Kaplan Software Requirements Specification for Conformer Searching

Jen Garner

May 1, 2019

1 Revision History

Date	Version	Notes
October 4, 2018 (Thursday)	1.0	first draft for submission
October 22, 2018 (Monday)	1.1	enumerate non-functional requirements, add dictionary to Texstudio and fix spelling errors
October 27, 2018 (Saturday)	1.2	fix some typos and consistency errors from github issues
November 14, 2018 (Wednesday)	1.3	remove original wss comments, move numerical constants table to appendix, fix traceability issues
November 21, 2018 (Wednesday)	1.4	fix some issues from Github

2 Reference Material

Units, symbols, abbreviations, and acronyms are abbreviated here, as used in this document. Numerical constants are found in the appendix, Section 9.1.

2.1 Table of Units and Constants

This section describes the units that are used in the document. For each unit, the symbol is given followed by a description of the unit and its name. Some common constants are given.

Symbol	**Unit**	**Name**
m	length	metre
Å	length	$1\text{Å} = 1\text{x}10^{-10}\text{m}$
kg	mass	kilogram
s	time	second
J	energy	Joule
F	electrical capacitance	farad
C	electric charge	coulomb (A s)

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 $n_a - 3$
S_E	J	sum of conformer potential energies
C_E	$\frac{1}{\text{J}}$	coefficient for S_E in the fitness function
S_{RMSD}	Å	sum of root-mean-square distances for all conformer geometries
C_{RMSD}	$\frac{1}{\text{Å}}$	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
Ψ	unitless	wavefunction

σ	unitless	spin (α or β)
ϕ	unitless	atomic orbital
r	Å	position of electron in Cartesian coordinates
R	Å	position of nuclei in Cartesian coordinates
c_i	tbd	basis set constant
c_N	tbd	normalization/principle quantum number constant
ζ	tbd	effective nuclear charge constant
e^-	N/A	electron
x, y, z	Å	components of atomic Cartesian coordinates
X, Y, Z	Å	components of electronic Cartesian coordinates
Z_a	unitless	atomic number
i, j, k	N/A	indexing variables (see text for specific use case)

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	Program Name
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
VSEPR	valence shell electron pair repulsion
VB	valence bond
3D	three-dimensional

Contents

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.

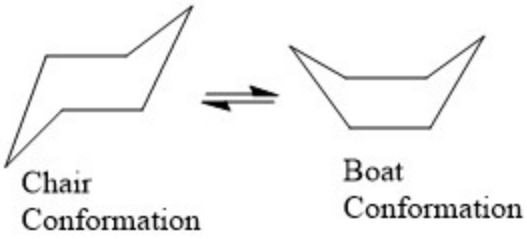

Figure 1: Two examples of conformational isomers of cyclohexane.

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. [I am confused by the difference between the global minimum and the local minima. Are you saying that there are times when the local minima is actually a better solution than the global one? If that is what you are saying, then I guess I'm not confused, but it goes against the usual rules of optimization. You are usually looking for the global minimum and trying to avoid getting stuck in a local minimum. —SS] 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.

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 [global minimum? — SS] 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); Smith et al. (2007). Section 1 is the revision history, including some updates as to how the document has changed over time. Section 2 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. Some constraints are also mentioned here for the design, and the user characteristics for the program are given.

Section 5 defines the problem to be solved, the terminology and definitions for the problem, the physical system that will be used to represent the problem, the goal statements, the characteristics of the solution, the models that are used to solve the problem, the assumptions about the problem, and the constraints on the output.

Section 6 lists the functional and non-functional requirements of the program. Section 7 gives the likely changes to the program, and Section 8 gives traceability tables showing how the portions of the document are connected. If changes are made to the document, then these traceability tables will be referenced to determine what other changes might be necessary in order to satisfy requirements.

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

Here the system context for Kaplan is shown (Figure 2). With respect to inputs, convergence conditions include 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 - third-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

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.

5.1 Problem Description

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. As calculated using a quantum mechanical approach, the energies of the conformers will be optimized when they have the lowest energy (largest negative value). The set of conformers will be optimized when the RMSD (root-mean-squared deviation) is maximized since this is a measure of how different the conformers are from one another. If the user specifies that they wish to obtain 3 conformers, the RMSD measure is necessary, otherwise the user could get 3 global minima of the potential energy space representing the exact same molecule.

5.1.1 Terminology and Definitions

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:

- **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. See Figure 3 for an example.
- **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.

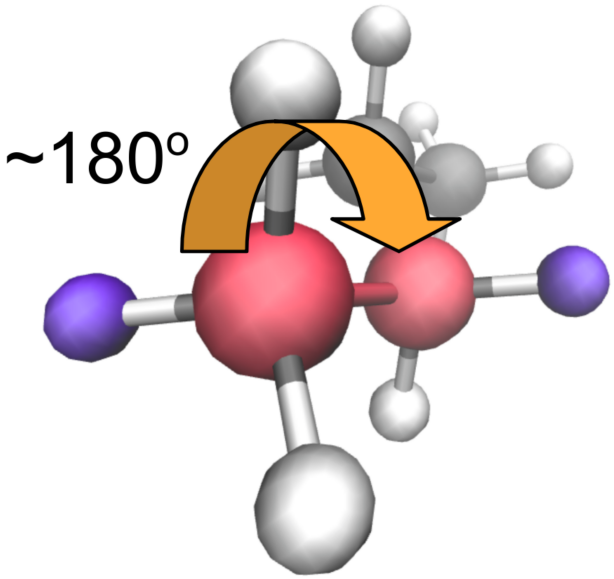

Figure 3: The dihedral angle formed between the 4 highlighted atoms is approximately 180°. Note that if the molecule was rotated such that the blue-highlighted hydrogen atom on the left went behind the blue-highlighted hydrogen atom on the right, then the dihedral angle would then be 0°.

An example of the physical system for the molecule butane (C_4H_{10}) can be found in Table 1. A visual for the molecule itself is displayed in Figure 3.

Z-Matrix	**XYZ (Cartesian) Coordinates**	**SMILES String**	**Dihedral 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.

5.1.3 Goal Statements

Given an input geometry and a list of convergence criteria for a molecule, the goal statement is:

GS1: Return n_G geometries that represent the maximum Fit_G value that was found after the optimization.

GS2: Return the energies of each geometry in the set that represents the maximum Fit_G value from the optimization.

[I'm confused by the output being a set. How many elements in the set? All of them? All of the ones the calculation can find? —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.

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 a linear combination of the sum of energies and the sum of RMSD values. [IM1][LC1]

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][LC2]

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.

Number	T1
Label	**Non-Relativistic Time-Independent Schrödinger Equation**
Equation	$\hat{H}\left\|\Psi\right\rangle = E\left\|\Psi\right\rangle$
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), Ψ 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 Couloumbic 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 (σ) of the electron (α or β) 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 $\Psi(r,\sigma) = \sum_i c_i \phi(r,\sigma)_i$, where $\phi(r,\sigma)_i$ is the ith AO and c_i is a constant described by the basis set.

There are two main types of AO in quantum chemistry - Slater-type orbitals (STO) and Gaussian-type orbitals (GTO). STO have the form: $\phi(r,\sigma) = c_N e^{-\zeta r}$, whereas GTO have the form: $\phi(r,\sigma) = c_N e^{-\zeta r^2}$, where c_N is a value that depends on normalization and the principle quantum number, ζ 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 ∞), 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.

Number	GD1
Label	**Root-mean square deviation**
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	The root-mean square deviation ($RMSD$) is an average distance between the atoms of conformer i and the atoms of conformer j. The inputs to this equation are the xyz coordinates for each atom (from both conformers) and the number of total atoms. n_a number of atoms in each conformer x_{ki} the x-coordinate of the k^{th} atom from conformer i (Å) y_{kj} the y-coordinate of the k^{th} atom from conformer j (Å) etc.
Source	
Ref. By	DD2

[Can GD1 be viewed as a refinement of T1? If so, this should be pointed out. If not, GD1 might make more sense as a data definition, or maybe as a theoretical model? —SS]

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.

Number	DD1
Label	**Sum of conformer energies**
Symbol	S_E
Units	J
Equation	$S_E = \left\lvert \sum_{i=1}^{n_G} E_i \right\rvert$
Description	E_i is the energy (J) of conformer i, as calculated by solving the Schrödinger equation (T1), and n_G 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	**Sum of root-mean square deviations**
Symbol	S_{RMSD}
Units	Å
Equation	$S_{\text{RMSD}} = \sum_{i \neq j} RMSD_{ij}$
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 is solved by IM1 . 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, Kaplan will 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.

Number	IM1
Label	**Fitness of conformer geometries** Fit_G
Input	C_E, C_{RMSD}, S_E, S_{RMSD} The input is constrained so that $C_E > 0$ and $C_{\text{RMSD}} > 0$
Output	$Fit_G = C_E S_E + C_{\text{RMSD}} S_{\text{RMSD}}$ 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 (1/J). S_E is the absolute value of the sum of conformer energies (DD1) (J). C_{RMSD} is root-mean square deviation coefficient (1/Å). S_{RMSD} is sum of root-mean square deviations for each conformer (DD2) (Å). The above equation applies when the number of conformers is greater than one. If the number of conformers is exactly one, then the $RMSD$ terms should be set to zero. [You mentioned multiple assumptions in the assumptions section that are relevant to this IM, but they do not actually appear in your description anywhere. All assumptions that are listed should be "invoked" somewhere in your documentation. —SS]
Sources	N/A
Ref. By	None

[I may have missed it, but I don't see an optimization anywhere. I thought I would see an objective function that you are minimizing? Something that will trace to your goal statement. I also thought that the output of the IM would be a set of conformers. —SS]

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
n_G	$n \geq 2^* \vert n \in \mathbb{Z}$	$n_G \leq n_{max}$	2-5	N/A
C_E	$C_E > 0$	$C_E > 0$	0.5	N/A
C_{RMSD}	$C_{\text{RMSD}} > 0$	$C_{\text{RMSD}} > 0$	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 n_G is equal to one, then the Fit_G function should be changed to only consider the energy term and not the RMSD term.

Table 3: Specification Parameter Values

Var	Value
None	-

5.2.7 Properties of a Correct Solution

This problem will return its "best guess" as to the most favoured conformations for the input molecule. The correct solution should return unique conformers, rather than n_G copies of

Table 4: Output Variables

Var	Physical Constraints
Fit_G	$Fit_G > 0$

the global minimum energy conformer. During the optimization, many potential solutions will be generated. A solution consists of n_G geometries, where each geometry G_i is generated from the dihedral angles list D_i and the initial geometry specification (which dictates bond angles and bond lengths). The energy should be calculated for each of these geometries. The best of these potential solutions exhibits the property that it has the biggest Fit_G value of all the sets of geometries currently available. The returned geometries should also converge when their energy calculations are performed.

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. See Table 1 for an example of what these molecular specifications look like.

symbol	unit	data type	description
n_G	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 the connectivity of the input molecule

R2: Given the inputs from R1, use IM1 to solve for Fit_G. Based on the initial geometry,

generate sets of random dihedral angles D_i, each with length of $n_a - 3$. The number of sets to generate should be equal to n_G.

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.

R4: Verify that the energy calculations converge (A1, A2) and that Fit_G satisfies the requirement in Table 4.

R5: Generate output geometries by combining the dihedral angles D_i for each conformer with the original geometric specifications (i.e. bond lengths, bond angles, and satisfying A8, A11).

R6: Program contains a set of template files that the user can modify to run their own calculations.

R7: The geometries for the conformers are optimized such that the returned value of Fit_G is maximized.

6.2 Nonfunctional Requirements

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 be easy to use and quick to explain to chemists.
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]

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.

	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	X
LC1					X						
LC2											X

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

	T1	DD1	DD2	GD1	IM1
T1					
DD1	X				
DD2				X	
GD1					
IM1		X	X		

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

	IM1	R1	Table 4	T1
IM1				
R1				
R2	X	X		
R3	X			
R4			X	X
R5	X			

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

References

W. Spencer Smith and Lei Lai. A new requirements template for scientific computing. In J. Ralyté, P. Àgerfalk, 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.

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

Number	T2
Label	**Electronic Hamiltonian**
Equation	$\hat{H_{ee}} = \sum_{i=1}^{N} \frac{-\hbar^2}{2m_e}\nabla_i^2 + \sum_{i=1}^{N}\sum_{\alpha=1}^{P} \frac{-Z_a q_e^2}{4\pi\epsilon_0 \lvert\overrightarrow{r_i} - \overrightarrow{R_\alpha}\rvert} + \sum_{i=1}^{N}\sum_{j=i+1}^{N} \frac{q_e^2}{4\pi\epsilon_0 \lvert\overrightarrow{r_i} - \overrightarrow{r_j}\rvert}$
Description	The above equation gives the electronic Hamiltonian operator for a molecule with N-e$^-$ and P-nuclei. $\hbar = \frac{h}{2\pi}$ is the reduced form of Plank's constant, m_e is the mass of an electron, ∇ is the gradient, Z_a is the atomic number for the given nuclei (a), ϵ_0 is the permittivity of free space, r_i are the xyz coordinates for the electrons, R_α are the xyz coordinates for the nuclei, and q_e is the charge of an electron.
Ref. By	T1

9.1 Symbolic Parameters

These symbols appear in the documentation and represent real numbers.

Symbol	Numerical Constant	Value
h	Planck's constant	6.62607004x10^{-34}J s
$\hbar$	reduced Planck's constant	$h/2\pi$J s
m_e	mass of an electron	9.10938x10^{-31}kg
q_e	charge of an electron	1.60217662x10^{-19}C
ϵ_0	permittivity of free space	8.854187817x10^{-12}F/m