Assignment 10: Experiments in Molecular Geometry Optimization: Biphenyl Minimization

See Insight II *and* Discover *manuals for reference.*

1. **Brief introduction to the Discover module of** Insight II.

   The Discover[7] software performs energy minimization and molecular dynamics simulations. This program constitutes a powerful modeling tool since it offers many features such as constrained and restrained minimization, calculation of vibrational frequencies, and analysis tools. Many variations of simulation conditions (e.g., constant temperature, constant pressure) are available.

   We will access the Discover software from the Insight II environment. The **Discover** module of Insight II is a convenient interface to the Discover program. This module builds Discover input files from information provided through graphical interfaces, and it allows users to run Discover jobs interactively. Though more advanced users may prefer to use the independent version of Discover, the Insight II environment is more appropriate for a novice.

   Before using Discover, make sure that Insight II contains all of the necessary information to define the topology, coordinates, and force field parameters. These include, for example, atom types and partial charges (see lecture notes for structure definitions).

   If you succeed in displaying the molecule correctly on the screen, the topological and coordinate information is most likely in order. However, selecting the appropriate force field and assigning atom types and parameters is a separate task.

   (a) To select the force field, use Forcefield / Select.
   (b) To assign atom types, use the Fix option for Potential Action in Forcefield / Potentials. Alternatively, first assign atom types with Atom / Potential in the **Biopolymer** module, and then use the Accept option for Potential Action in Forcefield / Potentials.
   (c) To assign charges, use the Fix option for both Partial Chg Action and Formal Chg Action under Forcefield / Potentials.

   Note that after each change in the force field you must assign atom types and charges anew.

[7] Note that the name *Discover* has two separate meanings. The first, Discover, stands for the software package with minimization and molecular dynamics routines. The second, typed in bold (**Discover**), refers to the module available in Insight II.

To check if the assigned atom types and partial charges are correct, you can select Potential or Partial_charge in Molecule / Label to label each atom. Once you specify the information about the structure and parameters, you are ready to move to the **Discover** module. (We will not use **Discover_3** in this course).

The Constraint pulldown menu contains various atom-constraining and restraining procedures that you can select. In Parameters, you select the simulation type for Discover (Minimize, Dynamics, etc.), as well as the choice for cutoff parameters for nonbonded interactions, periodic boundary conditions (Variables), and dielectric constant (Set). Take time to familiarize yourself with the first three pulldown menus Constraint, Parameters, and Run, with Insight_help active, to learn about the various commands they contain.

To start a simulation, go to Run / Run, select desired options, choose the object for calculations, and execute.
Each Discover run is assigned a number in the order of the execution start time. The files created during the execution are identified by the calculation object (molecular system name) and the job integer (appended to the name). The file extension specifies the file type. Examples are listed below.

Discover Input Files:

- Commands (`.inp`)
- Cartesian Coordinates (`.car`)
- Molecular Data (`.mdf`)
- Force field Parameters (`.frc`)
- Restraints (`.rstrnt`)

Discover Output Files:

- Standard Output (`.out`)
- Cartesian Coordinates (final structure) (`.cor`)
- Cartesian Coordinate Archive (multiple frames) (`.arc`)
- Automatic Potential Parameter Assignment (`.prm`)
- Discover Dynamics Restart Information (`.rst`)

You can specify the files to save with Run / Files. By default, all are saved.

2. **Setting up biphenyl minimization**.

We will begin to learn about potential energy minimization for a simple yet interesting system, biphenyl (see Fig. D.2).

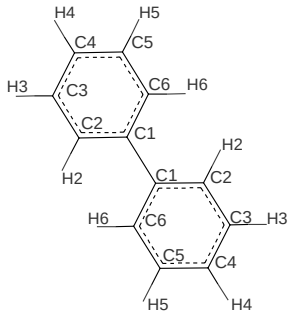

Figure D.2. Biphenyl

You will receive electronically two files containing the coordinates of biphenyl[8] The first, `biphenyl.car`, includes the structure with coplanar phenyl rings. This configuration was created with the **Builder** module by connecting two benzene rings.
The second file, `biphenyl_distorted.car`, contains the structure with each of the phenyl rings distorted from planarity.
Before displaying these structures, check that the AMBER force field is chosen. This will save work in assigning AMBER force field parameters. It will also permit you to proceed to Discover directly. (Note: For other force fields, you would have to assign parameters through Forcefield / Potentials). To open a coordinate file and display a structure, use Molecule / Get, specify Archive as the File Type, and select the desired file.

3. **Generation of energy profiles by restrained minimization**.

   A potential energy profile along some molecular coordinate, X (such as the rotamer dihedral angle χ_1), describes the dependence of the energy, minimized with respect to the remaining coordinates, on X. The simplest way to generate such a profile is to use minimization with *restraints*. Restraining a coordinate X to a specified value X^0 can be accomplished by adding harmonic penalty term,

   $$E_{\mathrm{rstr}} = K\,(X - X^0)^2,$$

   to the potential energy. After minimization, X should not deviate significantly from X^0 when the force constant K is large.[9] For a complete profile, minimum energy values must be calculated for a series of values $\{X_1^0, X_2^0,$

[8] Files can be obtained through the link to the course web site or directly from the author.

[9] *Constrained*, as opposed to restrained, minimization entails a more complex procedure to guarantee that $X = X^0$.

$X_3^0, \dots\}$ in the range of X.

For biphenyl, we will analyze the dependence of energy on the torsion angle between the planes of phenyl rings. Four dihedral angles are defined about the C1–C1 bond connecting the two rings. They are specified by the following atom quadruplets {1B:C2, 1B:C1, 1:C1, 1:C6}, {1B:C6, 1B:C1, 1:C1, 1:C2}, {1B:C2, 1B:C1, 1:C1, 1:C2}, and {1B:C6, 1B:C1, 1:C1, 1:C6}. Restraining only one of them will result in a nonplanar geometry of phenyl rings (since the remaining dihedral angles will tend to assume values associated with a lower energy). To ensure that the phenyl-ring planes are not distorted, it is necessary to restrain a *pair* of dihedral angles to the same value. You can choose the first two or the last two atom quadruplets from the list above.

The plot for the full range of the angle, $[-180.0°, 180.0°]$, can be created by first computing energy minima for a sequence of values in the range $[0.0°, 90.0°]$ (e.g., $0.0°$, $10.0°$, $20.0°$, ..., $90.0°$), and then using symmetry operations.

Start with the coplanar structure (`biphenyl.car`). Make sure that potential parameters are properly assigned.

Then select Constraint / TorsionForce. You can now proceed in different ways to calculate the energy values for the profile. For instance, you can make 10 separate minimization runs, each time specifying both restraints (Intervals set to 1). Alternatively, you can execute one run specifying the range of values for both restraints (Intervals set to 9, Starting_Angle set to 0.0, and Angle_Size set to 90.0). In the latter case, you must extract the appropriate energy values from the output file. For two restraints, defined at ten points each, 100 energy values (corresponding to all restraints) will be listed as output. Extract only those values for which the restraint targets on both angles are identical.

Use Force Constant set to the range of 2000–5000.

Switch to Parameters / Minimize and select Conjugate gradient algorithm with Gradient tolerance set to 0.001.

Note that Parameters / Set and Parameters / Variables are left at their default values. Now proceed to Run.

Another possibility is to use Run / Files to limit the number of output files. Before executing the Run / Run command check the restraints and selected minimization options using the List option.

After minimization, the dihedral angle might deviate somewhat from the value specified in the restraint. Save the final structure (both dihedral angles are around $90°$) to `biphenyl.psv` using File / Save_Folder.

You can view these structures (frames) with Trajectory / Get and Trajectory / Conformation from the **Analysis** module and determine the torsion angle value.

Plotting the profile should be done only after completing the next section of the assignment.

4. **Unrestrained minimization for biphenyl**.

   In addition to the restrained minimization calculations, perform unrestrained minimization to find the "global" energy minimum, $E_{\min}$, for biphenyl.

   Now express the profile energy E from the previous section relative to the $E_{\min}$ (i.e., $E - E_{\min}$), and plot against the dihedral angle for the full range $[-180.0°, 180.0°]$.

   Note that $E_{\min}$ may be larger than some E values. Why is that?

5. **Comparison of different force fields**.

   Repeat the energy profile calculations with the cff91 force field. Plot the results obtained with the AMBER and cff91 force fields on one plot and discuss your findings.

6. **Dependence on initial conditions**.

   Perform unrestrained minimization of biphenyl starting with the structure specified in the `biphenyl.psv` file from Part 3. Use the cff91 or AMBER force field and any minimization algorithm you wish, but use the Derivative tolerance of 0.001.

   Describe the minimization algorithm briefly and discuss your results.

7. **Assessment of the performance of various minimization algorithms in** Insight II.

   For each minimization algorithm offered in Discover record the CPU time required for convergence of the energy gradient to the target values of 10.0, 0.1, 0.001, and 0.00001 kcal/Å. Use the AMBER force field.

   For each algorithm, begin with the structure contained in the file `biphenyl_distorted.car`. Select the desired Algorithm from Parameters / Minimize; set Iterations to 5000; specify the first target value of the derivative; execute; and proceed to execute the Run/Run command. After this job is completed, change the derivative tolerance to the

next target value and repeat minimization. Extract the computational times and values of minima from each output file. Do not increase the number of Iterations above 5000. If the specified convergence is not reached with this threshold, note that in your report.

Repeat this procedure for each of the remaining algorithms. (Remember to start with the structure from `biphenyl_distorted.car` file.) Construct a table comparing the performance of minimization algorithms in the different regions of derivative tolerance (report the timing and energy minimum values).

On the basis of these results, and the information you have learned in class, suggest a simulation schedule to achieve an optimal minimization of a large molecule. Note that for our small system the gradient norm associated with the initial configuration of biphenyl is not extremely large.

Background Reading from Coursepack

- M. Karplus and G. A. Petsko, "Molecular Dynamics Simulations in Biology", *Nature* **347**, 631–639 (1990).

Assignment 11: A Global Optimization Contest!

Our goal is to compute the lowest energy structure for the pentapeptide met-enkephalin, whose sequence is **Tyr–Gly–Gly–Phe–Met**. Many local minima exist for this molecule, so it is a challenge to reach the global minimum. *The student who finds the structure of the lowest energy will receive a prize from the instructor.*

The rules of this contest are:

1. use a molecule with *charged* COO^- and NH_3^+ ends
2. use the AMBER force field
3. use the distance dependent dielectric constant (**Discover** module, Parameters / Set command, Dist_Dependent button on)
4. use $1/2$ as the scale factor for 1–4 nonbonded interactions (i.e., Parameters / Scale_Terms command, p1_4 button on, and specify 0.5)

You can use *any* technique mentioned in this course (energy minimization, molecular dynamics, Monte Carlo sampling), as well as any other resources (e.g., web and literature), to find the global minimum of the pentapeptide.

Be Creative.

Hand in a detailed report describing how you reached the minimum for met-enkephalin and any particular difficulties, or interesting observations, you encountered along the way. Attach the Cartesian coordinate file and the energy value reached.

Also submit a three-dimensional picture of the configuration of lowest energy along with a table specifying all associated bond lengths and bond angle values, and the $\{\phi, \psi\}$ and χ dihedral-angle values per residue.

To qualify for consideration of the prize, send electronically the coordinate file with the minimized structure to the instructor and TA.

Good Luck!

Background Reading from Coursepack

- K. A. Dill and H. S. Chan, "From Levinthal to Pathways to Funnels", *Nature Struc. Biol.* **4**, 10–19 (1997).
- T. Lazaridis and M. Karplus, " 'New View' of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations", *Science* **278**, 1928–1931 (1997).

Assignment 12: Monte Carlo Simulations

1. **Random Number Generators**. Investigate the types of random number generators available on: (a) your local computing environment and (b) a mathematical package that you frequently use. How good are they? Is either one adequate for long molecular dynamics runs? Suggest how to improve them and test your ideas.

   To understand some of the defects in linear congruential random number generators, consider the sequence defined by the formula $y_{i+1} = (a\, y_i + c) \bmod M$, with $a = 65539, M = 2^{31}$, and $c = 0$. (This defines the infamous random number generator known as RANDU developed by IBM in the 1960s, which subsequent research showed to be seriously flawed). A relatively small number of numbers in the sequence (e.g., 2500) can already reveal a structure in three dimensions when triplets of consecutive random numbers are plotted on the unit cube. Specifically, plot consecutive pairs and triplets of numbers in two and three-dimensional plots, respectively, for an increasing number of generated random numbers in the sequence, e.g., 2500, 50,000, and 1 million. (Hint: Figure D.3 shows results from 2500 numbers in the sequence).

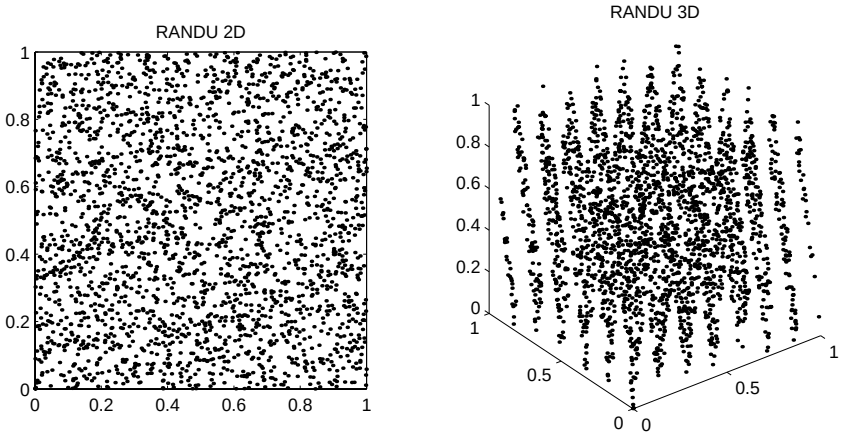

Figure D.3. Plots generated from pairs and triplets of consecutive points in the linear congruential generator known as RANDU defined by $a = 65539, M = 2^{31}$, and $c = 0$ when 2500 total points in the sequence are generated.

2. **MC Means**. Propose and implement a Monte Carlo procedure to calculate π based on integration. How many MC data points are needed to yield an answer correct up to 5 decimal places? What is the computational time

involved? Show a table of your results displaying the number of MC steps, the associated π estimate, and the calculated error.

3. **Gaussian Variates**. You are stranded in an airport with your faithful laptop with one hour to spare until the deadline for emailing your homework assignment to your instructor. The assignment (next item) relies on a *Gaussian random number generator*, but you have forgotten the appropriate formulas involved in the commonly used Box/Muller/Marsaglia transformation approach. Fortunately, however, you remember the powerful Central Limit Theorem in basic probability and decide to form a random Gaussian variate by sampling N uniform random variates $\{x_i\}$ on the unit interval as

$$\bar{y} = \sum_{i=1}^{N} x_i \,.$$

You quickly program the expression:

$$y = \sqrt{\frac{1}{\sigma^2(\bar{y})}} \sum_{i=1}^{N} [x_i - \mu(\bar{y})]$$

where above σ^2 is the standard deviation of $\bar{y} = N\sigma^2(x)$ and the mean $\mu(\bar{y}) = N\mu(x)$. [Recall that the uniform distribution has a mean of 1/2 and variance of 1/12].

How large should N be, you wonder. You must finish the assignment in a hurry. To have confidence in your choice, you set up some tests to determine when N is sufficiently large, and send your resulting routine, along with your testing reports, and results for several choices of N.

4. **Brownian Motion**. Now you can use the Gaussian variate generator above for propagating *Brownian motion* for a single particle governed by the biharmonic potential $U(x) = kx^4/4$. Recall that Brownian motion can be mimicked by simulating the following iterative process for the particle's position:

$$x^{n+1} = x^n + \frac{\Delta t}{m\gamma} F^n + R^n$$

where

$$\langle R^i R^j \rangle = \frac{2k_B T \Delta t}{m\gamma} \delta_{ij}, \qquad \langle R^i \rangle = 0 \,.$$

Here m is the particle's mass; γ is the collision frequency, also equal to ξ/m where ξ is the frictional constant; and F is the systematic force. You are required to test the obtained mean square atomic fluctuations against the

known result due to Einstein:

$$\langle x^2 \rangle = 2 \left(\frac{k_B T}{m\gamma} \right) t = 2D\, t\,,$$

where D is the diffusion constant.
The following parameters may be useful to simulate a single particle of mass $m = 4 \times 10^{-18}$ kg and radius $a = 100$ nm in water: by Stokes' law, this particle's friction coefficient is $\xi = 6\pi\eta a = 1.9 \times 10^{-9}$ kg/s, and $D = k_B T/\xi = 2.2 \times 10^{-12}\text{m}^2/\text{s}$. You may, however, need to scale the units appropriately to make the computations reasonable.

Plot the mean square fluctuations of the particle as a function of time, compare to the expected results, and show that for $t \gg 1/\gamma = 2 \times 10^{-9}$ s the particle's motion is well described by random-walk or diffusion process.

Background Reading from Coursepack

- T. Schlick, E. Barth, and M. Mandziuk, "Biomolecular Dynamics at Long Timesteps: Bridging the Timescale Gap Between Simulation and Experimentation", *Ann. Rev. Biophys. Biomol. Struc.* **26**, 179–220 (1997).
- E. Barth and T. Schlick, "Overcoming Stability Limitations in Biomolecular Dynamics: I. Combining Force Splitting via Extrapolation with Langevin Dynamics in LN", *J. Chem. Phys.* **109**, 1617–1632 (1998).
- L. S. D. Caves, J. D. Evanseck, and M. Karplus, "Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin", *Prot. Sci.* **7**, 649–666 (1998).

Assignment 13: Advanced Exercises in Monte Carlo and Minimization Techniques

1. Study the function:

$$E(x,y) = ax^2 + by^2 + c\,(1-\cos\gamma x) + d\,(1-\cos\delta y)\,. \qquad \text{(D.11)}$$

Note that is has many local minima and a global minimum at $(x,y) = (0,0)$. Minimize $E(x,y)$ with $a = 1, b = 2, c = 0.3, \gamma = 3\pi$, $d = 0.4$, and $\delta = 4\pi$ by the standard simulated annealing method. Use the starting point $(1,1)$ and step perturbations $\Delta x = 0.15$, and set β in the range of 3.5 to 4.5. Limit the number of steps to ~ 150. Now implement the *variant* of the simulated annealing method where acceptance probabilities for steps with $\Delta E < 0$ are proportional to $\exp(-\beta E^g \Delta E)$, with the exponent $g = -1$. Analyze and compare the efficiency of the searches in both cases. It will be useful to plot all pairs of points (x,y) that are generated by the method and distinguish 'accepted' from 'rejected' points.

2. Devise a different variant of the basic simulated annealing minimization method that would incorporate *gradient* information to make the searches more efficient.

3. Consider the following global optimization deterministic approach based on the *diffusion equation* as first suggested by Scheraga and colleagues (L. Piela, J. Kostrowicki, and H. A. Scheraga, "The Multiple-Minima Problem in Conformational Analysis of Molecules. Deformation of the Potential Energy Hypersurface by the Diffusion Equation Method", *J. Chem. Phys.* **93**, 3339–3346 (1989)).
The basic idea is to deform the energy surface smoothly. That is, we seek to make "shallow" wells in the potential energy landscape disappear iteratively until we reach a global minimum of the deformed function. Then we "backtrack" by successive minimization from the global minimum of the transformed surface in the hope of reaching the global minimum of the real potential energy surface. This idea can be implemented by using the heat equation where T represents the temperature distribution in space x, and t represents time:

$$\frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \qquad \text{(D.12)}$$

$$T(x,0) = E(x)\,. \qquad \text{(D.13)}$$

Here, the boundary condition at time $t = 0$ equates the initial temperature distribution with the potential energy function $E(x)$. Under certain conditions (e.g., E is bounded), a solution exists. Physically, the application of this equation exploits the fact that the heat flow (temperature distribution) should eventually settle down.

To formulate this idea, let us for simplicity consider first a one-dimensional problem where the energy function E depends on a scalar x. Let $E^{(n)}(x)$ represent the nth derivative of E with respect to x and define the transformation operator $\mathcal{S}$ on the energy function E for $\beta > 0$ as follows:

$$\mathcal{S}[E(x)] = E(x) + \beta E^{(2)}(x)\,. \tag{D.14}$$

That is, we have:

$$\begin{aligned}
\mathcal{S}^0 E &= E \\
\mathcal{S}^1 E &= E + \beta E^{(2)} \\
\mathcal{S}^2 E &= E + 2\beta E^{(2)} + \beta^2 E^{(4)} \\
\mathcal{S}^3 E &= E + 3\beta E^{(2)} + 3\beta^2 E^{(4)} + \beta^3 E^{(8)} \\
\vdots &\qquad\quad \vdots \\
\mathcal{S}^N E &= (1 + \beta d^2/dx^2)^N E\,.
\end{aligned}$$

Now writing $\beta = t/N$ where t is the time variable, and letting $N \to \infty$, we write:

$$\exp\left(td^2/dx^2\right) E \equiv \exp(A(t))\, E = \left[1 + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots\right]. \tag{D.15}$$

Thus we can define $T(t)$ as

$$T(t) = \exp((A(t)) = \exp(td^2/dx^2)\,. \tag{D.16}$$

In higher dimensions, let x represent the collective vector of n independent variables; we replace the differential operator above d^2/dx^2 by the *Laplacian operator*, that is

$$\Delta = \sum_{i=1}^{n} \partial^2/\partial x_i\,.$$

Using this definition, we can also write

$$T(t) = T_1(t)\, T_2(t)\, \cdots\, T_n(t)$$

where

$$T_i = \exp(t\partial^2/\partial x_i)\,.$$

This definition produces the heat equation (D.12, D.13) since

$$\begin{aligned}
\frac{\partial T(t)[E(x)]}{\partial t} &= \left[\frac{dA}{dt} + \frac{2A}{2}\frac{dA}{dt} + \frac{3A^2}{3!}\frac{dA}{dt} + \cdots\right][E] \\
&= \left[1 + A + \frac{A^2}{2} + \cdots\right]\frac{d^2}{dx^2}[E]
\end{aligned}$$

$$= \frac{\partial^2 T(t)}{\partial x^2}[E(x)]\,.$$

In practice, the diffusion equation method for global optimization is implemented by solving the heat equation by Fourier techniques (easy, for example, if we have dihedral potentials only) or by solving for T up to a sufficiently large time t. This solution, or approximate solution (representing $E(x,t)$ for some large t), is expected to yield a deformed surface with one (global) minimum. With a local minimization algorithm, we compute the global minimum x^* of the deformed surface, and then begin an iterative deformation/minimization procedure from x^* and $E(x,t)$ so that at each step we deform backwards the potential energy surface and obtain its associated global minimum ($E(x,t) \longrightarrow E(x,t-\Delta t)$ and x^* to x^1, $E(x,t-\Delta t) \longrightarrow E(x,t-2\Delta t)$ and x^1 to x^2, $\cdots$ $E(x,0) \longrightarrow x^k$). Of course, depending on how the backtracking is performed, different final solutions can be obtained.

(a) To experiment with this interesting diffusion-equation approach for global minimization, derive a general form for the deformation operator $T(t) = \exp(td^2/dx^2)$ on the following special functions $E(x)$: (*i*) polynomial functions of degree n, and (*ii*) trigonometric functions $sin\omega x$ and $cos\omega x$, where ω is a real-valued number (frequency). What is the significance of your result for (*ii*)?

(b) Apply the deformation operator $T(t) = \exp(td^2/dx^2)$ to the quadratic function

$$E(x) = x^4 + ax^3 + bx^2\,, \tag{D.17}$$

with $a = 3$ and $b = 1$. Evaluate and plot your resulting $T(t)E(x)$ function at $t = 0, \Delta t, 2\Delta t, ...$, for small time increments Δt until the global minimum is obtained.

(c) Apply the deformation operator $T(t)$ for the two-variable function in eq. (D.11). Examine behavior of the deformation as $t \to \infty$ as a function of the constants a and b. Under what conditions will a unique minimum be obtained as $t \to \infty$?

4. Use Newton minimization to find the minimum of the two-variable function in equation (D.11) and the one-variable function in equation (D.17). It is sufficient for the line search to use simple bisection: $\lambda = 1, 0.5$, etc., or some other simple backtracking strategy. For the quartic function, experiment with various starting points.