

# Kaplan Software Requirements Specification for Conformer Searching

Jen Garner

October 22, 2018

# 1 Revision History

| Date                          | Version | Notes   |
|-------------------------------|---------|---|
| October 4, 2018<br>(Thursday) | 1.0     | first draft for submission  |
| October 22,<br>2018 (Monday)  | 1.1     | enumerate non-functional requirements, add dictionary to<br>Texstudio and fix spelling errors |

## 2 Reference Material

Units, constants, symbols, abbreviations, and acronyms are abbreviated here, as used in this document.

### 2.1 Table of Units and Constants

Throughout this document SI (Système International d’Unités) is employed as the unit system. In addition to the basic units, several derived units are used as described below. For each unit, the symbol is given followed by a description of the unit and the SI name. Some common constants are given.

| symbol       | unit                       | SI                                     |
|--------------|----------------------------|--|
| m            | length                     | metre                                  |
| Å            | length                     | $1\text{Å} = 1\times 10^{-10}\text{m}$ |
| kg           | mass                       | kilogram                               |
| s            | time                       | second                                 |
| J            | energy                     | Joule                                  |
| F            | electrical capacitance     | farad                                  |
| C            | electric charge            | coulomb (A s)                          |
| $h$          | Planck’s constant          | $6.62607004\times 10^{-34}\text{J s}$  |
| $\hbar$      | reduced Planck’s constant  | $h/2\pi\text{J s}$                     |
| $m_e$        | mass of an electron        | $9.10938\times 10^{-31}\text{kg}$      |
| $q_e$        | charge of an electron      | $1.60217662\times 10^{-19}\text{C}$    |
| $\epsilon_0$ | permittivity of free space | $8.854187817\times 10^{12}\text{F/m}$  |

[Only include the units that your SRS actually uses —SS]

### 2.2 Table of Symbols

The table that follows summarizes the symbols used in this document along with their units.

| symbol | unit     | description   |
|--------|----------|---|
| $n_a$  | unitless | number of atoms in the input molecule   |
| $n_G$  | unitless | number of conformers (distinct geometries) being simultaneously optimised           |
| $G_i$  | Å        | geometry for conformer $i$ ; matrix of Cartesian coordinates with shape( $n_a$ , 3) |
| $D_i$  | °        | list of dihedral angles for conformer $i$ ; array of length $n_a - 3$               |

|            |                        |   |
|------------|------------------------|---|
| $E_i$      | J                      | potential energy for conformer $i$  |
| $S_E$      | J                      | sum of conformer potential energies   |
| $C_E$      | $\frac{1}{\text{J}}$   | coefficient for $S_E$ in the fitness function   |
| $S_{RMSD}$ | $\text{\AA}$           | sum of root-mean-square distances between all possible combinations of conformer geometries |
| $C_{RMSD}$ | $\frac{1}{\text{\AA}}$ | coefficient for $S_{RMSD}$ in the fitness function  |
| $Fit_G$    | unitless               | the fitness of the set of conformers  |
| $E$        | J                      | energy (of conformer)   |
| $\hat{H}$  | J                      | Hamiltonian operator  |
| $\Psi$     | unitless               | wavefunction  |
| $\sigma$   | unitless               | spin ( $\alpha$ or $\beta$ )  |
| $\phi$     | unitless               | atomic orbital  |
| $r$        | $\text{\AA}$           | position of electron; array of length 3   |
| $R$        | $\text{\AA}$           | position of nuclei; array of length 3   |
| $e^-$      | N/A                    | electron  |

---

[Use your problems actual symbols. The si package is a good idea to use for units. —SS]

## 2.3 Abbreviations and Acronyms

| Document-Specific Acronyms |  |
|----------------------------|--|
| Symbol                     | Description                                  |
| A                          | Assumption                                   |
| DD                         | Data Definition                              |
| GD                         | General Definition                           |
| GS                         | Goal Statement                               |
| IM                         | Instance Model                               |
| LC                         | Likely Change                                |
| PS                         | Physical System Description                  |
| R                          | Requirement                                  |
| SRS                        | Software Requirements Specification          |
| Kaplan                     | <a href="#">[Kaplan —SS]</a>                 |
| T                          | Theoretical Model                            |
| Chemical Acronyms          |  |
| Symbol                     | Description                                  |
| BO                         | Bond Order                                   |
| SMILES                     | Simplified molecular-input line-entry system |
| AO                         | atomic orbital                               |
| MO                         | molecular orbital                            |
| LCAO                       | linear combination of atomic orbitals        |
| STO                        | slater-type orbital                          |
| GTO                        | Gaussian-type orbital                        |
| QCM                        | quantum chemical method                      |
| BS                         | basis set                                    |
| RMSD                       | root-mean square deviation                   |

[\[Add any other abbreviations or acronyms that you add —SS\]](#)

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Revision History</b>                          | <b>i</b>  |
| <b>2</b> | <b>Reference Material</b>                        | <b>ii</b> |
| 2.1      | Table of Units and Constants . . . . .           | ii        |
| 2.2      | Table of Symbols . . . . .                       | ii        |
| 2.3      | Abbreviations and Acronyms . . . . .             | iv        |
| <b>3</b> | <b>Introduction</b>                              | <b>1</b>  |
| 3.1      | Purpose of Document . . . . .                    | 2         |
| 3.2      | Scope of Requirements . . . . .                  | 2         |
| 3.3      | Characteristics of Intended Reader . . . . .     | 2         |
| 3.4      | Organization of Document . . . . .               | 3         |
| <b>4</b> | <b>General System Description</b>                | <b>3</b>  |
| 4.1      | System Context . . . . .                         | 3         |
| 4.2      | User Characteristics . . . . .                   | 4         |
| 4.3      | System Constraints . . . . .                     | 4         |
| <b>5</b> | <b>Specific System Description</b>               | <b>4</b>  |
| 5.1      | Problem Description . . . . .                    | 4         |
| 5.1.1    | Terminology and Definitions . . . . .            | 4         |
| 5.1.2    | Physical System Description . . . . .            | 5         |
| 5.1.3    | Goal Statements . . . . .                        | 7         |
| 5.2      | Solution Characteristics Specification . . . . . | 8         |
| 5.2.1    | Assumptions . . . . .                            | 8         |
| 5.2.2    | Theoretical Models . . . . .                     | 9         |
| 5.2.3    | General Definitions . . . . .                    | 10        |
| 5.2.4    | Data Definitions . . . . .                       | 11        |
| 5.2.5    | Instance Models . . . . .                        | 12        |
| 5.2.6    | Data Constraints . . . . .                       | 13        |
| 5.2.7    | Properties of a Correct Solution . . . . .       | 14        |
| <b>6</b> | <b>Requirements</b>                              | <b>15</b> |
| 6.1      | Functional Requirements . . . . .                | 15        |
| 6.2      | Nonfunctional Requirements . . . . .             | 16        |
| <b>7</b> | <b>Likely Changes</b>                            | <b>16</b> |
| <b>8</b> | <b>Traceability Matrices and Graphs</b>          | <b>17</b> |

|          |                               |           |
|----------|-------------------------------|-----------|
| <b>9</b> | <b>Appendix</b>               | <b>19</b> |
| 9.1      | Symbolic Parameters . . . . . | 19        |

### 3 Introduction

Molecular geometry is a necessary piece of information with regards to running calculations. Most computational chemists will run a geometry optimization for their molecule of interest before they can investigate other properties, such as charge distribution and interaction energies. Without breaking any bonds, there are multiple ways in which the atoms can be positioned such that an optimal geometry is obtained - these geometries are called conformational isomers (sometimes abbreviated as conformers). Rather than performing an exhaustive search of all possible geometries, Kaplan is a package designed for the efficient exploration and optimization of molecular geometry with the end goal of procuring a set of conformers.

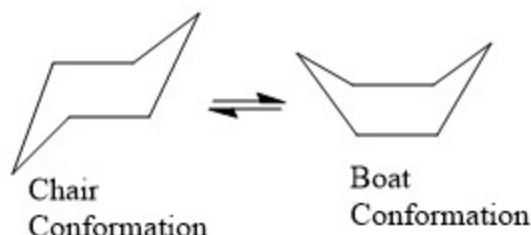


Figure 1: Two examples of conformational isomers of cyclohexane.

A classic example of conformational isomers is the boat versus the chair conformation of cyclohexane (see Figure 1). A simple geometry optimization of cyclohexane would only afford the global minimum to the potential energy landscape, but, in using Kaplan, it will be possible to find the local minima of this molecule. Knowing the local minima is useful in areas such as drug design, where the molecule is restricted by its environment (for example, in the interaction with an enzyme) to have a certain shape. Molecules can also adopt different conformations when they are exposed to certain solvents, especially when the molecule in question has charge separation.

[This SRS template is based on [Smith and Lai \(2005\)](#); [Smith et al. \(2007\)](#). It will get you started, but you will have to make changes. Any changes to section headings should be approved by the instructor, since that implies a deviation from the template. Although the bits shown below do not include type information, you may need to add this information for your problem. —SS]

[Feel free to change the appearance of the report by modifying the LaTeX commands. —SS]

[If you are documenting a family of models, you can start from this same template, but you will have to add a section for variabilities. For program families you should look at [Smith \(2006\)](#); [Smith et al. \(2017\)](#). You should be able to do one document that captures the commonality analysis and the requirements. —SS]



### 3.1 Purpose of Document

This document outlines the requirements that the software, Kaplan, must meet. It is an abstract document that does not describe the details of implementation. Rather, a discussion of its intended functionalities, performance, goals, and qualities will be presented. This document will be used as a reference guide when writing the software design specification and the software verification and validation plan. The inputs, expected outputs, and user characteristics for the program will be outlined, as well as the theory needed for searching a molecular geometry space. Any assumptions inherent to solving the problem will be mentioned in this document.

### 3.2 Scope of Requirements

Kaplan will optimize the geometry of an input molecule to find its conformers by searching the potential energy surface. The size of molecule that can be accommodated depends greatly on the other inputs of the system, such as level of theory and convergence criteria, but should be held within reasonable bounds such as a maximum of a couple hundred atoms. During optimization, the bond lengths and bond angles will be held fixed, and only the dihedral angles will be manipulated. As a result, molecules with less than 4 atoms will remain unchanged after optimization (no dihedral angles). Furthermore, the conformer with the lowest energy is not likely to represent the true minimum of the potential energy landscape without subsequent optimization of the other bond angles/lengths. Lastly, the system is not intended to be impervious to abnormal or forced bonding behaviour.

### 3.3 Characteristics of Intended Reader

The reader should have taken first-year undergraduate chemistry, physics, and mathematics. They should be familiar with some bonding theory in molecules (valence bond (VB) theory, molecular orbital (MO) theory, and valence-shell electron pair repulsion (VSEPR)) and comfortable with the concept of optimization (especially of a multi-dimensional surface).

To understand the energy calculations, the reader should have taken a quantum physics or quantum chemistry undergraduate course. Having a basic understanding the following terms will be useful:

- Schrödinger equation
- Hartree Fock
- wavefunction
- basis set
- Hamiltonian operator
- restricted versus unrestricted quantum chemistry calculations

### 3.4 Organization of Document

This document follows the template outlined in [Smith and Lai \(2005\)](#). Section 1 provides tables for all of the units, symbols, acronyms, and abbreviations used throughout the document. Section 3 is an overview of the purpose and scope of the system, including an explanation the intended reader for this document. Section 4 goes into more detail regarding the system and its inputs, and describes the responsibilities of the user versus the program.

## 4 General System Description

The interactions between the system and its environment are discussed in this section. The user characteristics and system constraints are also given.

### 4.1 System Context

[Your system context will likely include an explicit list of user and system responsibilities—SS]

Here the system context for Kaplan is shown (Figure 2). With respect to inputs, convergence conditions includes number of expected conformers, energetic requirements, and the method and basis set used to perform energy calculations. As for outputs, an energy will be returned for each conformer geometry.



Figure 2: The circles represent user interaction, and the rectangle represents the program. The inputs are fed to the program, and the outputs are given back to the user, as indicated by the arrows.

- User Responsibilities:
  - provide chemically and computationally reasonable input
  - set the convergence criteria (how long will the program search for conformers? What energetic requirements should the conformers possess? Estimate the number of expected conformers, etc.).
- Kaplan Responsibilities:
  - prepare molecular geometry for optimization

- find conformers and calculate their energies
- determine if the convergence conditions have been met
- ensure that returned conformers have distinct geometries

## 4.2 User Characteristics

The user of Kaplan should have taken first year undergraduate physics, chemistry, and mathematics. The user must understand the impact of changing the quantum mechanical method and basis set for energy calculations, which implies that they understand basic quantum mechanics (including how to solve the Schrödinger equation - 3rd year quantum physics course). The user should have a sense of whether their input geometry (where applicable) will converge under optimization

## 4.3 System Constraints

[You may not have any system constraints —SS] An evolutionary algorithm will be used to search the potential energy space. Energies will be calculated using an open-source quantum chemistry package.

# 5 Specific System Description

This section first presents the problem description, which gives a high-level view of the problem to be solved. This is followed by the solution characteristics specification, which presents the assumptions, theories, definitions and finally the instance models. [Add any project specific details that are relevant for the section overview. —SS]

## 5.1 Problem Description

[what problem does your program solve? —SS]

Kaplan is designed to search the potential energy surface for conformational isomers of an input molecule by manipulating dihedral angles. The solution to a fitness function, which is related to the energies of and spatial differences between conformers, will be found and maximized.

### 5.1.1 Terminology and Definitions

[Does the wording here imply that none of these definitions can be used in previous sections? Should the early sections of this document be so abstract as to not use any terminology? —JG]

This subsection provides a list of terms that are used in the subsequent sections and their meaning, with the purpose of reducing ambiguity and making it easier to correctly understand the requirements:

[Do I need to define a term if, as per the intended reader requirements, the reader should know what that terms means? For example, should I define covalent bond here? —JG]

- **molecule:** a collection of atoms that are related in space, either covalently or non-covalently.
- **bond angle:** the angle formed between three atoms.
- **bond length:** the distance between the centre of masses of two atoms.
- **dihedral angle** the angle between two intersecting planes, where each plane bisects 3 atoms.
- **bond order:** the number of chemical bonds connecting two atoms. This value sometimes depends on the theory used.
- **conformational isomer:** molecules that have the same number of atoms that are related by free rotation about single bonds. May also be referenced as conformer in the text.
- **quantum chemical method:** the strategies used to solve the Schrödinger equation.
- **basis set:** how to describe the molecule mathematically such that the Schrödinger equation can be evaluated.
- **z-matrix:** an input file type that uses dihedral angles and connectivities rather than Cartesian coordinates.

### 5.1.2 Physical System Description

The physical system of Kaplan includes the following elements:

PS1: molecule for which to find conformers. This molecule is specified by a file, such as xyz (Cartesian coordinates) or a z-matrix, a SMILES string, or a name. From this geometry, the dihedral angles can be obtained.

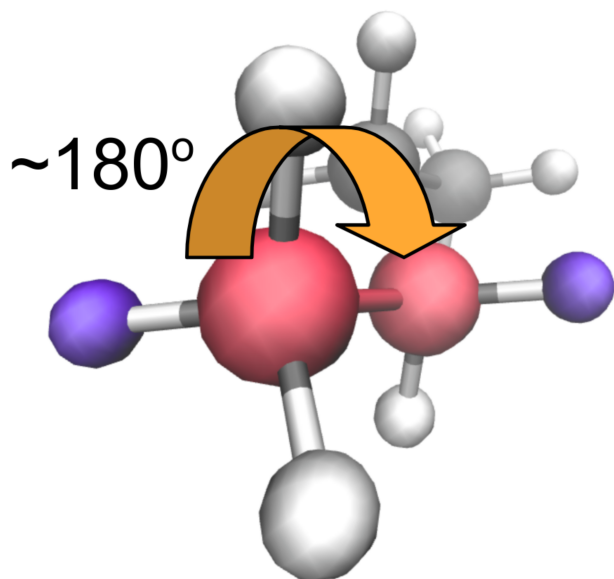


Figure 3: The dihedral angle formed between the 4 highlighted atoms is approximately  $180^\circ$ . Note that if the molecule was rotated such that the blue-highlighted hydrogen atom on the left went behind the blue-highlighted hydrogen atom on the right, then the dihedral angle would then be  $0^\circ$ .

An example of the physical system for the molecule butane ( $C_4H_{10}$ ) can be found in Table 1.

| Z-Matrix                    | XYZ (Cartesian)<br>Coordinates | SMILES<br>String | Dihedral<br>Angles |
|-----------------------------|--------------------------------|------------------|--------------------|
| C 1                         | 14                             | CCCC             | 178.8              |
| C 1 1.527                   | butane                         |                  | 238.9              |
| C 1 1.520 2 111.489         | C -0.5630 0.5160 0.0071        |                  | 121.2              |
| C 2 1.520 1 111.497 3 178.8 | C 0.5630 -0.5159 0.0071        |                  | 300.1              |
| H 1 1.096 2 109.879 3 238.9 | C -1.9293 -0.1506 -0.0071      |                  | 57.8               |
| H 1 1.096 2 109.864 3 121.2 | C 1.9294 0.1505 -0.0071        |                  | 299.7              |
| H 2 1.096 1 109.866 3 300.1 | H -0.4724 1.1666 -0.8706       |                  | 60.2               |
| H 2 1.096 1 109.874 3 57.8  | H -0.4825 1.1551 0.8940        |                  | 180.0              |
| H 3 1.095 1 111.018 2 299.7 | H 0.4825 -1.1551 0.8940        |                  | 299.7              |
| H 3 1.095 1 110.996 2 60.2  | H 0.4723 -1.1665 -0.8706       |                  | 180.0              |
| H 3 1.095 1 110.265 2 180.0 | H -2.0542 -0.7710 -0.9003      |                  | 60.2               |
| H 4 1.095 2 111.009 1 299.7 | H -2.0651 -0.7856 0.8742       |                  |                    |
| H 4 1.095 2 110.262 1 180.0 | H -2.7203 0.6060 -0.0058       |                  |                    |
| H 4 1.095 2 110.989 1 60.2  | H 2.0542 0.7709 -0.9003        |                  |                    |
|                             | H 2.7202 -0.6062 -0.0059       |                  |                    |
|                             | H 2.0652 0.7854 0.8743         |                  |                    |

Table 1: The physical system of Kaplan, using butane as an example molecule. The first three columns represent equivalent geometries, given in three different formats. The last column is the list of dihedral angles to be optimised by the program. During optimization, it is important that the dihedral angles map back to the input geometry using the same ordering for the atoms.

[A figure here may make sense for most SRS documents —SS]

### 5.1.3 Goal Statements

Given an input geometry and a list of convergence criteria for a molecule [inputs —SS], the goal statement is:

GS1: To create a set of conformational isomers that represent the local minima of the potential energy surface by optimizing the molecule’s dihedral angles. Alternatively, this GS can be described as: maximize the fitness function,  $Fit_G$ . [One sentence description of the goal. There may be more than one. Each Goal should have a meaningful label. —SS]

## 5.2 Solution Characteristics Specification

The instance models that govern Kaplan are presented in Subsection 5.2.5. The information to understand the meaning of the instance models and their derivation is also presented, so that the instance models can be verified.

### 5.2.1 Assumptions

This section simplifies the original problem and helps in developing the theoretical model by filling in the missing information for the physical system. The numbers given in the square brackets refer to the theoretical model [T], general definition [GD], data definition [DD], instance model [IM], or likely change [LC], in which the respective assumption is used.

[Short description of each assumption. Each assumption should have a meaningful label. Use cross-references to identify the appropriate traceability to T, GD, DD etc., using commands like dref, ddref etc. —SS]

- A1: The initial bond lengths and bond angles, regardless of the type of input specification, will allow for the energy calculations to converge. [T1]
- A2: The energy of the molecule is calculable (using quantum mechanics). [T1]
- A3: The potential energy surface can be represented by a real-valued, continuous function that contains at least one minimum. [IM1]
- A4: The viability of a set of conformer geometries can be evaluated by solving for the fitness function,  $Fit_G$ , whose inputs depend on the RMSD distances between the conformers and the energies of those conformers. A bigger value for this function implies a better set of conformers. [IM1][DD1][DD2]
- A5: The fitness function is linear combination of the sum of energies and the sum of RMSD values. [IM1]
- A6: Conformers correspond to local minima on the potential energy surface. [IM1]  
[A6 might be too similar to A7. —JG].
- A7: Within one potential well, the molecular geometry corresponding to a lower (more negative) energy will be more stable than a geometry of a higher energy. [IM1]
- A8: Conformers can be found by manipulating dihedral angles only (i.e. bond lengths and bond angles will be held fixed). [IM1]
- A9: The molecular composition does not change during the optimization [T1]
- A10: The molecule is the only calculable object in the system (no solvent, other molecules, etc.). This assumption does not exclude guest-host chemistry, or non-bonded “molecules”, from the list of permissible inputs. [IM1]

A11: The conformer space is independent of the ordering of the input atoms. This may change (subject to proof/benchmarking). [IM1]

### 5.2.2 Theoretical Models

This section focuses on the general equations and laws that Kaplan is based on. A quantum chemistry approach is used since this level of theory allows for a more accurate calculation of the molecular energy. Some assumptions are needed for these energy calculations in order to make the equations solvable, which will be discussed briefly here.

[Modify the examples below for your problem, and add additional models as appropriate. —SS]

|             |   |
|-------------|---|
| Number      | T1  |
| Label       | <b>Non-Relativistic Time-Independent Schrödinger Equation</b>   |
| Equation    | $\hat{H} \Psi\rangle = E \Psi\rangle$   |
| Description | Most quantum chemical methods are focused on solving the above equation for a molecule with some number of nuclei and electrons. This equation is essentially an eigenvalue problem, where E gives the energy of a system (the eigenvalues), $\Psi$ is the wavefunction, and $\hat{H}$ is the Hamiltonian operator. An example of the electronic Hamiltonian operator can be found in the Appendix (Section 9). |
| Source      | Szabo and Ostlund (1996)  |
| Ref. By     | DD1   |

The Hamiltonian, in its most explicit form, is a sum of the kinetic energies of the electrons ( $\hat{T}_e$ ), the kinetic energies of the nuclei ( $\hat{T}_n$ ), the electron-electron potential energy ( $\hat{V}_{ee}$ ), the electron-nuclear potential energy ( $\hat{V}_{en}$ ), and the nuclear-nuclear potential energy ( $\hat{V}_{nn}$ ). The potential energy terms arise from Coulombic repulsive and attractive forces. The way in which the Schrödinger equation is solved and how the Hamiltonian is constructed are described by the quantum chemical method (QCM) that is used. Some QCM examples include Hartree-Fock, Coupled-Cluster, Configuration-Interaction, perturbation theory, and density functional theory.

The Born-Oppenheimer approximation is commonly used in quantum chemical methods. Since the nuclei are much larger than the electrons, they move much more slowly. As a result, the nuclei are considered fixed and only the electrons move in the system. Therefore, the  $\hat{T}_n$  term of the Hamiltonian is neglected and the  $\hat{V}_{nn}$  term becomes constant. When a constant is added to an operator, the eigenfunctions do not change (wavefunction is the same). The eigenvalues (the energies) are added to the constant to get the final result. The electronic Hamiltonian in the Appendix is shown explicitly after these assumptions have been applied.



The wavefunction has no physical meaning; it is a complex-valued function for a single particle where the input is the position vector. The wavefunction squared ( $|\psi(r)|^2$ ) is proportional to the probability of finding the particle at position  $r = x, y, z$ . That is, the integration over all space for  $|\psi(r)|^2$  is equal to one, because the particle (if it exists) must be somewhere in space.

For an electron, the wavefunction describes a hydrogen atomic orbital (AO), and its inputs are the spin ( $\sigma$ ) of the electron ( $\alpha$  or  $\beta$ ) and the position ( $r$ ). For a multi-electron wavefunction, we have a linear combination of atomic orbitals (LCAO) to give the molecular orbital (MO). Since the number of orbitals needed to exactly describe a molecule is infinite, the MO is approximated by contracting the number of AO to a finite set (and thus a finite space), affording  $\Psi(r, \sigma) = \sum_{i=1}^k c_i \phi(r, \sigma)_i$ , where  $\phi(r, \sigma)_i$  is the AO,  $c_i$  is a constant described by the basis set, and  $k$  is the number of AO.

There are two main types of AO in quantum chemistry - Slater-type orbitals (STO) and Gaussian-type orbitals (GTO). STO have the form:  $\phi(r, \sigma) = ce^{-\zeta r}$ , whereas GTO have the form:  $\phi(r, \sigma) = ce^{-\zeta r^2}$ , where  $c$  is a value that depends on normalization and the principle quantum number,  $\zeta$  is a constant related to the effective nuclear charge, and  $r$  is the distance of the electron from the nucleus. STO are more accurate, have poor near-nuclear behaviour (approach  $\infty$ ), but are expensive in terms of the evaluation of the integrals. GTO, on the other hand, are much faster to evaluate, but are less accurate. When a basis set is chosen for a quantum chemical calculation, the set of one-particle functions (AO) used to build the MO are described. Most quantum chemistry packages use GTO, and often the solution for a more accurate MO is to use a linear combination of GTO to mimic one STO.

### 5.2.3 General Definitions

This section collects the laws and equations that will be used in deriving the data definitions, which in turn are used to build the instance models. [\[Some projects may not have any content for this section, but the section heading should be kept. —SS\]](#) [\[Modify the examples below for your problem, and add additional definitions as appropriate. —SS\]](#)

|             |   |
|-------------|---|
| Number      | GD1   |
| Label       | <b>Root-mean square deviation</b>   |
| SI Units    | Å   |
| Equation    | $RMSD_{ij} = \sqrt{\frac{1}{n_a} \sum_{k=1}^{n_a} ((x_{ki} - x_{kj})^2 + (y_{ki} - y_{kj})^2 + (z_{ki} - z_{kj})^2)}$   |
| Description | <p>The root-mean square deviation (<math>RMSD</math>) is an average distance between the atoms of conformer <math>i</math> and the atoms of conformer <math>j</math>. The inputs to this equation are the xyz coordinates for each atom (from both conformers) and the number of total atoms.</p> <p><math>n_a</math> number of atoms in each conformer</p> <p><math>x_{ki}</math> the x-coordinate of the <math>k^{th}</math> atom from conformer <math>i</math> (Å)</p> <p><math>y_{kj}</math> the y-coordinate of the <math>k^{th}</math> atom from conformer <math>j</math> (Å)</p> <p>etc.</p> |
| Source      |   |
| Ref. By     | DD2   |

#### 5.2.4 Data Definitions

This section collects and defines all the data needed to build the instance models. The dimension of each quantity is also given. [\[Modify the examples below for your problem, and add additional definitions as appropriate. —SS\]](#)

|             |  |
|-------------|--|
| Number      | DD1  |
| Label       | <b>Sum of conformer energies</b>   |
| Symbol      | $S_E$  |
| SI Units    | J  |
| Equation    | $\left  \sum_{i=1}^{G_n} E_i \right $  |
| Description | $E_i$ is the energy (J) of conformer $i$ , as calculated by solving the Schrödinger equation (T1), and $G_n$ is the number of conformers being simultaneously optimised by the system. The individual energies should be negative; the absolute value bars imply that, when this summation is included in the overall fitness function, we are maximizing the value of the fitness function. |
| Sources     | N/A  |
| Ref. By     | IM1  |
| Number      | DD2  |
| Label       | <b>Sum of root-mean square deviations</b>  |
| Symbol      | $S_{RMSD}$   |
| SI Units    | Å  |
| Equation    | $\sum_{i \neq j} RMSD_{ij}$  |
| Description | $RMSD_{ij}$ is the root-mean square deviation between conformers $i$ and $j$ (Å), as calculated by using the distance formula given in GD1. This distance represents how different two conformer geometries are from one another on average.   |
| Sources     | N/A  |
| Ref. By     | IM1  |

### 5.2.5 Instance Models

This section transforms the problem defined in Section 5.1 into one which is expressed in mathematical terms. It uses concrete symbols defined in Section 5.2.4 to replace the abstract symbols in the models identified in Sections 5.2.2 and 5.2.3.

The goal GS1 [reference your goals —SS] is solved by IM1 [reference your instance models —SS]. IM1 is an empirical function designed to explore the potential energy space for a molecule. Given a multi-variate surface that depends on dihedral angles, manipulate those dihedral angles and solve for  $Fit_G$ . This fitness function has two arbitrary coefficients,  $C_E$

and  $C_{RMSD}$  that the user must determine through experimentation with their molecule. [other details, with cross-references where appropriate. —SS] [Modify the examples below for your problem, and add additional models as appropriate. —SS]

|             |  |
|-------------|--|
| Number      | IM1  |
| Label       | <b>Fitness of conformer geometries</b> $Fit_G$   |
| Input       | $C_E, C_{RMSD}, S_E, S_{RMSD}$<br>The input is constrained so that $C_E > 0$ and $C_{RMSD} > 0$  |
| Output      | $Fit_G = C_E S_E + C_{RMSD} S_{RMSD}$<br>This equation is entirely empirical and its output does not have any physical meaning. The purpose of the equation is to represent the optimal set of conformers for a given input molecule.  |
| Description | $C_E$ is the energy coefficient water temperature (1/J).<br>$S_E$ is the absolute value of the sum of conformer energies (DD1) (J).<br>$C_{RMSD}$ is root-mean square deviation coefficient (1/Å).<br>$S_{RMSD}$ is sum of root-mean square deviations for each conformer (DD2) (Å).<br>The above equation applies when the number of conformers is greater than 1. If the number of conformers is exactly one, then the $RMSD$ terms should be set to zero. |
| Sources     | N/A  |
| Ref. By     | None   |

### 5.2.6 Data Constraints

Tables 2 and 4 show the data constraints on the input and output variables, respectively. The column for physical constraints gives the physical limitations on the range of values that can be taken by the variable. The column for software constraints restricts the range of inputs to reasonable values. The constraints are conservative, to give the user of the model the flexibility to experiment with unusual situations. The column of typical values is intended to provide a feel for a common scenario, but for this project the typical values are very specific to the input geometry. The uncertainty column provides an estimate of the confidence with which the physical quantities can be measured. This information would be part of the input if one were performing an uncertainty quantification exercise.

The specification parameters in Table 2 are listed in Table 3.

[There is no associated uncertainty with any of my values. Also, in most cases the typical values will be specific to the molecule (specifically, number of atoms). Should these two columns still be in this table? —JG]

For  $C_E$  and  $C_{RMSD}$ , the values for the coefficients depend on the shape of the potential energy surface. For example, in a surface where the minima are close together and the potential wells for the conformers are very high, placing more emphasis on  $C_E$  (i.e. making  $C_E$  bigger) would enable a better  $Fit_G$ . In the case where there are many low-lying conformers with wide energy basins, then the emphasis on  $C_{RMSD}$  would afford a better result. The number of conformers may not be known at the start of the program; the user may have to determine this value through experimentation with the code.

Table 2: Input Variables

| Var        | Physical Constraints            | Software Constraints          | Typical Value | Uncertainty |
|------------|---------------------------------|-------------------------------|---------------|-------------|
| $G_n$      | $n \geq 2^*   n \in \mathbb{Z}$ | $G_n \leq G_{max}$            | 2-5           | N/A         |
| $C_E$      | $C_E > 0$                       | $C_E > 0$                     | 0.5           | N/A         |
| $C_{RMSD}$ | $C_{RMSD} > 0$                  | $C_{RMSD} > 0$                | 0.5           | N/A         |
| $BS$       | available for molecule          | available in software package | cc-pVTZ       | N/A         |
| $QCM$      | available for molecule          | available in software package | CCSD          | N/A         |

(\*) if  $G_n$  is equal to one, then the  $Fit_G$  function should be changed to only consider the energy term and not the RMSD term. [you might need to add some notes or clarifications —SS]

Table 3: Specification Parameter Values

| Var  | Value |
|------|-------|
| None | -     |

### 5.2.7 Properties of a Correct Solution

A correct solution must exhibit [fill in the details —SS]

Table 4: Output Variables

| Var     | Physical Constraints |
|---------|----------------------|
| $Fit_G$ | $Fit_G > 0$          |

## 6 Requirements

This section provides the functional requirements, the business tasks that the software is expected to complete, and the nonfunctional requirements, the qualities that the software is expected to exhibit.

### 6.1 Functional Requirements

R1: Input the following parameters for the molecule whose conformers should be found. The molecule can be given as a SMILES string (example: C[N]1C=NC2=C1C(N(C(N2C)=O)C)=O), name (example: "caffeine"), xyz coordinates file, or z-matrix file. For example, the pubchem database can be queried with a name and is able to return 3D coordinates for that molecule. The other inputs that are required include the basis set and method; these inputs specify exactly how the Schrödinger Equation (T1) should be solved in order to get the energy. See Table 1 for an example of what these molecular specifications look like. [\[Requirements for the inputs that are supplied by the user. This information has to be explicit. —SS\]](#)

| symbol             | unit     | data type      | description   |
|--------------------|----------|----------------|---|
| $G_n$              | unitless | integer        | number of conformers (distinct geometries) to search for  |
| $C_E$              | J        | floating point | coefficient for energy term in fitness function   |
| $C_{RMSD}$         | m        | floating point | coefficient for RMSD term in fitness function   |
| BS                 | unitless | string         | basis set   |
| QCM                | unitless | string         | quantum chemical method   |
| molecular geometry | Å, °     | file, string   | name, SMILES string, xyz file, z-matrix file - a way to specify to connectivity of the input molecule |

R2: Given the inputs from R1, use IM1 to solve for  $Fit_G$ . Based on the initial geometry, generate a set of random dihedral angles  $D_i$ , whose size should be equal to  $n_G$ . [\[It](#)

isn't always required, but often echoing the inputs as part of the output is a good idea. —SS]

- R3: Calculate the energy  $E_i$  for each conformer and the RMSD distance between each conformer pair and use these values to solve for IM1. [Calculation related requirements. —SS]
- R4: Verify that the energy calculations converge (A1, A2) and that  $Fit_G$  satisfies the requirement in Table 4. [Verification related requirements. —SS]
- R5: Convert the dihedral angles  $D_i$  for each conformer back to a xyz geometry (using A8, A11). [Output related requirements. —SS]

## 6.2 Nonfunctional Requirements

[List your nonfunctional requirements. You may consider using a fit criterion to make them verifiable. —SS]

1. Given that the fitness function is empirical, Kaplan should be relatively robust with regards to changes made to the definition of  $Fit_G$ .
2. Maintainability is also important for other students who will use the project later-on.
3. The program should be parallelisable and capable of running on high-performance computing servers without a difficult install process (usability, portability).
4. The program should have some suggestions for inputs (example, default values for  $n_G$ ,  $C_E$ , and  $C_{RMSD}$ ) to help the user get started.
5. The program should work well with other quantum chemistry packages.

## 7 Likely Changes

- LC1: The assumption that the fitness function is linear with respect to energies and distances has not been verified; the function could be exponential, sinusoidal, etc. A5. [Give the likely changes, with a reference to the related assumption (aref), as appropriate. —SS]
- LC2: The assumption that the ordering of the atoms does not change the conformer space has not been verified A11.

## 8 Traceability Matrices and Graphs

The purpose of the traceability matrices is to provide easy references on what has to be additionally modified if a certain component is changed. Every time a component is changed, the items in the column of that component that are marked with an “X” may have to be modified as well.

[You will have to modify these tables for your problem. —SS]

|     | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 |
|-----|----|----|----|----|----|----|----|----|----|-----|-----|
| T1  | X  | X  |    |    |    |    |    |    | X  |     |     |
| DD1 |    |    |    | X  |    |    |    |    |    |     |     |
| DD2 |    |    |    | X  |    |    |    |    |    |     |     |
| GD1 |    |    |    |    |    |    |    |    |    |     |     |
| IM1 |    |    |    | X  | X  | X  | X  | X  |    | X   | X   |
| LC1 |    |    |    |    | X  |    |    |    |    |     |     |
| LC2 |    |    |    |    |    |    |    |    |    |     | X   |

Table 5: Traceability Matrix Showing the Connections Between Assumptions and Other Items

|     | T1 | DD1 | DD2 | GD1 | IM1 |
|-----|----|-----|-----|-----|-----|
| T1  |    | X   |     |     |     |
| DD1 | X  |     |     |     |     |
| DD2 |    |     |     | X   |     |
| GD1 |    |     |     |     |     |
| IM1 |    | X   | X   |     |     |

Table 6: Traceability Matrix Showing the Connections Between Items of Different Sections



|     |     |    |
|-----|-----|----|
|     | IM1 | R1 |
| IM1 |     |    |
| R1  |     |    |
| R2  | X   | X  |
| R3  | X   |    |
| R4  |     |    |
| R5  |     |    |

Table 7: Traceability Matrix Showing the Connections Between Requirements and Instance Models

## References

- W. Spencer Smith. Systematic development of requirements documentation for general purpose scientific computing software. In *Proceedings of the 14th IEEE International Requirements Engineering Conference, RE 2006*, pages 209–218, Minneapolis / St. Paul, Minnesota, 2006. URL <http://www.ifi.unizh.ch/req/events/RE06/>.
- W. Spencer Smith and Lei Lai. A new requirements template for scientific computing. In J. Ralyté, P. Ågerfalk, and N. Kraiem, editors, *Proceedings of the First International Workshop on Situational Requirements Engineering Processes – Methods, Techniques and Tools to Support Situation-Specific Requirements Engineering Processes, SREP’05*, pages 107–121, Paris, France, 2005. In conjunction with 13th IEEE International Requirements Engineering Conference.
- W. Spencer Smith, Lei Lai, and Ridha Khedri. Requirements analysis for engineering computation: A systematic approach for improving software reliability. *Reliable Computing, Special Issue on Reliable Engineering Computation*, 13(1):83–107, February 2007.
- W. Spencer Smith, John McCutchan, and Jacques Carette. Commonality analysis for a family of material models. Technical Report CAS-17-01-SS, McMaster University, Department of Computing and Software, 2017.
- Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, Inc., 1996. ISBN 0-486-69186-1.

## 9 Appendix

[Your report may require an appendix. For instance, this is a good point to show the values of the symbolic parameters introduced in the report. —SS]

|             |   |
|-------------|---|
| Number      | T2  |
| Label       | <b>Electronic Hamiltonian</b>   |
| Equation    | $\hat{H}_{ee} = \sum_{i=1}^N \frac{-\hbar^2}{2m_e} \nabla_i^2 + \sum_{i=1}^N \sum_{\alpha=1}^P \frac{-Z_{\alpha} q_e^2}{4\pi\epsilon_0  \vec{r}_i - \vec{R}_{\alpha} } + \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_e^2}{4\pi\epsilon_0  \vec{r}_i - \vec{r}_j }$  |
| Description | The above equation gives the electronic Hamiltonian operator for a molecule with N-e <sup>-</sup> and P-nuclei. $\hbar = \frac{h}{2\pi}$ is the reduced form of Plank's constant, $m_e$ is the mass of an electron, $\nabla$ is the gradient, $Z_{\alpha}$ is the atomic number for the given nuclei, $\epsilon_0$ is the permittivity of free space, $r_i$ are the xyz coordinates for the electrons, $R_{\alpha}$ are the xyz coordinates for the nuclei, and $q_e$ is the charge of an electron. |
| Ref. By     | T1  |

### 9.1 Symbolic Parameters

[The definition of the requirements will likely call for SYMBOLIC\_CONSTANTS. Their values are defined in this section for easy maintenance. —SS]