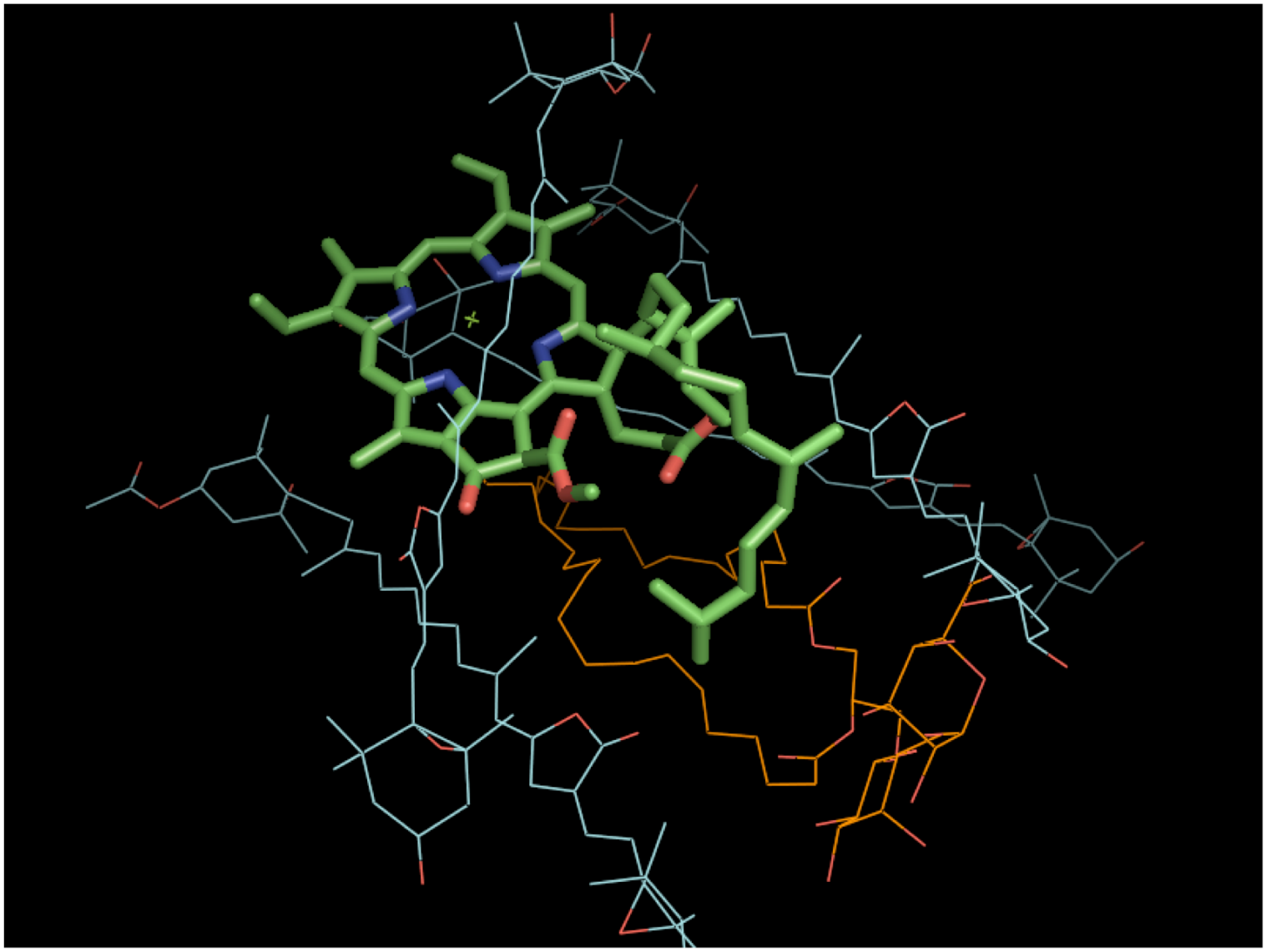
PCP cont.



PCP cont.



PCP cont.



Plan

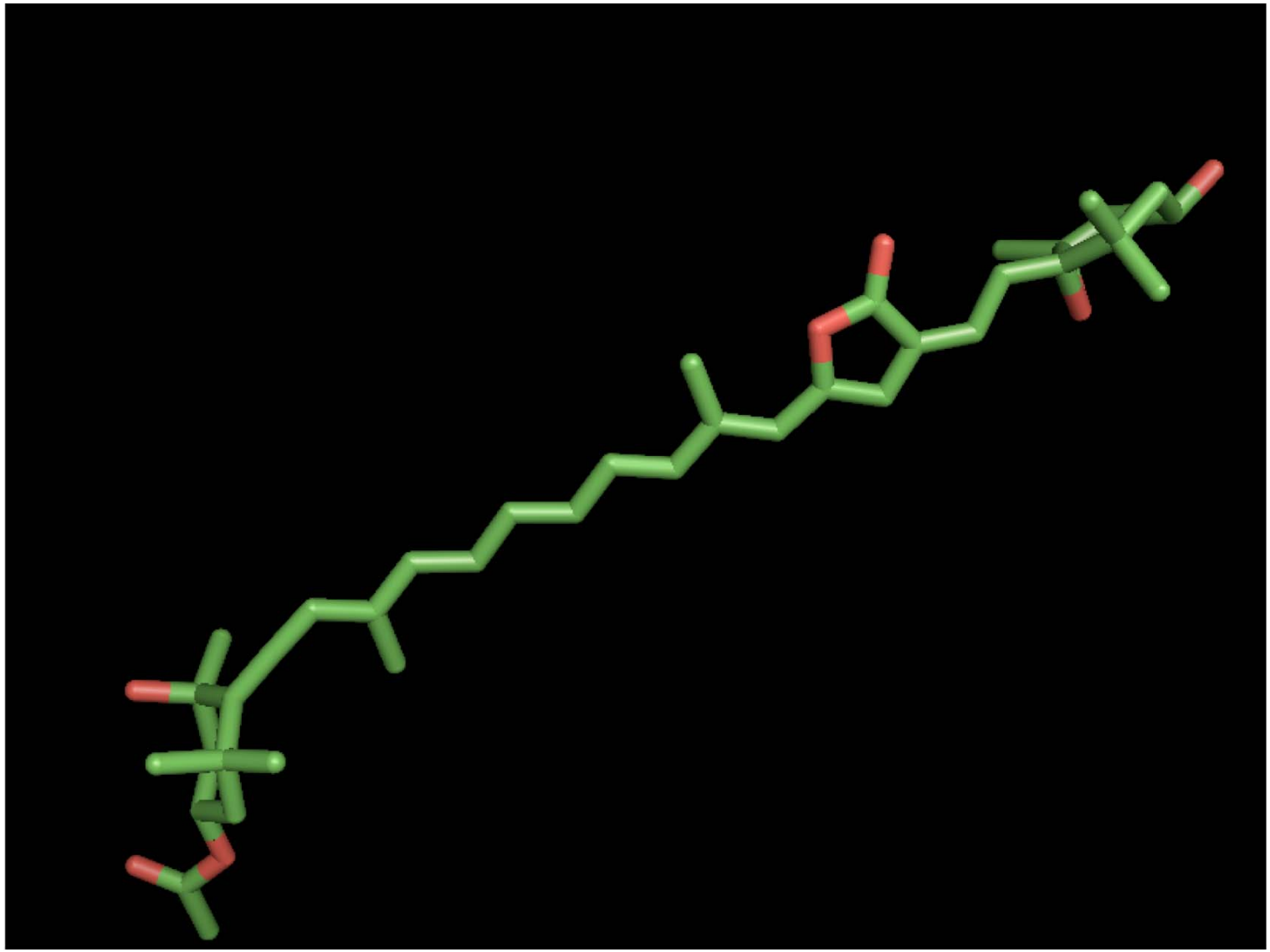
- **Computations of the whole molecule are not possible**
- **System needs to be modeled**
  - For this we need:
    - Chlorophyll A and peridinin ground and excited states geometries and normal modes
    - The geometry of whole complex
    - A little programming (tetramer model, energy transfer model)

Plan

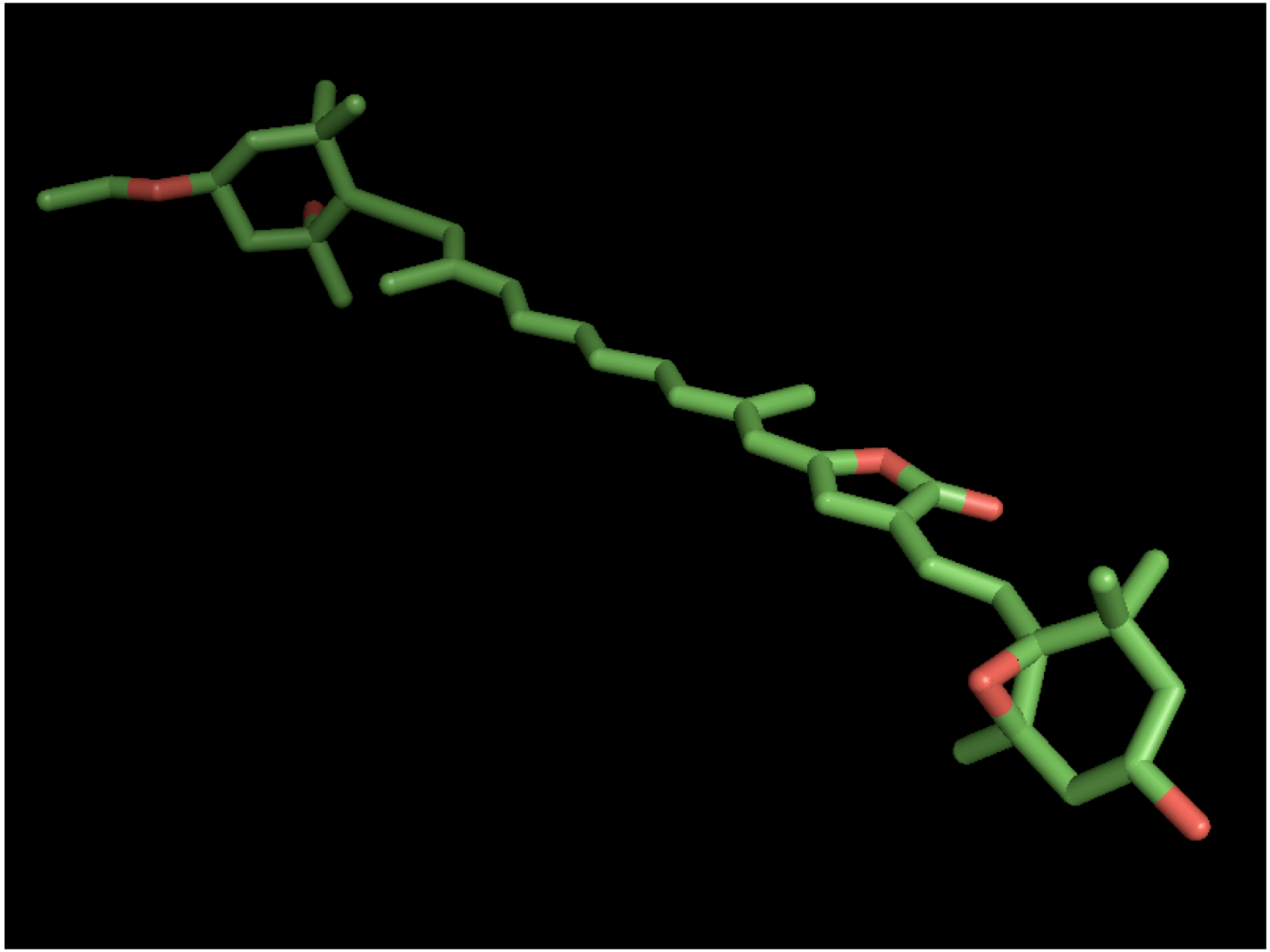
- **Computations of the whole molecule are not possible**
- **System needs to be modeled**
  - For this we need:
    - Chlorophyll A and peridinin ground and excited states geometries and normal modes
    - The geometry of whole complex
    - A little programming (tetramer model, energy transfer model)

… a lot of luck

First step – Peridinin



First step – Peridinin



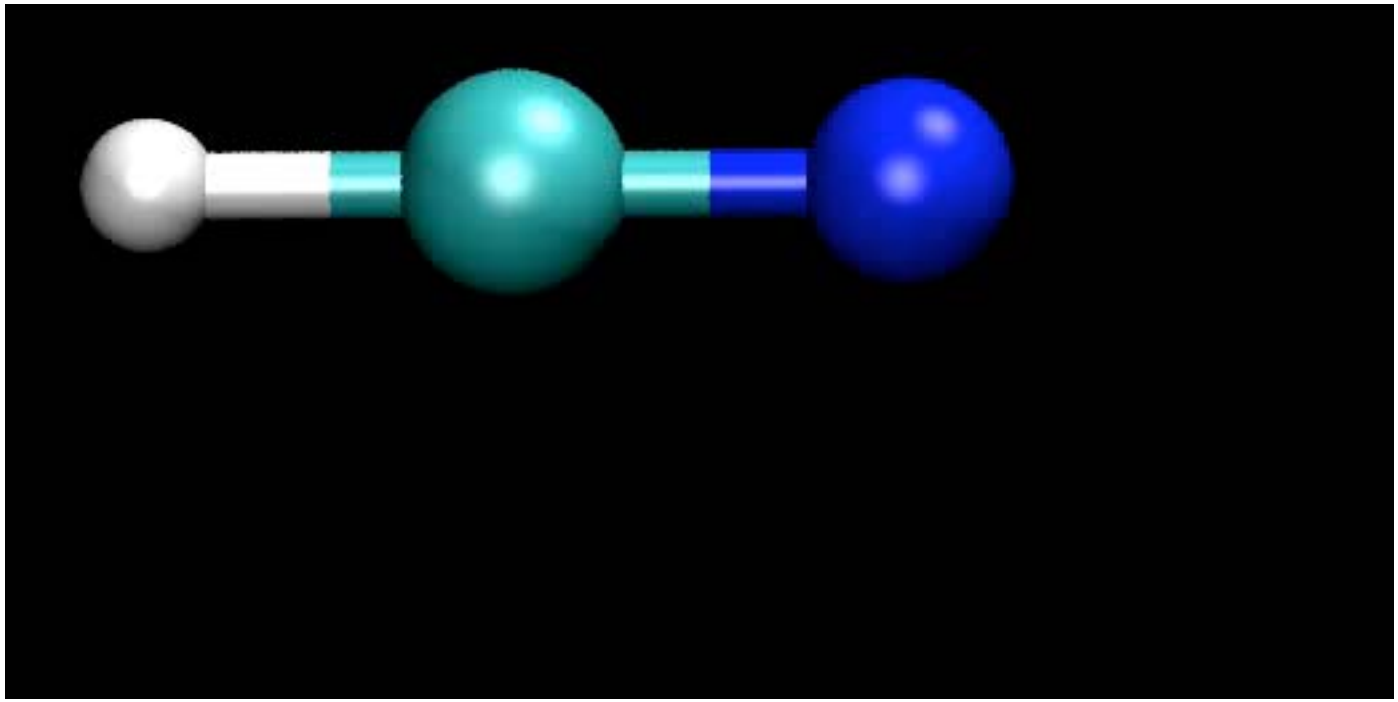
Peridinin geometry

- **A long peridinin chain makes usual gradient based optimization inefficient**
- **Instead we propose following scheme:**
  - MD for peridinin (force field level)
  - Geometry preoptimization for series of MD snapshots (semi empirical or ~ RHF/STO-3G level)
  - “Final” geometry optimization for few lowest energy MD snapshots (CASSCF/PT2 level)
  - Verification of the minima by frequency calculations
  - Excited state geometry optimization with ground state as a starting point
  - Again, minima verification via vibrational analysis
  - Verification of obtained data by comparison with experiment

Chemical reactions

... a process that results interconversion of molecules

H—CN ⟶ CN—H



Chemical reactions

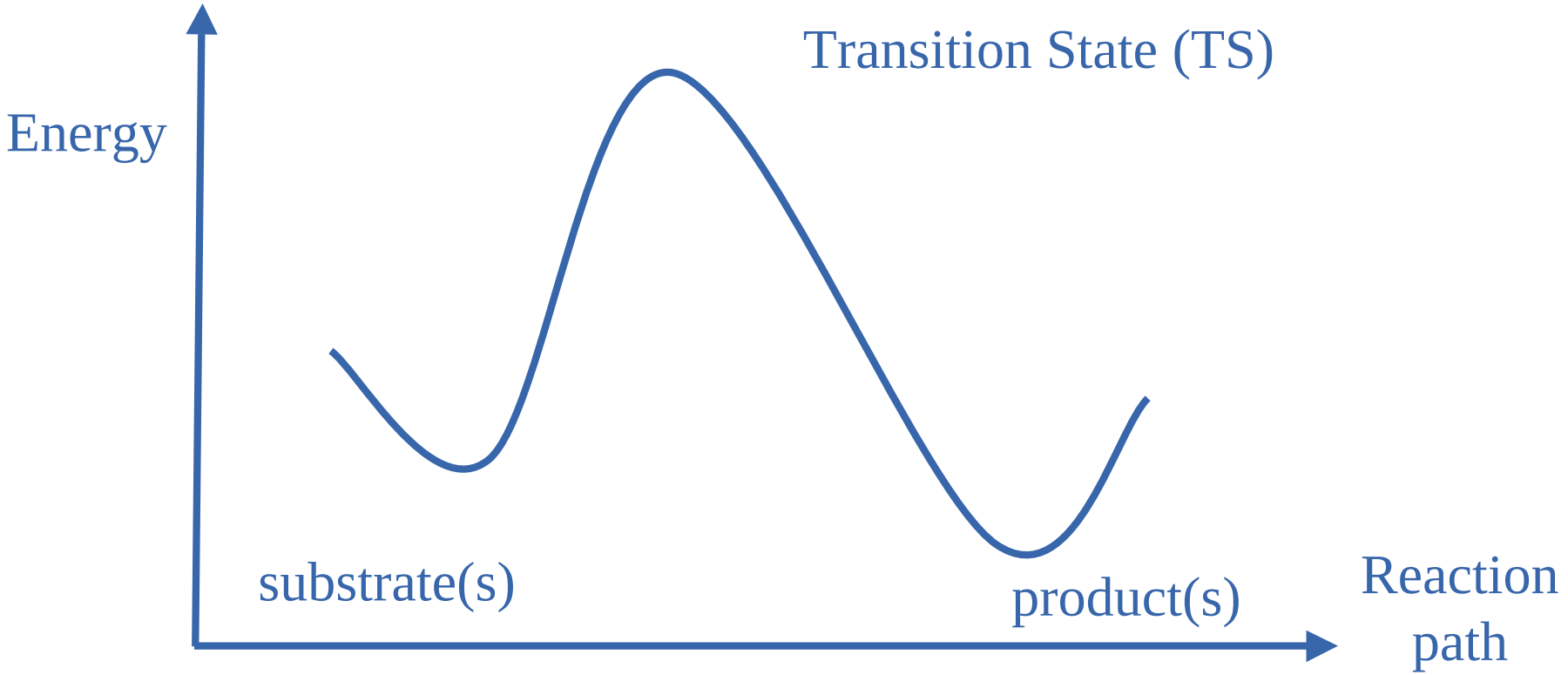
Points of interest:

- Structure of substrates and products
- Structure of active complex at TS
- Reaction path

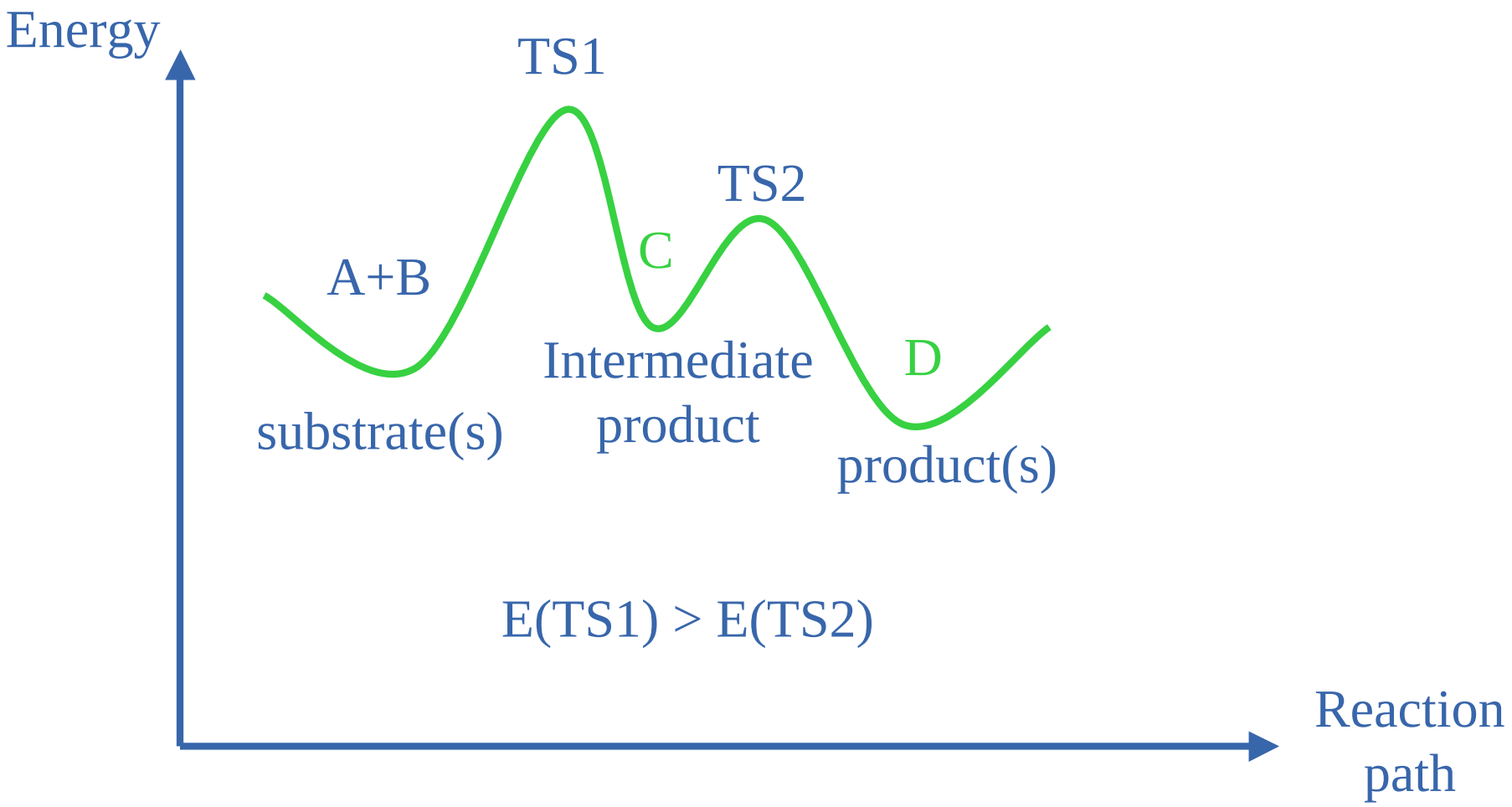
Chemical reactions

Points of interest:

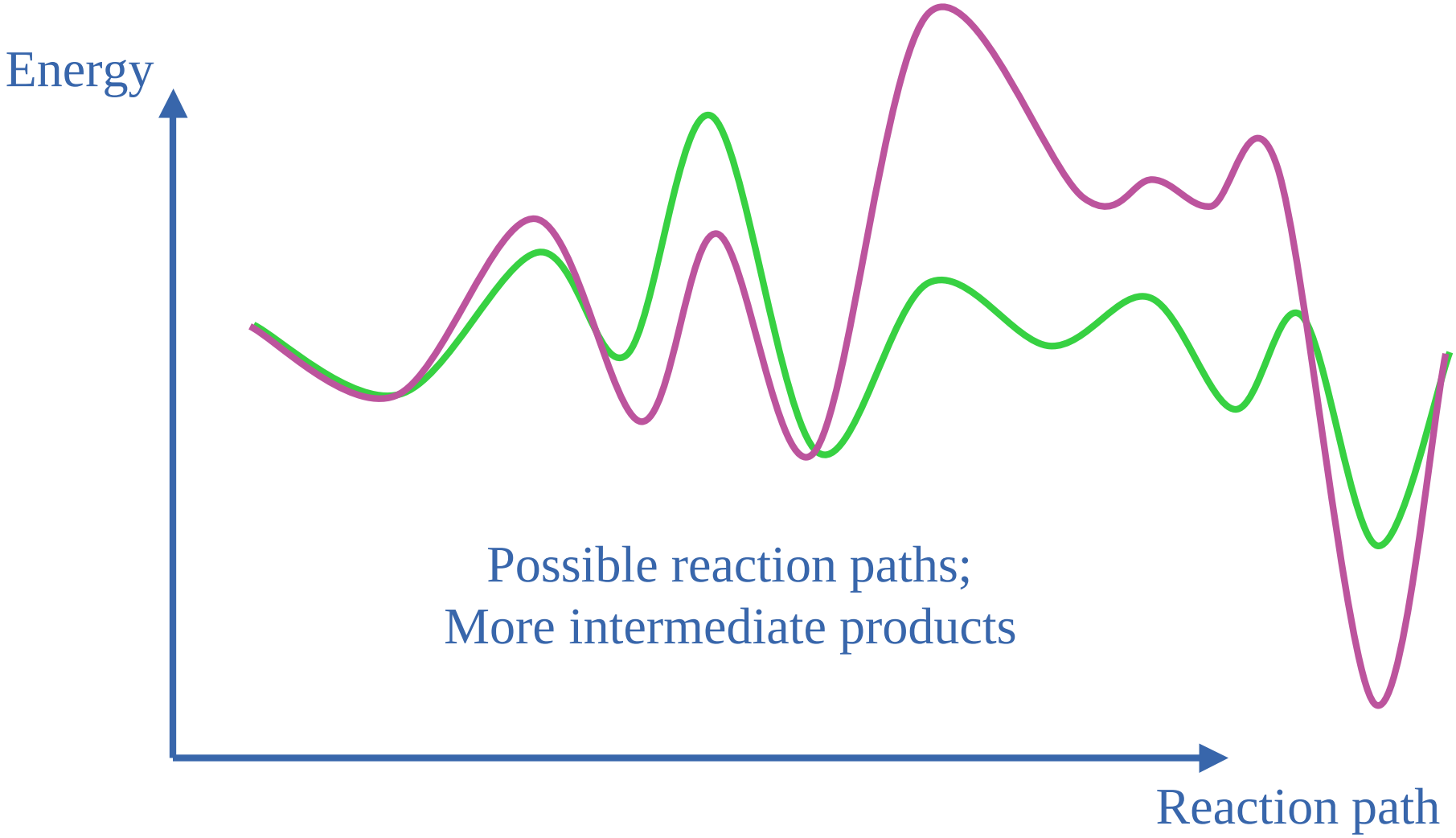
- Structure of substrates and products
- Structure of active complex at TS
- Reaction path(s)



Chemical reactions



Chemical reactions



Chemical reactions

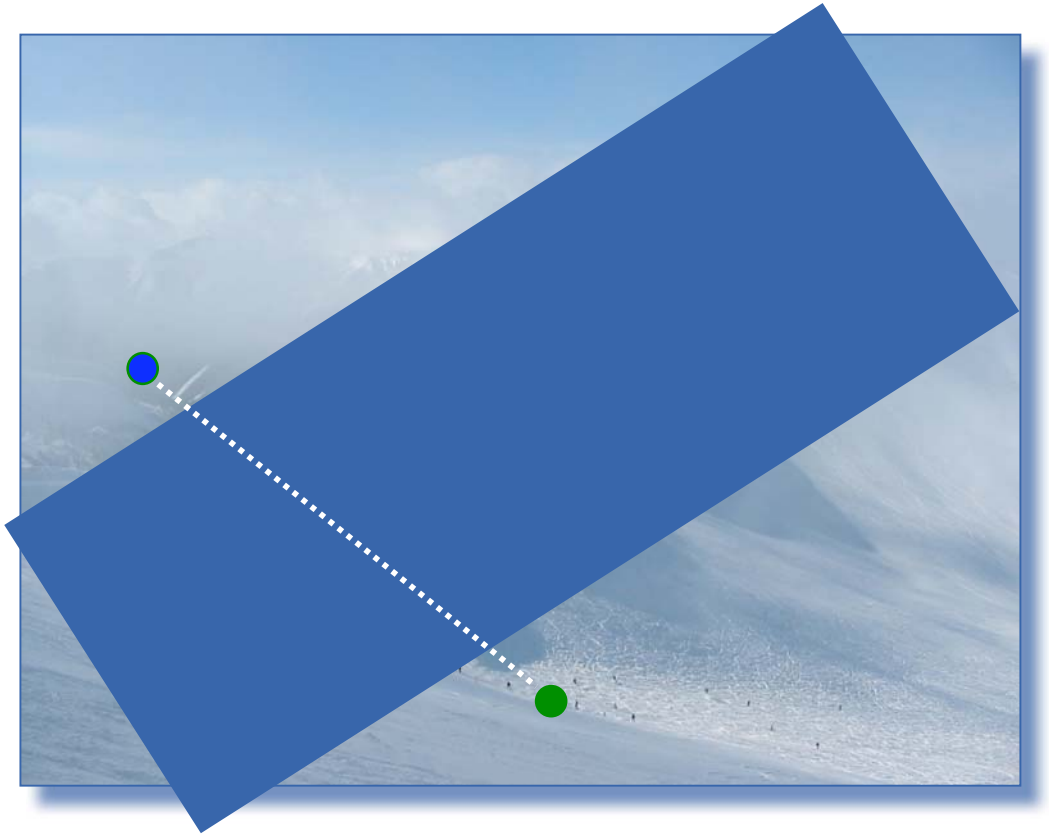
Main problem – TS determination

Chemical reactions

Main problem – TS determination

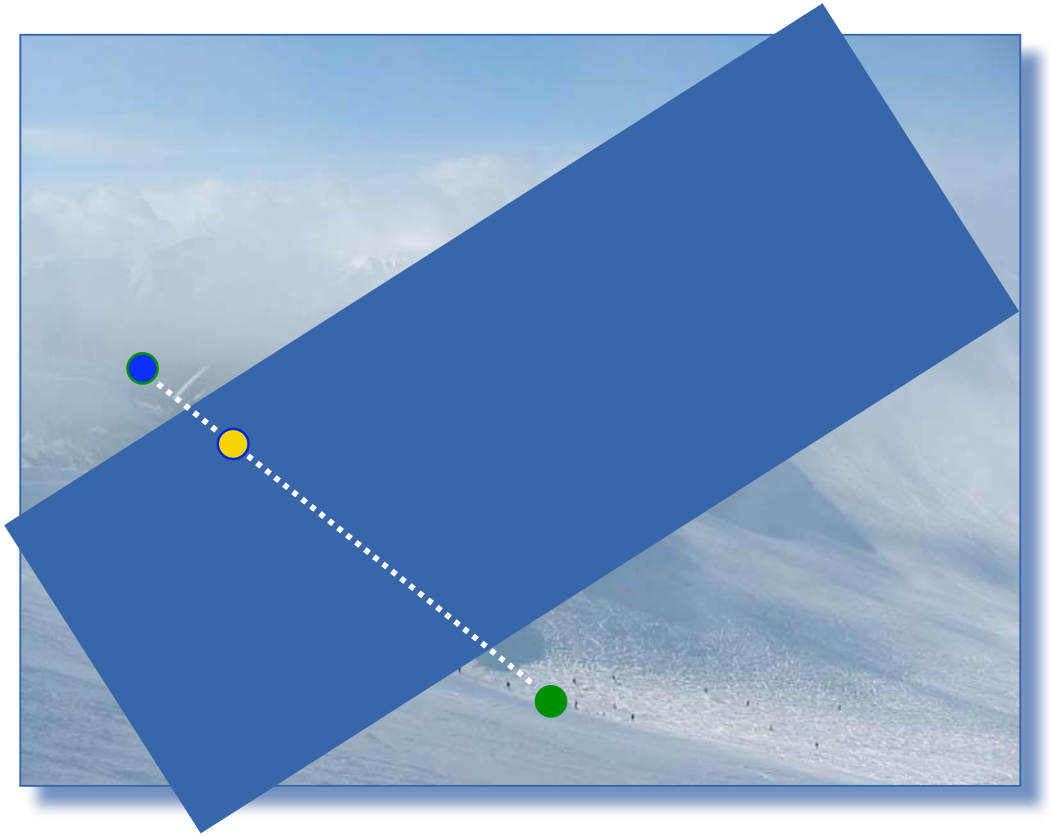
Chemical reactions

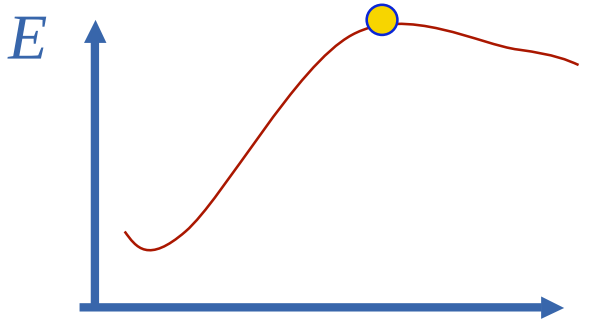
Main problem – TS determination



Chemical reactions

Main problem – TS determination

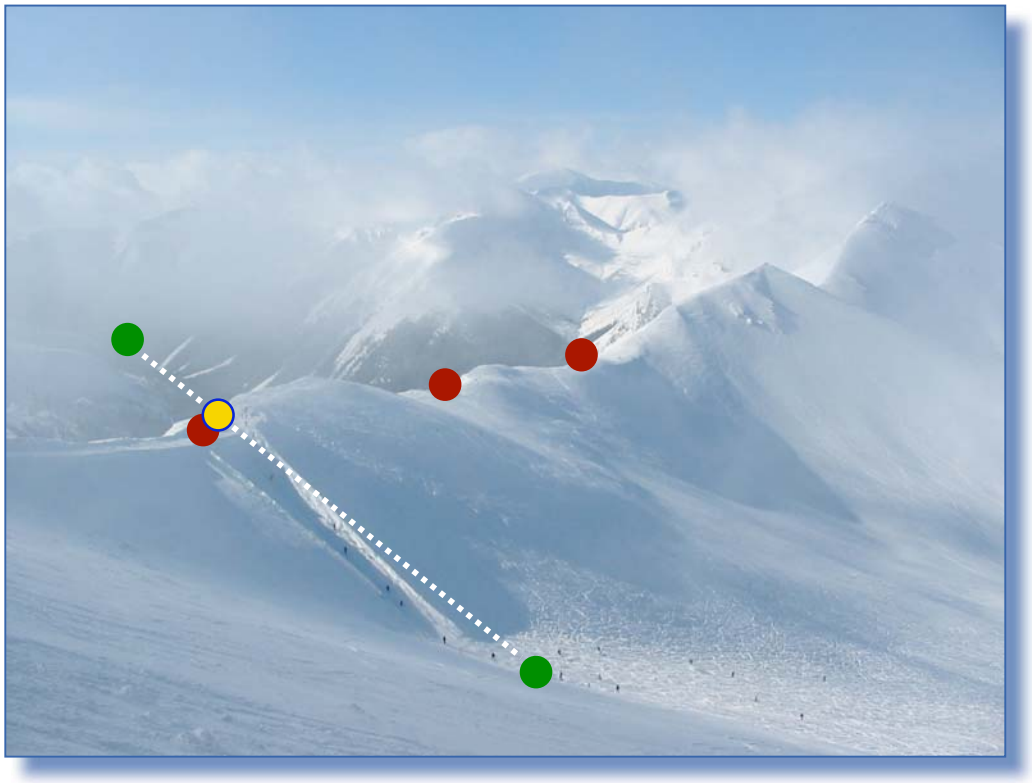


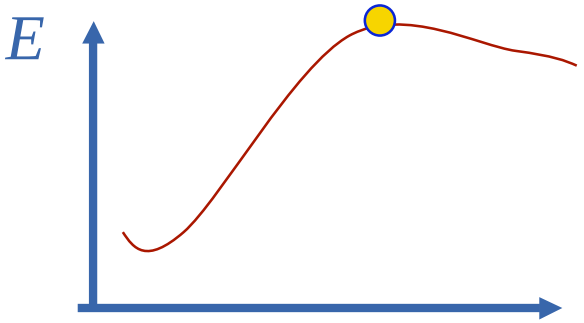


‘reaction coordinate’

Chemical reactions

Main problem – TS determination

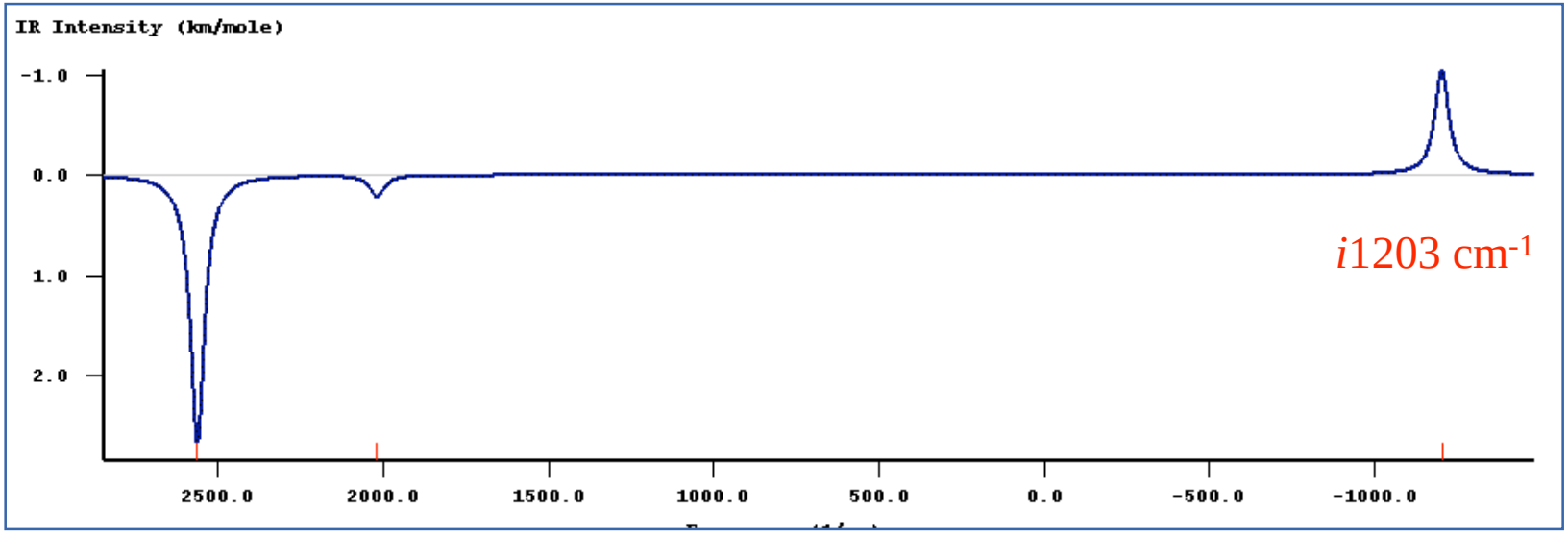


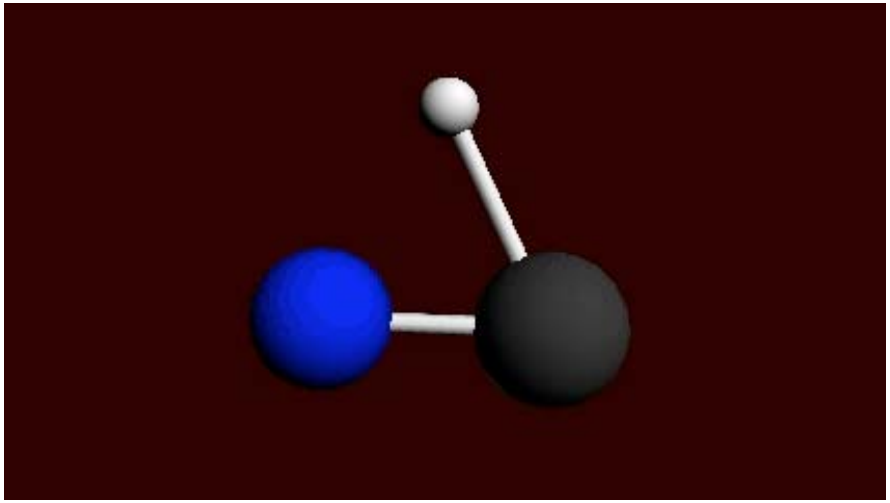


‘reaction coordinate’

Chemical reactions

TS verification – vibrational analysis – one imaginary frequency!

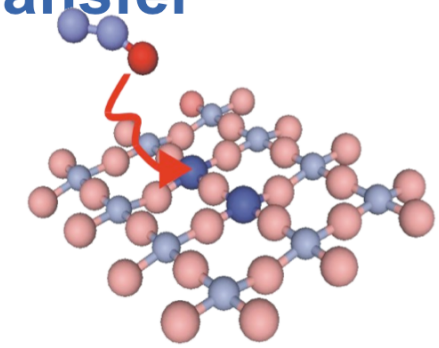




Sample calculations

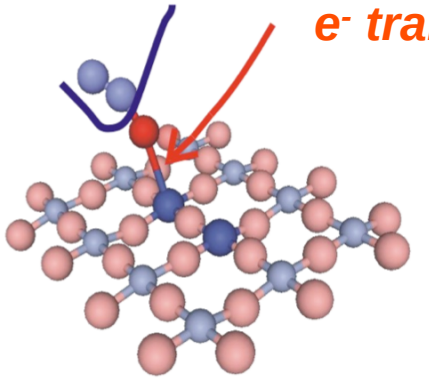
N_2O braking on oxide surfaces – possible mechanisms

- **Electron transfer**



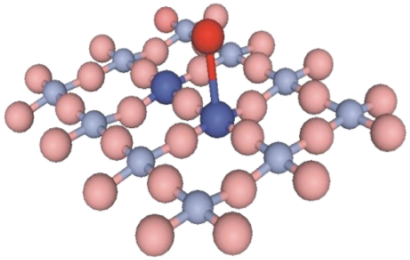
1. $N_2O + X$

e^- transfer

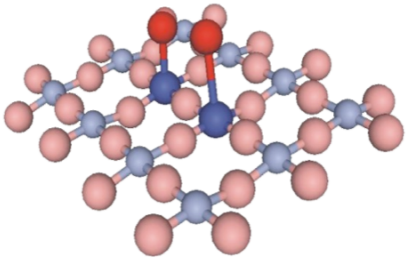


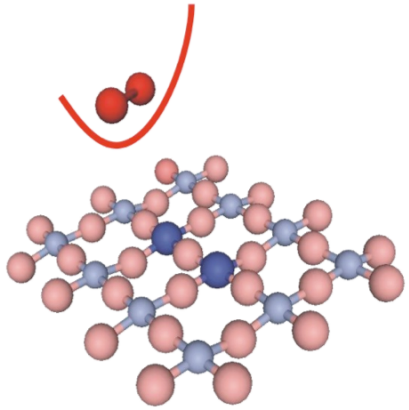
2. $N_2O_{(s)} + X \rightarrow X^+ \text{--} N_2O^-$

4. $X^+ \text{--} N_2O^- \rightarrow X^+ \text{--} (O^-) + N_2$



5. $(O^-) \text{--} X^+ \text{--} (O^-)$



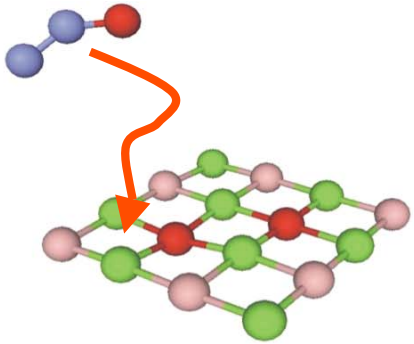


6. $(O^-) \text{--} X^+ \text{--} (O^-)\ X + O_2$

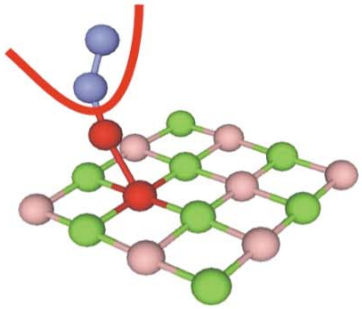
Sample calculations

N_2O braking on oxide surfaces – possible mechanisms

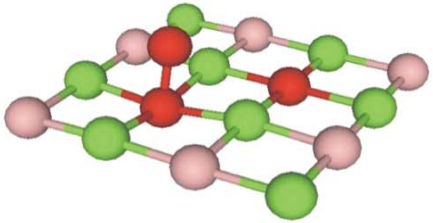
- An Oxygen transfer



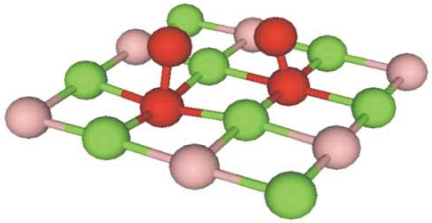
1. $N_2O_{(g)} \rightarrow N_2O_{(s)}$



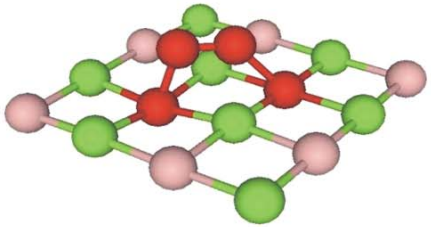
2. $N_2O_{(s)} + X \rightarrow X\text{--}O\text{--}N_2$



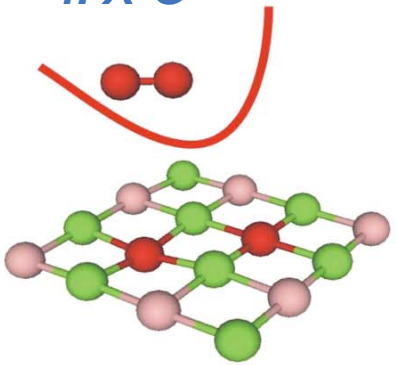
4. X-O



4. O-X-O

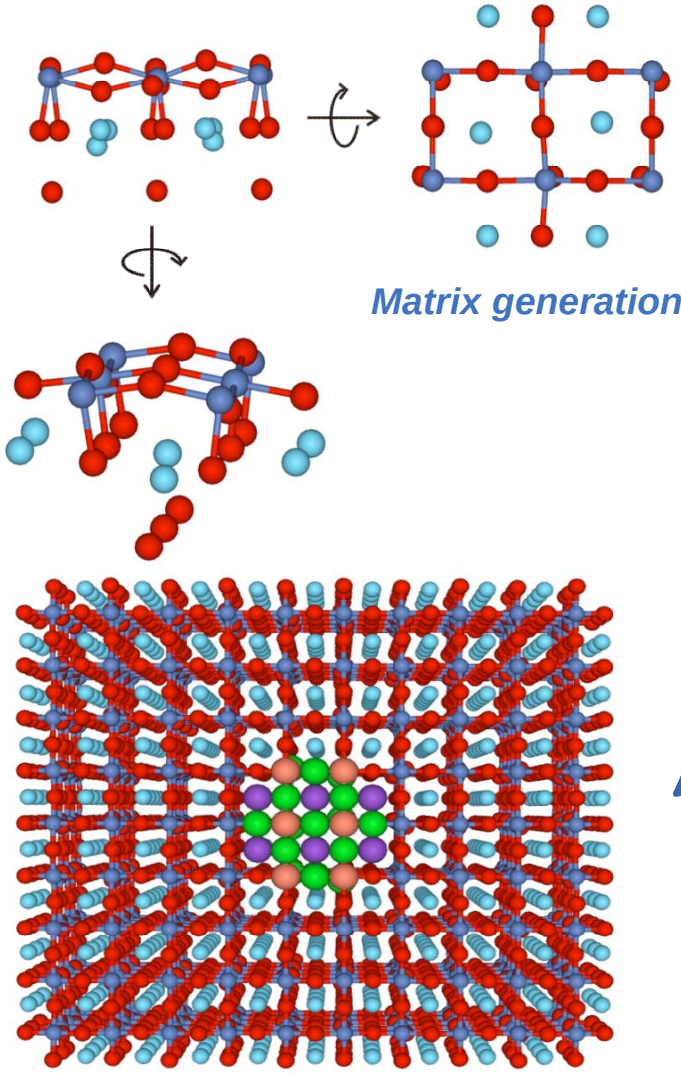


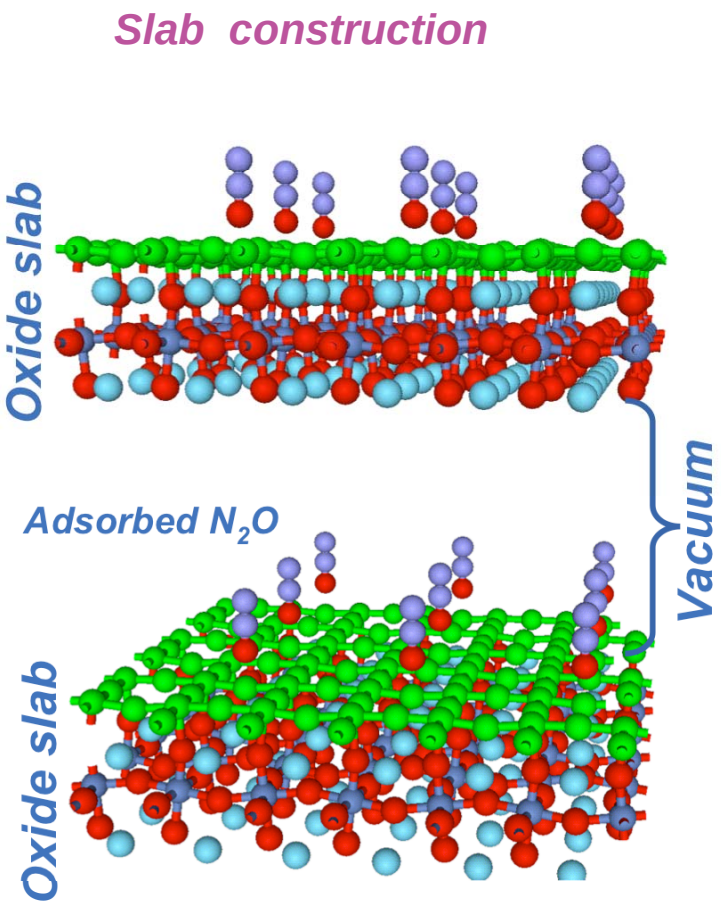
5. O--O-X



6. $X + O_2$

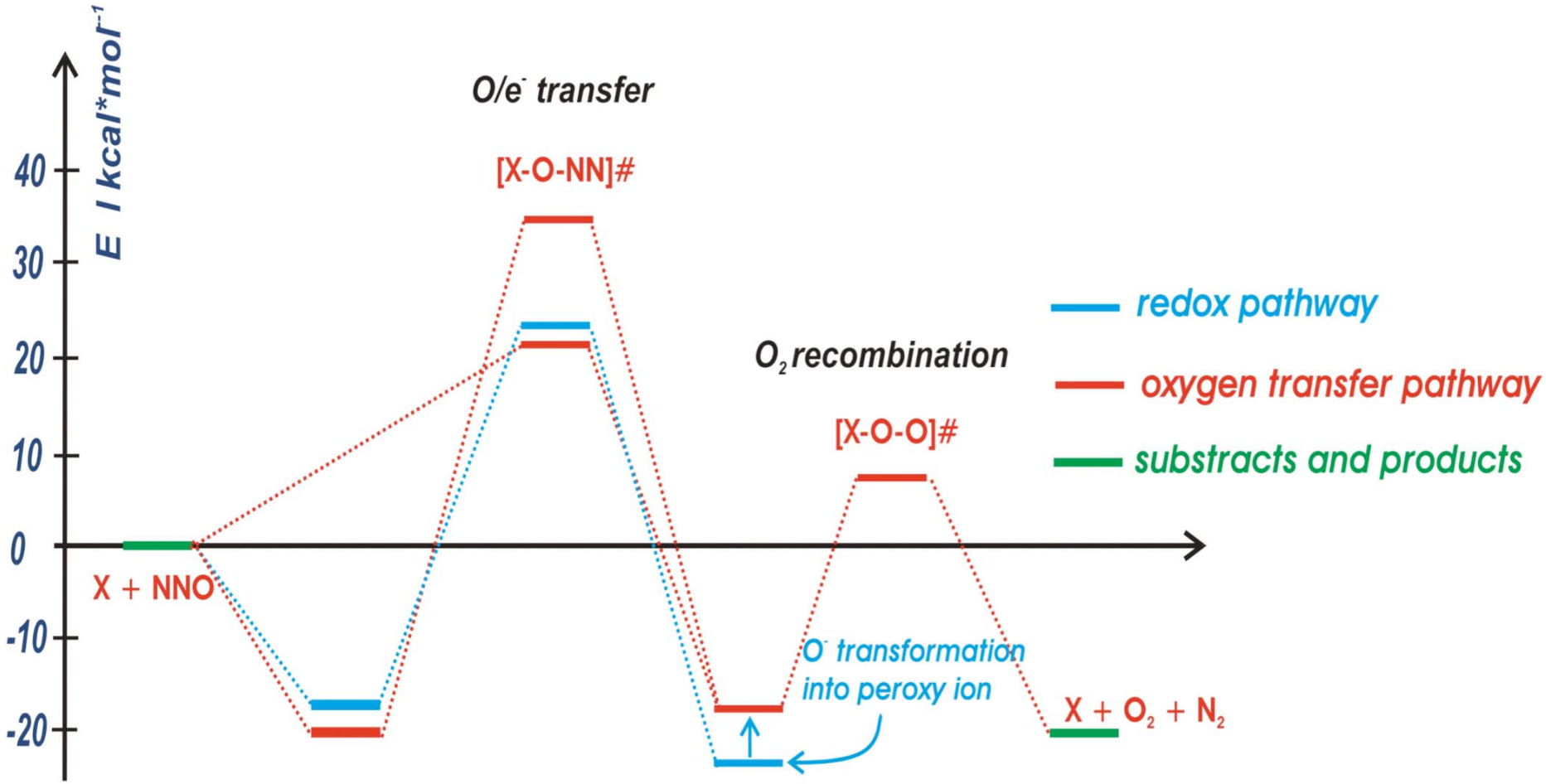
Sample calculations





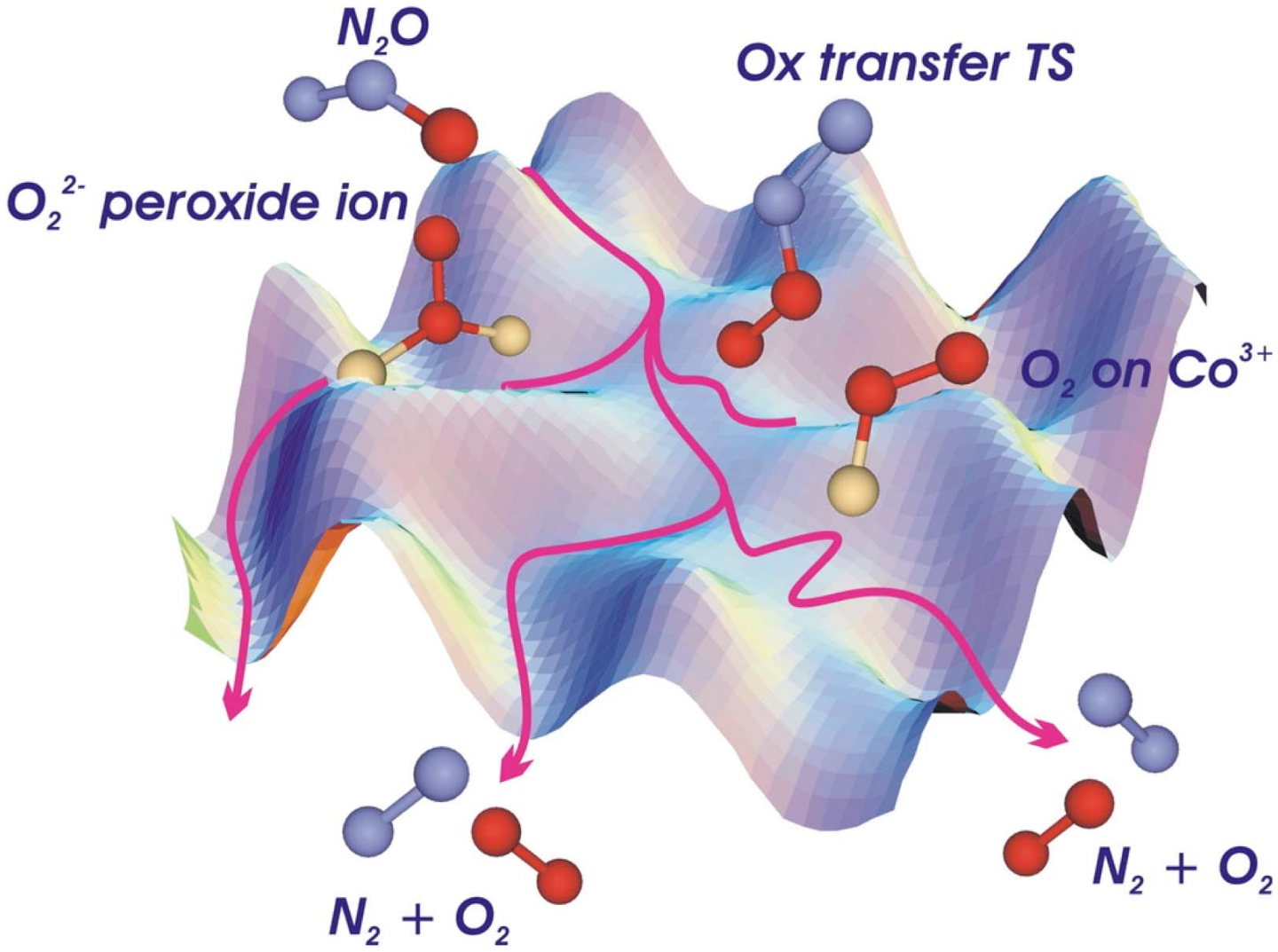
Sample calculations

- **Reaction paths:**



Sample calculations

- **Reaction paths:**



Sample calculations

Computational details:

- Gaussian 03 D.01
- BP86 functional
- Basis set of double-ζ quality

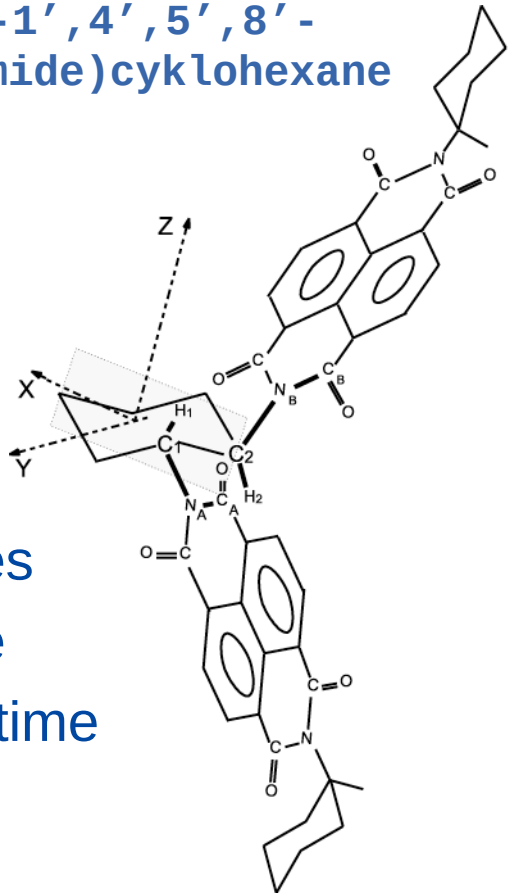
Timings:

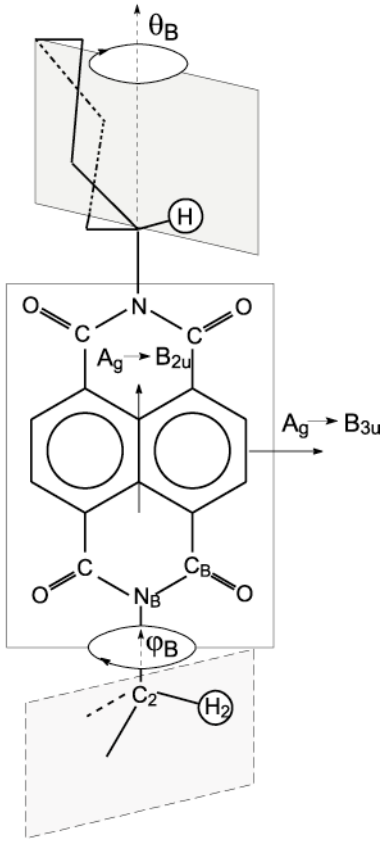
- CPU time for SP calculation – approx 7 hours
- 15 paths – 10-50 energy points on each path
- In total about 250 energy points calculated

Conformational searchers

- **To determine lowest energy structure**

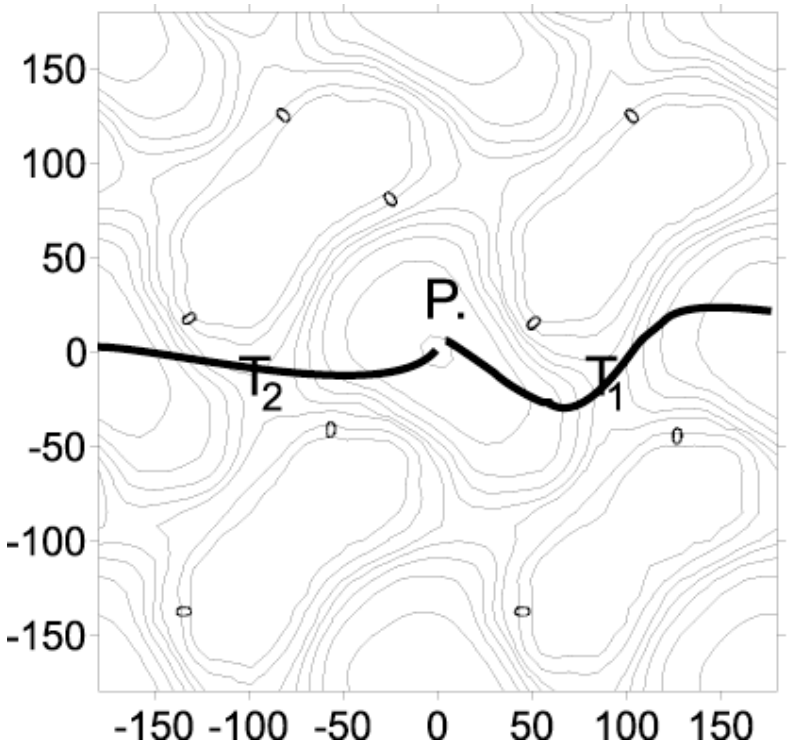
(1R,2R)-1,2-bis-(N'-cyklohexy-1',4',5',8'-naphthalenetetracarboxydiimide)cyklohexane

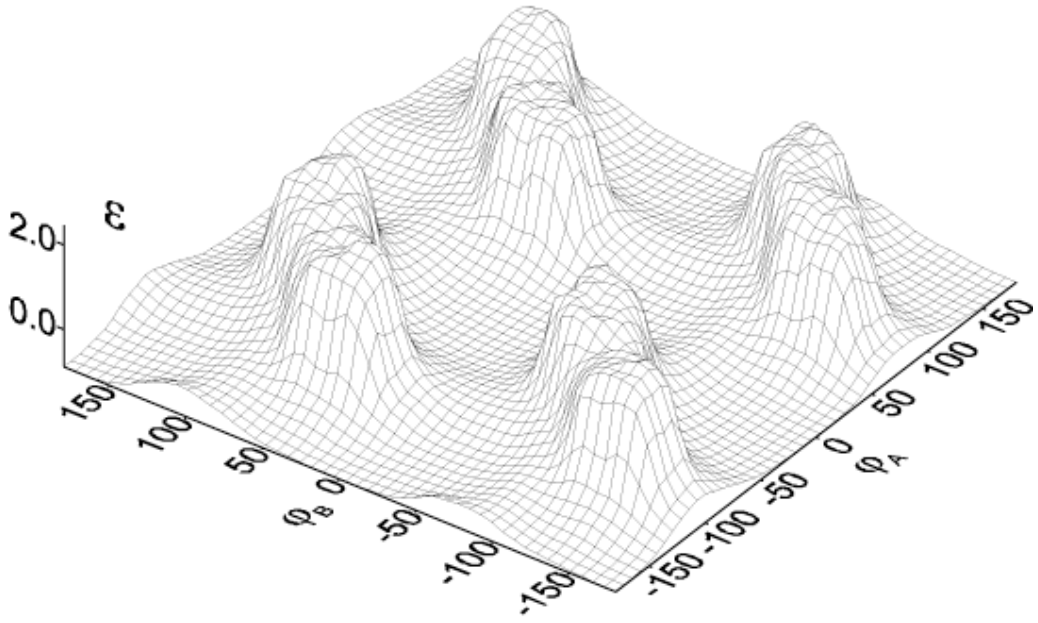




- **Possible issues**
  - For short time jobs it is convenient to group several geometries in to one job to minimize average job scheduling time

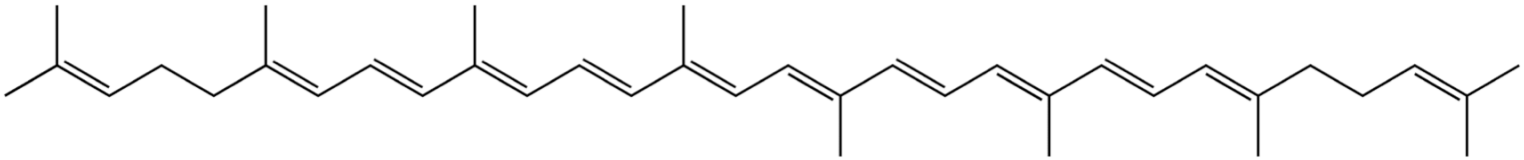
Conformational searches





Harmonic frequencies

Numerical frequency computations for lycopene



- 96 atoms, 2·3·96+1=577 independent computation steps
- Software: GAMESS version June 2005
- VOCE VO resources used
- Methodology: B3LYP, cc-pVDZ basis set
- Computations done on approx. 100 processors (ia64 and i386)
- CPU time for single computation:
  - Intel Xeon 2.8Gh – 14h 39’
  - Intel Itanium 2 1.3Gh – 12h 8’
- Total time:
  - single CPU – 330 days (estimated)
  - EGEE Grid – 3 days

Summary

- Access to the Grid is easy, does not differ to much from queue system usage
- EGEE Grid offers variety of software packages for chemical computations. A parallel execution can be made as simple for the user as a serial execution is now
- It is always possible to find solution for parallel execution even if computational platform does not directly support certain parallelization model
- With a little of planning every computational chemistry task may benefit from the Grid platform