Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project

Final Report

Calculating reduced partition function ratios of metal-ligand complexes

by

Devang Patel

devang.patel20@imperial.ac.uk

GitHub login: acse-dp1820

Supervisors:

Professor Dominik J Weiss

Dr James Percival

Professor Patricia Hunt

August 2021

Contents

Abstract

Isotopes such as carbon and oxygen are important for determining the paleoclimate, and heavier metal isotopes such as Zn are especially important for air quality and determining nutrient uptake in plants. Differences in isotopic ratios and fractionations are determined by calculating reduced partition function ratios, $1000\ln\beta$, using experimental observations, or computational methods, such as Density Functional Theory. The code introduced in this paper aims to automatically calculate the reduced partition function ratio and test the following hypothesis – the difference between the reduced partition function ratio using all vibrational modes and a secondary RPFR using only those with isotopic movement increases as the size of the molecule increases – by ascertaining the vibrational modes that cause isotopic movement. It is found that an increased difference occurs with increased molecular size, with the main differences of 0.28% for $Zn(H_2O)_6$ and 7.14% for $Zn(C_2H_8O_8)$.

Keywords: Gaussian 16, Python, reduced partition function ratio, isotopes, geochemistry, DFT

Introduction

Background

Isotopes are different atoms of an element that have a different mass number (number of neutrons). Isotopologues are molecules that contain different isotopes - $^{13}CO_2$ and $C^{18}O^{16}O$ are isotopologues of CO_2. Fractionation is the relative partitioning of isotopes in a chemical system/reaction, and is a quantum mechanical effect caused by differences in vibrational energies of molecules containing isotopes (Urey, 1947).

^{18}O and ^{13}C have been used to track the paleoclimate (Drever, 1982) for the past 5 million years (Lisiecki and Raymo, 2005). This has been accomplished by drilling into ice cores, as the fractionation of ^{18}O indicates the temperature of the climate when the snow fell (Lowe and Walker, 2014). Carbon isotopes are mainly fractionated by photosynthesis, and the change in the ratio within plants informs how favourable the environment was (Faure and Mensing, 2005), also allowing for reconstruction of past diets (Schwarcz and Schoeninger, 1991). ^{66}Zn plays a major role in studies of the current climate – specifically air studies. Anthropogenic contributions lead to large isotope enrichments (Gonzalez et al., 2016). They also pose a significant health risk, with soluble zinc in atmospheric dust leading to lung inflammation (Adamson et al., 2000), oxidative stress, and chronic diseases. (Prasad et al., 2004) The largest isotope fractionations for Zn occur in chemical exchange between species (Fujii et al., 2014, Fujii et al., 2011, Moynier et al., 2013). In plant geochemistry, it can be seen that Zn isotopes are used to determine plant uptake mechanisms due to roots having heavy zinc phosphates, and aerial parts having light zinc malates. (Fujii and Albarede, 2012) Deoxymugineic acid (DMA) has been proposed as a pathway for the transfer of zinc from soil to rice plants (Arnold et al., 2015, ARNOLD et al., 2010), with further developments on the mechanism and fractionation of Zn isotopes undertaken by Weiss et al. (2021).

Primarily experimental observations / calculations have been carried out, but in recent years, Density Functional Theory (DFT) has been used to calculate isotopic ratios of metals. This is due to there being comparatively less experimental information available for metal isotopes, which is why computational methods are necessary for predicting fractionation of metal isotope systems. Fractionation of aqueous Zn complexes and metal-ligand systems have been carried out (Black et al., 2011) for use in determining fractionation in seawater/chemical sediments, plant life and mammalian body fluids and organs (Fujii et al., 2014).

Mathematical Derivation of Reduced Partition Function Ratio

Table 1: Mathematical Constants and Definitions

Symbol	Definition	Value	Units
R	Molar gas constant	8.31	m^2 kg s^{-2} K^{-1} mol^{-1}
T	Temperature	-	K
n_i	^{i}th quantum state	0,1,2….	-
v_i	^{i}th vibrational frequency (or wavenumber)	-	Hz (or cm^{-1})
k_s	Force constant	-	N m^{-1}
μ	Reduced mass	-	kg
m_Y	Mass of species Y	-	kg
k	Boltzmann constant	1.38×10^{-23}	m^2 kg s^{-2} K^{-1}
M	Mass of molecule	-	kg
h	Planck's constant	6.63×10^{-34}	m^2 kg s^{-1}
I	Moment of inertia	-	kg m^2
c	Speed of light	3×10^8	m s^{-1}

The following mathematical background of isotopic fractionation is taken from Schauble (2004). Light isotopes, and related variables are denoted by a prime e.g. – X'.

The simplest isotopic exchange reaction between a diatomic molecule, AX' and the heavy isotope X, is shown below:

$$AX' + X \leftrightarrow AX + X'$$
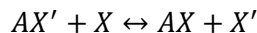
We define β_{AX} as a theoretical fractionation between the molecule AX, and non-interacting, dissociated atoms of X, also known as the reduced partition function ratio (RPFR). For the specific exchange reaction above, β is equal to the isotopic fractionation value, $\alpha_{AX-X'}$.

We see that the expression for β is also equivalent to the equilibrium constant expression, K_{eq}, which simplifies the derivation.

$$\alpha_{AX-X'} = \beta_{AX} = \frac{(X/X')_{AX}}{(X/X')_{X'}} = \frac{(AX/AX')}{(X/X')} = \frac{(AX)(X')}{(AX')(X)} = K_{eq}$$

The Gibbs free energy, ΔG, of a reaction at equilibrium is related to K_{eq} by the following:

$$\Delta G = -RTlnK_{eq}$$

Given the number of molecules on both sides are the same, the temperature is constant, and isotopic substitution has a minimal effect on molar volume, thus the Gibbs energy can be approximated by the Helmholtz free energy, F.

The equation to solve then becomes:

$$\beta = K_{eq} = e^{\frac{-\Delta F}{RT}}$$

To calculate the Helmholtz free energy, we must think about atomic motion. Atomic motion is split into 3 sections – translational, rotational and vibrational, each of which have discrete energies. Only vibrational quanta are significantly large enough for chemical reactions at ambient temperatures – for a visual depiction, see Appendix 1. The vibrational quantum energy levels, E, are evenly spaced: $E_i = (n_i + 0.5)h\nu_i$, with *i* being an individual energy state, and the vibrational frequency *ν_i*:

$$\nu_i = \frac{1}{2\pi}\sqrt{\frac{k_s}{\mu}}$$

where k_s is the force constant, and μ the reduced mass of the molecule: $\mu_{AX} = \frac{m_A m_x}{m_A + m_X}$.

From this, we see that vibrations are influenced by the mass of atoms – if the heavier X replaces X', the reduced mass will increase, and the vibrational frequency will decrease. The most basic case of this is in the totally symmetric (A_1) methane vibration, where only the hydrogen atoms are moving. If all the hydrogens were replaced by deuterium, this decreases the corresponding vibrational mode from 2914.2 cm^{-1} (34.86 kJ/mol) to 2084.7 cm^{-1} (24.94 kJ/mol) (Wilson et al., 1980).

For a molecule with N atoms, there are 3N degrees of freedom (comprised of translational, rotational and vibrational), as the position of the N nuclei depend on 3N coordinates. Translation is described by the movement of the molecule's centre of mass along the 3 Cartesian coordinates. Rotational degrees of freedom depend on the geometry of the molecule - a non-linear molecule can rotate freely about the 3 Cartesian axes, giving 3 degrees of freedom. However, a linear molecule can only rotate around 2 axes, as the molecular axis does not involve a change of nucleus movement, giving 2 degrees of freedom. Thus, the number of vibrations is equal to 3N-6 for a non-linear molecule, and 3N-5 for a linear molecule.

The total energy of a system is found out by using partition functions, represented by Q. Partition functions are thought of as a sum, over all molecular energy states, of the probability that the molecule will be in a particular energy state.

The Helmholtz energy is related to the partition function as follows:

$$F = -RTlnQ$$

The total partition function, Q, consists of 4 components, the 3 atomic motions (Q_T, Q_R, Q_V), and the electronic partition function (Q_E) – which is independent of isotopic substitution.

$$Q = Q_T \times Q_R \times Q_V \times Q_E$$

Substituting for Q, rearranging and ignoring the electronic contribution, we obtain:

$$F = -RTln(Q_T \times Q_R \times Q_V)$$

The equilibrium constant becomes:

$$K_{eq} = e^{\frac{-\Delta F}{RT}}$$

$$K_{eq} = e^{\sum_{products} \ln Q_T Q_R Q_V - \sum_{reactants} \ln Q_T Q_R Q_V}$$

$$K_{eq} = \frac{\prod_p Q_T Q_R Q_V}{\prod_r Q_T Q_R Q_V}$$

The expressions for the individual partition functions are as follows:

$$Q_T = V \left(\frac{2MkT}{h^2}\right)^{3/2}$$

$$Q_R = \frac{8\pi^2 I k T}{h^2}$$

$$Q_V = \prod_i \frac{e^{-\frac{hv_i}{2kT}}}{1 - e^{-\frac{hv_i}{kT}}}$$

where *i* is the number of vibrational frequencies of the molecule.

Substituting the above expressions into the equation for β, the constants in the rotational and translational components cancel out, hence, the reduced partition function ratio.

$$\beta = K_{eq} = \left(\frac{m_{X'}}{m_X}\right)^{\frac{3}{2}} \times \left(\frac{m_{AX}}{m_{AX'}}\right)^{\frac{3}{2}} \times \frac{I_{AX}}{I_{AX'}} \times \left[\prod_i \frac{e^{-\frac{hv_i}{2kT}}}{1 - e^{-\frac{hv_i}{kT}}} \times \frac{1 - e^{\left(-\frac{hv_i'}{kT}\right)}}{e^{\left(-\frac{hv_i'}{2kT}\right)}}\right]$$

Applying the Redlich-Teller product rule (Redlich (1935), Wilson et al. (1980)), which relates the vibrational frequencies of a molecule to its mass and moments of inertia:

$$\left(\frac{m_{X'}}{m_X}\right)^{\frac{3}{2}} \times \left(\frac{m_{AX}}{m_{AX'}}\right)^{\frac{3}{2}} \times \frac{I_{AX}}{I_{AX'}} = \frac{v_{AX}}{v_{AX'}}$$

The following expression for β incorporates site specific substitutions - n is the number of exchangeable atoms in the molecule. The following expressions use wavenumbers (cm^{-1}) so the speed of light (cm s^{-1}) is needed.

Substituting the following into the expressions for the vibrational partition function and the RPFR, now 1000*lnβ, as convention:

$$u_i^{(\prime)} = \frac{hcv_i^{(\prime)}}{kT}$$

Equation 1

$$Q_i^{(\prime)} = \frac{e^{-\frac{u_i^{(\prime)}}{2}}}{1 - e^{-u_i^{(\prime)}}}$$

Equation 2

$$\ln\beta = \left(\sum_i \ln\frac{v_{i,AX}}{v_{i,AX\prime}} + \frac{e^{-\frac{u_i}{2}}}{1 - e^{-u_i}} + \frac{1 - e^{-u_i'}}{e^{-\frac{u_i'}{2}}}\right)^{\frac{1}{n}}$$

Equation 3

$$1000\ln\beta = 1000 \times \left(\sum_i \ln\frac{v_i}{v_i'} + \ln Q_i - \ln Q_i'\right)^{\frac{1}{n}}$$

Equation 4

Thus, we see that β only depends on the vibrational frequencies of a molecule, which can be found computationally by carrying out electronic structure calculations. For a brief introduction to electronic structure theory, please see Appendix 2.

Introduction to Gaussian

Gaussian 16 (Frisch et al., 2016), is used to carry out electronic structure calculations, with GaussView used as a pre-processor to specify the geometry of the system. Deep within the code, the single electron Schrödinger equation is used to get energies for all the electrons within the system (see Appendix 2).

There are a couple of ways to calculate the energy of a molecule within Gaussian – sp (single point calculation) and opt (optimisation). The first takes the geometry provided from GaussView and calculates the energy, whereas optimisation uses the potential energy surface and attempts to find a minimum energy structure from the geometry provided. The energy is a function of the 3N-6 nuclear coordinates, q.

$$E = f(q_1 \ldots q_{3N-6})$$

The aim is to find points on the molecule's potential energy surface (PES) where the first derivative of the energy with respect to the nuclear coordinates, q, is 0. The first derivative is equal to the negative of the forces on the molecule.

$$\frac{\partial E(q)}{\partial q} = -F_q = 0$$

This is accomplished using gradient descent methods until the minimum is found. However, there are a variety of minima on a potential energy surface – for a reaction system, there may exist a saddle point, which corresponds to a transition state between reactants and products. Thus, the second derivatives of the energy must be calculated, to obtain the curvature of the PES around the minima. This is accomplished by carrying out frequency analysis (freq). The second derivatives are equal to the force constant, k_s, which can then be used to calculate the vibrational frequencies.

$$\frac{\partial E^2(q)}{\partial q^2} = k_s$$

If all the vibrational frequencies are real (positive), this confirms that the structure is at a minimum. If there are any imaginary (negative) frequencies, the structure is not at a minimum, but likely at a transition state, with the number of imaginary frequencies corresponding to the order of the transition state. (Foresman and Frisch, 2015). For a given imaginary frequency, distorting the geometry along the vibrational mode will lead to a lower energy. Given that the structures being optimised for this project are only molecules, and not full reaction systems, there are unlikely to be transition states.

Problem Description

Given that the reduced partition function ratio depends on vibrations of the isotopic atoms in question (Wilson et al., 1980), the following hypothesis has been proposed.

Hypothesis

The difference between 1000lnβ calculated using all vibrational modes, and modes that only involve movement of the isotopic atom, increases as the size of the molecule increases.

Code Objectives

The aim of the code introduced in this paper is to firstly automatically calculate the reduced partition function ratio, by extracting vibrational frequencies / wavenumbers from Gaussian 16 opt+freq log files (Fig. 1), and then ascertain the vibrational frequencies that contain isotopic movement, in order to calculate the secondary partition function ratio and test the hypothesis.

Left — ^{64}ZnCl$_4$ isotopologue

	1 E			2 E			3 T2		
Frequencies --	78.5452			78.5452			126.6746		
Red. masses --	34.9689			34.9689			38.8186		
Frc consts --	0.1271			0.1271			0.3670		
IR Inten --	0.0000			0.0000			9.0997		
Atom AN	X	Y	Z	X	Y	Z	X	Y	Z
1 30	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.35	-0.10
2 17	-0.11	0.28	0.40	0.39	0.29	-0.10	-0.37	-0.07	-0.26
3 17	0.11	-0.28	0.40	-0.39	-0.29	-0.10	-0.21	-0.23	0.35
4 17	0.11	0.28	-0.40	-0.39	0.29	0.10	0.22	-0.25	0.33
5 17	-0.11	-0.28	-0.40	0.39	-0.29	0.10	0.39	-0.09	-0.24

	4 T2			5 T2			6 A1		
Frequencies --	126.6746			126.6746			232.1478		
Red. masses --	38.8186			38.8186			34.9689		
Frc consts --	0.3670			0.3670			1.1104		
IR Inten --	9.0997			9.0997			0.0000		
Atom AN	X	Y	Z	X	Y	Z	X	Y	Z
1 30	-0.10	0.09	0.34	0.35	0.04	0.10	0.00	0.00	0.00
2 17	0.25	-0.24	-0.32	-0.11	-0.39	0.22	0.29	-0.29	0.29
3 17	-0.31	0.33	0.01	-0.28	-0.23	-0.30	-0.29	0.29	0.29
4 17	0.41	0.15	-0.16	-0.04	0.36	0.28	-0.29	-0.29	-0.29
5 17	-0.16	-0.41	-0.14	-0.21	0.19	-0.37	0.29	0.29	-0.29

	7 T2			8 T2			9 T2		
Frequencies --	232.2047			232.2047			232.2047		
Red. masses --	44.3617			44.3617			44.3617		
Frc consts --	1.4093			1.4093			1.4093		
IR Inten --	93.4247			93.4247			93.4247		
Atom AN	X	Y	Z	X	Y	Z	X	Y	Z
1 30	0.57	0.00	0.00	0.00	0.55	0.16	0.00	-0.16	0.55
2 17	-0.26	0.23	-0.23	0.15	-0.19	0.14	-0.28	0.29	-0.31
3 17	-0.26	0.22	0.22	0.28	-0.32	-0.29	0.15	-0.14	-0.19
4 17	-0.26	-0.23	-0.23	-0.28	-0.31	-0.29	-0.15	-0.14	-0.18
5 17	-0.26	-0.22	0.22	-0.15	-0.18	0.14	0.28	0.29	-0.32

Right — ^{66}ZnCl$_4$ isotopologue

	1 E			2 E			3 T2		
Frequencies --	78.5452			78.5452			126.2453		
Red. masses --	34.9689			34.9689			39.0504		
Frc consts --	0.1271			0.1271			0.3667		
IR Inten --	0.0000			0.0000			9.4990		
Atom AN	X	Y	Z	X	Y	Z	X	Y	Z
1 30	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	0.01	0.36
2 17	-0.17	0.24	0.41	0.37	0.33	-0.04	0.31	-0.30	-0.19
3 17	0.17	-0.24	0.41	-0.37	-0.33	-0.04	-0.30	0.32	-0.15
4 17	0.17	0.24	-0.41	-0.37	0.33	0.04	0.32	0.29	-0.18
5 17	-0.17	-0.24	-0.41	0.37	-0.33	0.04	-0.29	-0.32	-0.16

	4 T2			5 T2			6 T2		
Frequencies --	126.2453			126.2453			230.5599		
Red. masses --	39.0504			39.0504			44.5255		
Frc consts --	0.3667			0.3667			1.3945		
IR Inten --	9.4990			9.4990			92.2714		
Atom AN	X	Y	Z	X	Y	Z	X	Y	Z
1 30	-0.12	0.34	-0.01	0.34	0.12	0.01	-0.05	0.01	0.55
2 17	-0.24	-0.05	-0.39	-0.25	-0.36	0.18	-0.20	0.20	-0.24
3 17	-0.22	-0.07	0.40	-0.28	-0.34	-0.19	0.26	-0.25	-0.29
4 17	0.34	-0.27	0.19	-0.05	0.24	0.39	-0.21	-0.21	-0.25
5 17	0.36	-0.26	-0.18	-0.07	0.22	-0.40	0.24	0.24	-0.28

	7 T2			8 T2			9 A1		
Frequencies --	230.5599			230.5599			232.1478		
Red. masses --	44.5255			44.5255			34.9689		
Frc consts --	1.3945			1.3945			1.1104		
IR Inten --	92.2714			92.2714			0.0000		
Atom AN	X	Y	Z	X	Y	Z	X	Y	Z
1 30	0.55	0.05	0.05	-0.05	0.55	-0.02	0.00	0.00	0.00
2 17	-0.26	0.22	-0.23	0.26	-0.29	0.26	0.29	-0.29	0.29
3 17	-0.22	0.18	0.18	0.24	-0.27	-0.24	-0.29	0.29	0.29
4 17	-0.30	-0.27	-0.27	-0.20	-0.23	-0.20	-0.29	-0.29	-0.29
5 17	-0.26	-0.23	0.23	-0.21	-0.25	0.21	0.29	0.29	-0.29

Figure 1: Frequencies from Gaussian log files - left - ^{64}ZnCl$_4$ isotopologue, right - ^{66}ZnCl$_4$ isotopologue. Each numbered frequency has a symmetry label (E, T2, A1), the frequency (in cm^{-1}), the reduced mass of the molecule (AMU), force constants (mDyne/Angstrom), IR intensities (KM/Mole) and normal atomic coordinates. The coordinates indicate movement of the atoms in the frequency.

Constants							-0.03	2.70	2.73		2.49 = 1000 *ln(beta)
h	6.63E-34		Number	Light'	Heavy	u'	u	ln(h/l)	lnQ'	lnQ	ln beta
c*100	3E+10		1	78.5452	78.5452	0.38	0.38	0.00E+00	9.64E-01	9.64E-01	0.00E+00
k	1.38E-23		2	78.5452	78.5452	0.38	0.38	0.00E+00	9.64E-01	9.64E-01	0.00E+00
			3	126.6746	126.2453	0.61	0.61	-3.39E-03	4.77E-01	4.80E-01	1.05E-04
Temperature	298.15		4	126.6746	126.2453	0.61	0.61	-3.39E-03	4.77E-01	4.80E-01	1.05E-04
			5	126.6746	126.2453	0.61	0.61	-3.39E-03	4.77E-01	4.80E-01	1.05E-04
Number of atoms	5		6	232.1478	230.5599	1.12	1.11	-6.86E-03	-1.65E-01	-1.58E-01	6.99E-04
Number of frequencies (non linear)	9		7	232.2047	230.5599	1.12	1.11	-7.11E-03	-1.66E-01	-1.58E-01	7.24E-04
			8	232.2047	230.5599	1.12	1.11	-7.11E-03	-1.66E-01	-1.58E-01	7.24E-04
Author: Devang Patel			9	232.2047	232.1478	1.12	1.12	-2.45E-04	-1.66E-01	-1.65E-01	2.51E-05
GitHub login: acse-dp1820											

Figure 2: Image of worksheet used to calculate the RPFR for ZnCl4 (calculations.xlsx). The Light' column are the frequencies of ^{64}ZnCl$_4$, and the Heavy column is the frequencies of ^{66}ZnCl$_4$. The u' and u columns correspond to Equation 1, with the next 4 columns being the natural log of the isotope ratio, and the natural log of the light and heavy partition functions (Equation 2) and the individual ln beta calculation (Equation 3). The individual contributions are summed together and multiplied by 1000 (Equation 4) to give the final value of 2.49. We can see that the first 2 frequencies are identical for both isotopologues, and this is due to no movement of the Zn isotope, only the chlorine ligands.

This is relatively simple, and takes a short amount of time, as ZnCl$_4$ has 9 vibrational modes, but for significantly complex molecules, the task of manually inputting the frequencies becomes tedious, and thus, a computational solution was sought out to accurately calculate the reduced partition function ratio without the need of a spreadsheet.

In Figure 1, we can see that each atom has a set of Cartesian coordinates for each mode. By identifying the isotope and taking the absolute sum of the relevant coordinates, we gain an estimate on how much the isotope moves for that vibration. For example, to calculate the approximate movement of the isotope for the 3rd vibration for ^{64}ZnCl$_4$, the coordinates to be added together are: -0.01, 0.35 and -0.10 giving a value of 0.24. If this estimate is greater than 0, the vibration involves isotopic movement. This provides further insight into how much the

movement of the isotope contributes to the RPFR of a given molecule – if the isotope moves a significant amount, the contribution of that vibration to the RPFR will be greater.

Software Description

CODE METADATA

CODE VERSION / REPOSITORY	https://github.com/acse-2020/acse2020-acse9-finalreport-acse-dp1820
LEGAL LICENSE	MIT
VERSION CONTROL SYSTEM	git
OPERATING SYSTEM	Windows, Linux (Ubuntu)
DEPENDENCY LIST	https://github.com/acse-2020/acse2020-acse9-finalreport-acse-dp1820/blob/master/requirements.txt

Python was used as the programming language due to its relative ease of understanding, ability to scrape through data files, and use for numerical operations – the NumPy library can perform operations on each element of an array, and hence is an efficient way of performing calculations on large arrays, such as frequency arrays. A compiled language like C/C++ would have been too high level for the operations required.

To extract information from log files, regular expressions were implemented. Regular expressions are a sequence of characters that allow you to match strings, and thus are useful when extracting data.

The list of extracted quantities from a log file is as follows:

- List of vibrational frequencies
- Temperature of simulation
- Isotopic information
- Item convergence table
- List of low frequencies
- Coordinate lists for frequency contribution analysis

See Appendix 3 for the list of regular expressions.

The software was first and foremost designed with the user in mind, as the program only needs the required inputs, and a command line interface shows what is occurring during each step of the program (see "Description of main program – script.py"). The user can also run a testing suite, *tests.py*, which tests the main functionalities - frequency extraction and calculation codes - against manually determined RPFR values in *calculations.xlsx* (see "Testing").

Module Breakdown

The code itself is split into 4 submodules: *file_io*, *extractions*, *calculations* and *isotope_contribution*, with the following hierarchy.
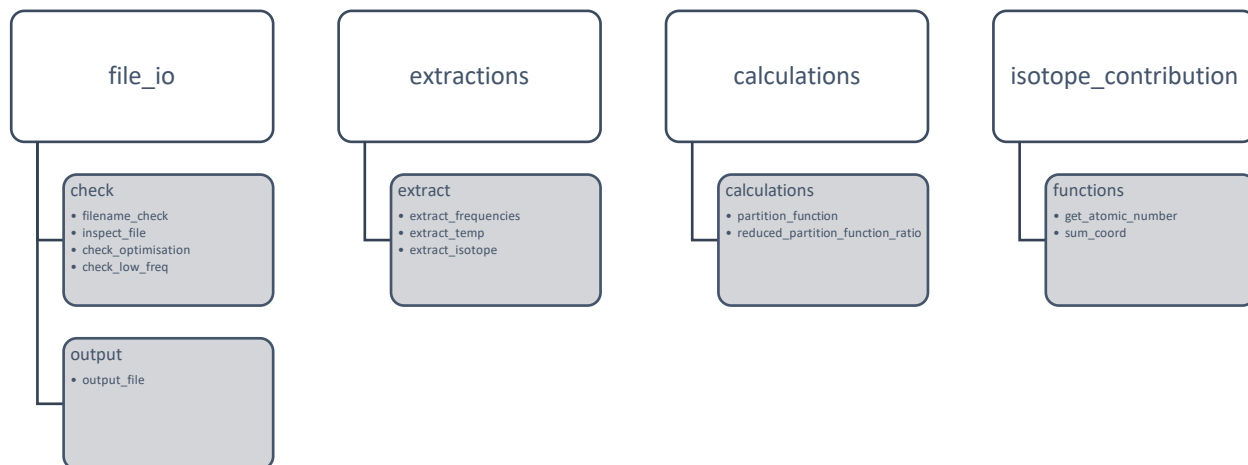
Figure 3: Module Hierarchy

Module 1: file_io

This module contains 2 submodules: *check* and *output*.

1.1 check

This submodule contains the following functions:

1.1.1 filename_check(filename)

This function checks if the filename provided exists and exits the program if it does not.

1.1.2 inspect_file(filename, pattern)

This function checks for the provided regex pattern within the file. The regex patterns are used by other functions, so an initial check through the file is necessary. A Boolean is returned, indicating the presence/absence of the regex pattern.

1.1.3 check_optimisation(filename)

This function checks for the last occurrence of the item convergence table, extracts and returns it. This function also informs the user via the terminal output if the forces + displacements have converged – if there are any values that have not converged, the user is asked if they wish to continue, and exits if not.

1.1.4 check_low_freq(filename)

This function extracts the low frequencies from the log file, returns them in a list, and checks if they meet a threshold of +/- 30 cm^{-1}. If they exceed the threshold, a warning is printed to the terminal.

1.2 output

1.2.1 output_file(filename)

This function checks if the output file exists and creates any required directories.

Module 2: extractions

2.1: extract

This module consists of the major extraction code – extracting the frequencies, temperature and isotopic information.

2.1.1 *extract_frequencies(filename, linear_check)*

This function firstly extracts the number of atoms in the molecule and uses it to calculate the expected number of frequencies - this distinction is made by the *linear_check*.

Once the expected number has been calculated, the frequencies are extracted, and stored in a NumPy array. If the length of the frequency list and the number of expected frequencies does not match, the program exits.

2.1.2 *extract_temp(filename)*

This function extracts and returns the temperature value from the log file.

2.1.3 *extract_isotope(filename)*

Extracts the element name and the isotope number from the log file.

Module 3: calculations

3.1 calculations

This module contains 2 functions:

3.1.1 *partition_function(array, temp)*

Using an array of vibrational frequencies, this function calculates the sum of the natural log of the vibrational partition function (Equations 1 and 2). Constants were imported from *scipy.constants*. The array is converted to a NumPy array, so the elementwise operation functionality can be used to quickly carry out calculations on the array itself. This function is only directly called in *reduced_partition_function_ratio.*

3.1.2 *reduced_partition_function_ratio(light_freq, heavy_freq, temp)*

This function carries out the main calculation of the RPFR and returns it as 1000lnβ (Equation 4). First, the lengths of the light isotopologue and heavy isotope frequency arrays are compared – if they are not the same, the program exits, informing the user that the lengths do not match. The ratio of the heavy and light frequencies is calculated, and the length of the ratio array is checked against both frequency arrays, and their sum is taken. Next, the vibrational partition functions are calculated by calling *partition_function* on both sets of frequencies. Finally, the RPFR is calculated and returned.

Module 4: isotope_contribution

4.1 functions

This module ascertains, for each individual frequency, the approximate movement of the isotope, by summing over the Cartesian coordinates in an individual frequency's atomic matrix. Each individual line of the distance matrix is identified by atomic number, so *atomic_number.csv*

is used as a lookup table to confirm the atomic number is present (*get_atomic_number)*, and the sum of coordinates extracted (*sum_coord).*

atomic_number.csv is a simple csv file with 2 columns, "Element" and "Atomic Number".

4.1.1 get_atomic_number(element)

The atomic number is extracted from *atomic_number.csv* by selecting the row that corresponds to the element extracted from the log file if it is in the csv file, and then indexing the "Atomic Number" column.

4.1.2 sum_coord(filename, atomic_number)

First, the atomic distance matrices are found, and then using the atomic number obtained from *get_atomic_number*, the coordinate line is extracted, and the index and atomic number removed. This is done for all coordinate lines to get a list of isotope coordinates. Once this list is obtained for all frequencies, the list separated into chunks of 3 to get the isotope's Cartesian coordinates for a single frequency. This list is then iterated over in steps of 3, summing the coordinates in each chunk together, to finally return a list of the extent of isotopic movement for each frequency.

Description of main program – script.py

The modules described above all come together within the main program, script.py. The program takes in 5 command line arguments:

- Paths to the 2 input files
- Linear check – to determine the expected number of frequencies
- Path to the output file
- Print variable – determines the level of information saved to the output file
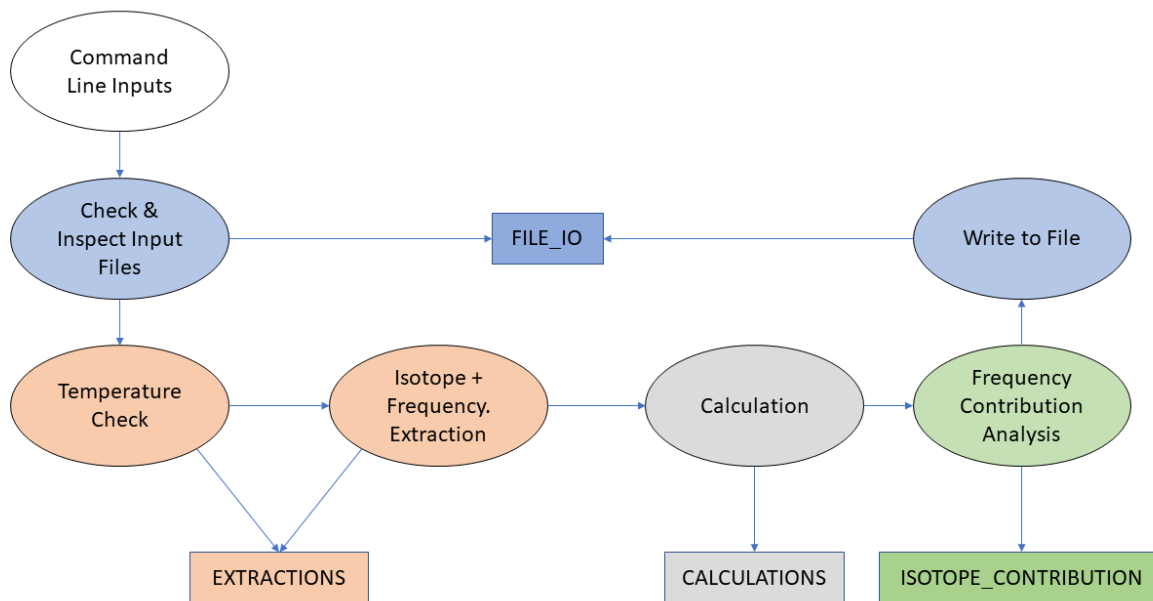

Figure 3: Basic process of script.py and the modules associated with each section

Command Line Arguments

Initially, the program checks that the correct number of command line arguments has been passed, and if not, a SystemExit is raised, with a message indicating correct usage of the program. If correct, the command line arguments are stored in their respective variables.

Check & Inspect Input Files

Firstly, *filename_check* is called on both the light and heavy isotopologue files. Next, *inspect_file* checks the presence of temperature, isotopes, optimisation/convergence, low frequency and frequency information. Any further functions will only run if the inspections pass (i.e., the Boolean variable returned is True).

Temperature Check

If the light and heavy temperature inspections pass, *extract_temp* is called for both input files, and the values returned compared to each other. If they are the same, the temperature variable used in the calculations is set to the temperature in the light log file. If the temperatures are not the same, or not found at all, the program exits, as the RPFR is temperature dependent.

In practice, the 2 inputs would not have different temperatures, so this serves as an initial comparison.

Isotope and Frequency Extraction

Extractions are carried out for the light and then heavy isotope file. The required inspections are isotope information (*extract_isotope)*, convergence information (*check_optimisation)* and frequency information (*check_low_freq, extract_frequencies*). However, if the frequencies are not found, the user is asked if they wish to continue, and if not, the program exits.

Calculation

The calculation requires the existence of both frequency sets, and a temperature value. If either of the three requirements are not present, the calculation will not be carried out.

Frequency Contribution Analysis

The atomic number is extracted and stored, and the extent of isotopic movement is calculated for all frequencies in both log files, with the lengths of the coordinate sum arrays checked against the length of the respective arrays. If the lengths are not equal, an error is raised.

Next, the frequencies that contain isotopic movement are isolated from the initial frequency array and stored in their own list. If the number of frequencies that have isotopic movement are the same for both isotopologues, a second RPFR calculation is carried out. This gives a brief insight into how the frequencies with isotopic movement contribute to the RPFR value for individual molecules.

Write to File

First, the output file path is checked to see if the file already exists, and any necessary directories created if not.

The level of information that will be saved to the output file is shown to the user.

Basic Information:

- Input file paths
- If inspections pass:
  - Isotopic information
  - Temperature of simulation
  - RPFR for all frequencies
  - RPFR for selected frequencies with isotopic movement, if applicable
  - Frequency lists – full and isotopic movement

Full Information - as above, but including the following:

- Contributions to RPFR
- Full isotope contribution arrays
- Item convergence tables
- Low frequencies

Testing

Testing the software consisted of 2 sections – testing the calculation and extraction code – seen in *tests.py,* allowing the user to confirm the validity of the code, and testing the functionality of *script.py*, detailed in the *tests/* folder, allowing for testing of the program with a variety of input files.

Testing calculation and extraction code - tests.py

This python script makes use of *reduced_partition_function_ratio, extract_frequencies* and *extract_temp*, as they implement the basic functionality of the program. The testing code is setup as a module and a script, by using 'if __name__ == "__main__":'

Functions:

1.1 test_extraction(l_filename, h_filename, linear_check, m_light_freq, m_heavy_freq)

> This function compares the frequencies extracted using *extract_frequencies* (*l_filename, h_filename, linear_check)* against the manually extracted frequency lists (*m_light_freq, m_heavy_freq)* and prints whether the values are identical.

1.2 test_calculation(l_filename, h_filename, linear_check, m_beta, m_ratio, m_Qlight, m_Qheavy)

> This function tests the RPFR returned by *reduced_partition_function_ratio* against the manually calculated values (*m_beta, m_ratio, m_Qlight, m_Qheavy)* and informs the user if the values agree with each other with an absolute tolerance of 10^{-8}.

test_extraction is run with ZnCl4 only, as there are only 9 frequencies to extract, and the calculation code is tested for both ZnCl4 and $Zn(H_2O)_6$ log files.

Testing functionality of script.py

The tests folder contains a *log_files/* folder. There are 3 subfolders for testing script.py – *linear*, *water* and *zinc*. The *outputs/* subfolder within contains both the output files that script.py generate – "CO2.txt", and the associated command line outputs (what the user sees when the program is running) - "CO2_cmd.txt". Any user inputs were disabled when running functionality tests.

1. Linear

This folder contains 2 carbon dioxide log files to illustrate the simple running of script.py with a linear molecule, as most of the molecules used during the development of the program are non-linear. The rationale behind this was to confirm that the code to check the number of frequencies, and the *linear_check* variable performed correctly.

2. Water

This folder contains most of the testing functionality, as optimisation of a H_2O molecule takes about 30 seconds. The isotopes used were ^{16}O and ^{18}O.

2.1. Convergence

> This folder contains 2 genuine log files and 2 log files that have had their convergence tables removed. This will not occur normally, as the item convergence table is always present within Gaussian log files. The associated tests check the functionality of script.py if convergence tables were not present.

If no convergence tables are present, the inspection for convergence fails, and the script informs the user that the convergence information is unavailable and cannot extract and write them to the output file.

2.2. Frequency

The log files here consist of 2 exemplary log files and 2 log files that just contain optimisation information.

The program was tested on 4 permutations of log files: no frequency information, light frequency only, heavy frequency only, and both sets of frequencies.

For the complete lack of frequency information, the script ran in its entirety, illustrating that only the convergence tables and isotopic information is written to the output file.

For the instances where one set of frequencies exist, the program exits if frequencies weren't found in either file.

2.3. Isotopes

This folder contains log files that do not have any isotopic information specified – by default, the most abundant isotopes are assumed.

The program was tested on complete lack of isotopic information, light isotope, heavy isotope only, and both isotopes specified.

When testing the complete lack of isotopes, the molecules are identical, and thus the RPFR calculated as 0, and no isotopic information was extracted. This also occurred for specifying the light isotope, but the output file correctly stated that the light isotope, ^{16}O was extracted from the log file.

A similar situation occurred for the other 2 tests, with the RPFR having the same value of 63.8.

2.4. Temperature

This has log files that test the functionality of the program if two log files with different temperatures are chosen as input files. Optimisation and frequency analysis simulations were run on a water molecule at two different temperatures, i.e., 298.15K and 300K. Given that the calculation of the RPFR is temperature dependent, the program should exit, and does so.

3. Zinc

$ZnCl_4$ and $Zn(H_2O)_6$ log files are used in *tests.py*. The $ZnCl_2$ log files are used to illustrate what occurs when input files that do not contain any optimisation / frequency analysis information - which is unlikely in practice - is passed to the program. In this testing case all that occurs is the output file simply states the input files used to generate it, with no other information present.

Results + Discussion

Results

The molecules in Table 2 were chosen as they have a widespread molecular size, enabling the hypothesis to be tested properly.

Table 2: Difference in manual and computationally calculated RPFR's

Molecule	# Atoms (# Freq)	Manual RPFR	Computational RPFR	Absolute Difference (10^{-13})
CO_2	3 (4)	177.36	177.36	0.14
H_2O	3 (3)	63.80	63.80	2.70
$ZnCl_4$	5 (9)	2.49	2.49	0.53
$ZnSO_4$	18 (48)	4.08	4.08	6.13
$Zn(H_2O)_6$	19 (51)	3.90	3.90	12.05
Zinc oxalate - $Zn(C_2H_8O_8)$	19 (51)	4.87	4.87	7.36
ZnDMA	39 (111)	4.12	4.12	23.30

Table 3 shows the difference between the full RPFR and the secondary RPFR, for the isotopologues that had the same number of frequencies with isotopic movement.

Table 3: Difference in RPFR using full / selected frequencies

Molecule	Full RPFR (# of atoms, # of freq)	RPFR using selected frequencies (# of freq)	Absolute % difference
CO_2	177.36 (3, 4)	177.36 (3)	0
H_2O	63.80 (3, 3)	63.80 (3)	0
$ZnCl_4$	2.49 (5, 9)	2.49 (6)	2.21×10^{-11} (≈ 0)
$Zn(H_2O)_6$	3.90 (19, 51)	3.89 (15)	0.28
$Zn(C_2H_8O_8)$	4.87 (19, 51)	4.52 (26)	7.14

Testing the hypothesis further by calculating the secondary partition function ratio for the following alkanes - methane, ethane, butane, octane - with the same carbon atom substituted for ^{13}C – was attempted, but only values for methane and ethane were obtained, which were identical to the full RPFR.

Discussion

Each table addresses one of the code objectives – automatic calculation of the reduced partition function and testing the hypothesis.

In Table 2, we can see that there are minute differences between the manual and computationally calculated partition function ratios, on the order of 10^{-13} – the differences are so minute that the manual and computational values can be considered equal.

Table 3 compares the full and secondary partition function ratios – for small molecules there is no difference between the two, but for $Zn(H_2O)_6$ and $Zn(C_2H_8O_8)$, there are differences of 0.28%

and 7.14% respectively. This confirms the hypothesis, as the greater the number of atoms in the molecule, the more vibrations that contribute to the partition function ratio.

Code Evaluation

The above results confirm that the code does automatically calculate the reduced partition function ratio from Gaussian 16 log files, thereby achieving the goal of reducing time spent on calculations and allowing geochemists to focus on analysing what the RPFR means in context.

Using the function, *inspect_file*, and encompassing the rest of the program within if / else statements is arguably one of the most important aspects of the code design, as it ensures that the program can only run if the necessary parts of the input files are present, mitigating any input file errors that may occur.

On the other hand, the method used to calculate the secondary reduced partition function ratio, whilst useful, is not the best. This is because the main criterion for calculation is the same number of frequencies that cause isotopic movement, as the reduced partition function calculates the ratios of heavy and light frequencies. This was highlighted as an issue when attempting to run the program to calculate the reduced partition function ratio for alkanes. A better way to compare the full and secondary RPFR's would be to remove the ratio of frequencies and only look at the partition function difference.

When testing the hypothesis using alkanes, the *sum_coord()* function proved to be problematic, due to the specificity of the function. The extracted frequency arrays and the frequency contribution arrays were different lengths, as *sum_coord()* was extracting the coordinates for all carbon atoms in the molecule, not just the ^{13}C isotope.

Future Work

The code as written currently serves a very simple purpose, calculating reduced partition function ratios for isotopologues with a single isotope. A natural progression would be to address the *sum_coord()* issue described above and further adapt the code for more than 1 isotope within a molecule.

The code can also be utilised further to ascertain if there are atomic / molecular controls that impact the reduced partition function ratio. Gaussian 16 provides information such as atomic angles/distances, dipole moments and atomic charges and the necessary functions can be added to extract these.

Furthermore, Gaussian 16 has an interface, gauopen, which allows Python to interface with Gaussian and extract force constants and reduced masses to calculate the frequencies, as well as other quantities - removing the necessity for a simulation log file.

Introducing more frequency analysis, such as calculating the RPFR using subsets of frequencies that have the same symmetries and/or only frequencies with wavenumbers below a cut-off value, is also worthwhile.

The output file can be further developed from a plain text file, by generating a heatmap of the molecule and indicating frequency contributions visually.

Conclusion

The Python code introduced in this report automatically calculates the reduced partition function, $1000\ln\beta$, using frequencies extracted from Gaussian 16 log files, thus removing the need for manual calculations. A secondary RPFR using only frequencies that cause isotopic movement was also calculated, and the hypothesis of an increased difference between full and secondary ratios as molecular size increases was tested and confirmed, with the greatest difference being 7.14% for the $Zn(C_2H_8O_8)$ species.

Bibliography

ADAMSON, I. Y. R., PRIEDITIS, H., HEDGECOCK, C. & VINCENT, R. 2000. Zinc Is the Toxic Factor in the Lung Response to an Atmospheric Particulate Sample. *Toxicology and Applied Pharmacology,* 166**,** 111-119.

ARNOLD, T., KIRK, G. J. D., WISSUWA, M., FREI, M., ZHAO, F.-J., MASON, T. F. D. & WEISS, D. J. 2010. Evidence for the mechanisms of zinc uptake by rice using isotope fractionation. *Plant, Cell & Environment,* 33**,** 370-381.

ARNOLD, T., MARKOVIC, T., KIRK, G. J. D., SCHÖNBÄCHLER, M., REHKÄMPER, M., ZHAO, F. J. & WEISS, D. J. 2015. Iron and zinc isotope fractionation during uptake and translocation in rice (Oryza sativa) grown in oxic and anoxic soils. *Comptes Rendus Geoscience,* 347**,** 397-404.

BLACK, J. R., KAVNER, A. & SCHAUBLE, E. A. 2011. Calculation of equilibrium stable isotope partition function ratios for aqueous zinc complexes and metallic zinc. *Geochimica Et Cosmochimica Acta,* 75**,** 769-783.

BORN, M. & OPPENHEIMER, R. 1927. Zur Quantentheorie der Molekeln. *Annalen der Physik,* 389**,** 457-484.

BYTAUTAS, L. & RUEDENBERG, K. 2004. Correlation energy extrapolation by intrinsic scaling. II. The water and the nitrogen molecule. *The Journal of Chemical Physics,* 121**,** 10919-10934.

DREVER, J. I. 1982. The geochemistry of natural waters.

FAURE, G. & MENSING, T. M. 2005. *Principles and applications*, John Wiley & Sons, Inc.

FORESMAN, J. & FRISCH, A. 2015. Exploring Chemistry with Electronic Structure Methods, 3rd edn., Wallingford, CT USA, Gaussian. Inc.

FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., SCALMANI, G., BARONE, V., PETERSSON, G. A., NAKATSUJI, H., LI, X., CARICATO, M., MARENICH, A. V., BLOINO, J., JANESKO, B. G., GOMPERTS, R., MENNUCCI, B., HRATCHIAN, H. P., ORTIZ, J. V., IZMAYLOV, A. F., SONNENBERG, J. L., WILLIAMS, DING, F., LIPPARINI, F., EGIDI, F., GOINGS, J., PENG, B., PETRONE, A., HENDERSON, T., RANASINGHE, D., ZAKRZEWSKI, V. G., GAO, J., REGA, N., ZHENG, G., LIANG, W., HADA, M., EHARA, M., TOYOTA, K., FUKUDA, R., HASEGAWA, J., ISHIDA, M., NAKAJIMA, T., HONDA, Y., KITAO, O., NAKAI, H., VREVEN, T., THROSSELL, K., MONTGOMERY JR., J. A., PERALTA, J. E., OGLIARO, F., BEARPARK, M. J., HEYD, J. J., BROTHERS, E. N., KUDIN, K. N., STAROVEROV, V. N., KEITH, T. A., KOBAYASHI, R., NORMAND, J., RAGHAVACHARI, K., RENDELL, A. P., BURANT, J. C., IYENGAR, S. S., TOMASI, J., COSSI, M., MILLAM, J. M., KLENE, M., ADAMO, C., CAMMI, R., OCHTERSKI, J. W., MARTIN, R. L., MOROKUMA, K., FARKAS, O., FORESMAN, J. B. & FOX, D. J. 2016. Gaussian 16 Rev. C.01. Wallingford, CT.

FUJII, T. & ALBAREDE, F. 2012. Ab initio calculation of the Zn isotope effect in phosphates, citrates, and malates and applications to plants and soil. *PLoS One,* 7**,** e30726.

FUJII, T., MOYNIER, F., BLICHERT-TOFT, J. & ALBAREDE, F. 2014. Density functional theory estimation of isotope fractionation of Fe, Ni, Cu, and Zn among species relevant to geochemical and biological environments. *Geochimica Et Cosmochimica Acta,* 140**,** 553-576.

FUJII, T., MOYNIER, F., PONS, M.-L. & ALBARÈDE, F. 2011. The origin of Zn isotope fractionation in sulfides. *Geochimica et Cosmochimica Acta,* 75**,** 7632-7643.

GONZALEZ, R. O., STREKOPYTOV, S., AMATO, F., QUEROL, X., RECHE, C. & WEISS, D. 2016. New Insights from Zinc and Copper Isotopic Compositions into the Sources of Atmospheric Particulate Matter from Two Major European Cities. *Environmental Science & Technology,* 50**,** 9816-9824.

HARTREE, D. R. 1928. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Mathematical Proceedings of the Cambridge Philosophical Society,* 24**,** 89-110.

HOHENBERG, P. & KOHN, W. 1964. Inhomogeneous Electron Gas. *Physical Review,* 136**,** B864-B871.

JR., C. W. B. & TAYLOR, P. R. 1986. Benchmark full configuration-interaction calculations on H2O, F, and F−. *The Journal of Chemical Physics,* 85**,** 2779-2783.

KNOWLES, P. J. & HANDY, N. C. 1984. A new determinant-based full configuration interaction method. *Chemical Physics Letters,* 111**,** 315-321.

KOHN, W. & SHAM, L. J. 1965. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review,* 140**,** A1133-A1138.

KRISHNAN, R., SCHLEGEL, H. B. & POPLE, J. A. 1980. Derivative studies in configuration–interaction theory. *The Journal of Chemical Physics,* 72**,** 4654-4655.

LISIECKI, L. E. & RAYMO, M. E. 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography,* 20**,** n/a-n/a.

LOWE, J. J. & WALKER, M. 2014. *Reconstructing quaternary environments*, Routledge.

MØLLER, C. & PLESSET, M. S. 1934. Note on an Approximation Treatment for Many-Electron Systems. *Physical Review,* 46**,** 618-622.

MOYNIER, F., FUJII, T., SHAW, A. S. & LE BORGNE, M. 2013. Heterogeneous distribution of natural zinc isotopes in mice. *Metallomics,* 5**,** 693-699.

POPLE, J. A., SEEGER, R. & KRISHNAN, R. 1977. Variational configuration interaction methods and comparison with perturbation theory. *International Journal of Quantum Chemistry,* 12**,** 149-163.

PRASAD, A. S., BAO, B., BECK, F. W. J., KUCUK, O. & SARKAR, F. H. 2004. Antioxidant effect of zinc in humans. *Free Radical Biology and Medicine,* 37**,** 1182-1190.

REDLICH, O. 1935. Eine allgemeine Beziehung zwischen den Schwingungsfrequenzen isotoper Molekeln. *Zeitschrift für Physikalische Chemie,* 28**,** 371-382.

SCHAUBLE, E. A. 2004. Applying stable isotope fractionation theory to new systems. *Geochemistry of Non-Traditional Stable Isotopes,* 55**,** 65-111.

SCHRÖDINGER, E. 1926. Quantisierung als Eigenwertproblem. *Annalen der Physik,* 384**,** 361-376.

SCHWARCZ, H. P. & SCHOENINGER, M. J. 1991. Stable isotope analyses in human nutritional ecology. *American Journal of Physical Anthropology,* 34**,** 283-321.

UREY, H. C. 1947. The thermodynamic properties of isotopic substances. *Journal of the Chemical Society (Resumed)***,** 562-581.

WEISS, D., NORTHOVER, G., HANIF, M., GARCÍA-ESPAÑA, E., VILAR, R., ARNOLD, T., MARKOVIC, T., WISSUWA, M. & DELGADO, E. 2021. Isotope fractionation of zinc in the paddy rice soil-water environment and the role of 2'deoxymugineic acid (DMA) as zincophore under Zn limiting conditions. *Chemical Geology,* 577**,** 120271.

WILSON, E. B., DECIUS, J. C. & CROSS, P. C. 1980. *Molecular vibrations: the theory of infrared and Raman vibrational spectra*, Courier Corporation.

Appendices

- Appendix 1: Relative sizes of translational, rotational and vibrational quanta (taken from Schauble (2004))
- Appendix 2: Introduction to Electronic Structure Theory
- Appendix 3: Table of regular expressions

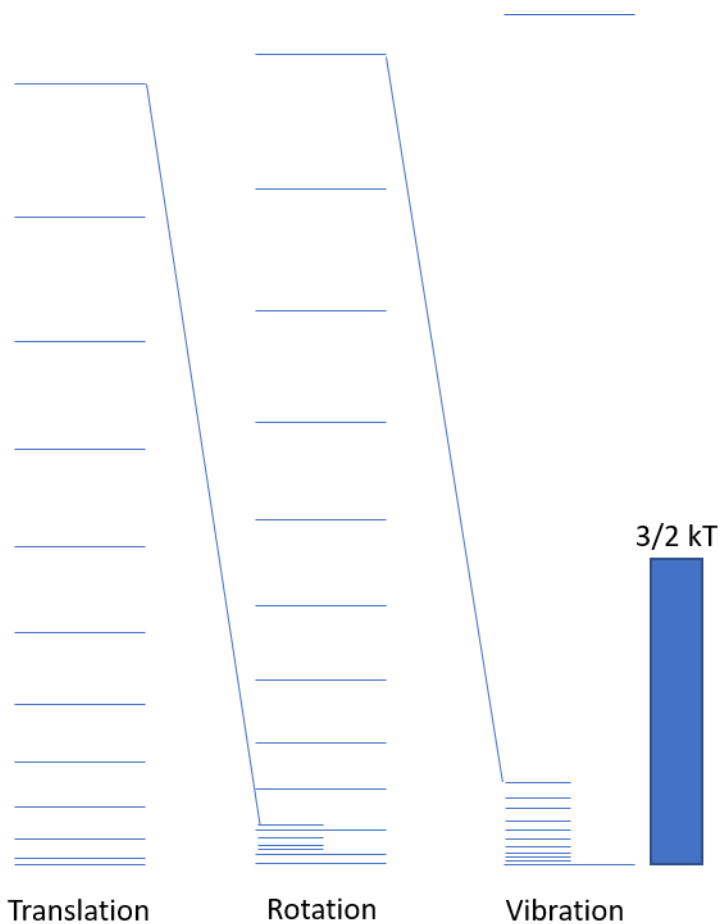## Appendix 1: Relative sizes of translational, rotational and vibrational quanta

Figure 1: Relative energies of translational, rotational and vibrational quanta (reproduced from the original - Fig. 2 in Schauble (2004)). 3/2kT is the classical thermal energy of a particle at 298K, shown at the same scale as vibrational quanta. Not all the translational energies are shown at the rotational scale, and due to their size, rotational and translational quanta are not needed when calculating fractionation values. For further details on calculation of the energy levels see the original image.

Appendix 2: Introduction to Electronic Structure Theory

Electronic structure theory applies quantum mechanics to atoms/molecules to predict their energies and properties. These are also known as "ab initio" methods, and they solve the time independent Schrödinger equation (Schrödinger, 1926) to obtain the wavefunction for the system in question.

$$H\Psi = E\Psi$$

Ψ is the wavefunction, E is the energy, and H is the Hamiltonian operator.

$$H = -\sum_i^{electrons}\frac{\hbar}{2m_e}\nabla_i^2 - \sum_A^{nuclei}\frac{\hbar}{2m_A}\nabla_A^2 - \sum_i^{electrons}\sum_A^{nuclei}\frac{e^2 Z_A}{r_{iA}} + \sum_{i>j}^{electrons}\frac{e^2}{r_{ij}} + \sum_{A>B}^{nuclei}\frac{e^2 Z_A Z_B}{r_{AB}}$$

$\nabla^2 = \frac{\partial}{\partial x^2} + \frac{\partial}{\partial y^2} + \frac{\partial}{\partial z^2}$, $\hbar = \frac{h}{2\pi}$, r_{xy} is the distance between x and y, e denotes electrons, A and B denote nuclei, m is the mass of the species, and Z is the charge on the nuclei.

The first 2 terms in the Hamiltonian are the kinetic energy of the electrons and nuclei respectively. The other 3 terms are the potential energy of the nuclei due to the attraction of the electrons, electron-electron repulsions and nuclear repulsion. Given that the nuclei are much heavier and move significantly slower than electrons, their positions can be frozen and only consider the positions of electrons when solving for energy and wavefunctions. This is known as the Born-Oppenheimer Approximation. (Born and Oppenheimer, 1927). The consequences for the Hamiltonian are:

- the second term, involving only the nuclei, can be ignored
- the final term is trivial to evaluate, and can be done so separately, as all that is needed is the charge on the nuclei and the distances from one another. This is known is the nuclear repulsion energy. (E_{NRE})
- The remaining terms (1,3 and 4) are the ones that need to be solved. The energy that is found can be added to E_{NRE} to get a total energy for the system.

The Hamiltonian incorporating the Born-Oppenheimer approximation is as follows:

$$H_{BO} = -\sum_i^{electrons}\frac{\hbar}{2m_e}\nabla_i^2 - \sum_i^{electrons}\sum_A^{nuclei}\frac{e^2 Z_A}{r_{iA}} + \sum_{i>j}^{electrons}\frac{e^2}{r_{ij}}$$

The equations to solve to get molecular properties are now:

$$E_{NRE} = \sum_{A>B}^{nuclei}\frac{e^2 Z_A Z_B}{r_{AB}}$$

$$H_{BO}\Psi = E_{BO}\Psi$$

$$E = E_{NRE} + E_{BO}$$

Because of the Born-Oppenheimer approximation, the total energy of a molecule can change with small adjustments to nuclei positions – the best way to visualise this is a potential energy surface which links molecular geometry and energy.

The Schrödinger equation can only be solved exactly for 1 electron systems – as the exact wavefunction can be found – for all other systems, inter electron repulsions exist, so approximations to the wavefunction are needed. The wavefunction is approximated by basis sets, and the Hamiltonian is approximated by various methods - from the simple Hartree-Fock (Hartree, 1928), to Density Functional Theory (Hohenberg and Kohn, 1964, Kohn and Sham, 1965).

The basic, single electron Hartree-Fock method attempts to solve the Schrödinger equation by reducing it to a set of single electron wavefunctions. This forms the basis of all major levels of theory.

The single electron orbitals can be found by solving the single electron Schrödinger equation:

$$h_i \psi_i = \epsilon_i \psi$$

$$h_i = -\frac{1}{2}\nabla_i^2 - \sum_{J=1}^{N} \frac{q_J}{r_{ij}} + V_i^{eff}(r)$$

The first term is the kinetic energy of electrons, the second term is the attraction of electron I to all nuclei, and the final term is the effective potential acting on electron i:

$$V_i^{eff}(r) = \sum_{j \neq i} \int \frac{\psi_j^2}{r_{ij}} dr$$

The only way to find the single electron orbitals is to use an iterative procedure starting with an initial guess for the orbitals, calculate the effective potential, solve the one electron equations, and iterate until convergence.

For a more comprehensive look, see Foresman and Frisch (2015).

Electron correlation methods include the correlation between methods of opposite spin and solve for energies with high accuracy, but at a high computational cost.

The Configuration Interaction method (Pople et al., 1977, Krishnan et al., 1980) starts off with by expanding around Hartree wavefunction and sequentially adds wavefunctions for electronically exicted states. If this is taken to the full extent, the full CI method is obtained. (Knowles and Handy, 1984, Jr. and Taylor, 1986, Bytautas and Ruedenberg, 2004) However, CI is only practical for small systems.

Moller-Plesset perturbation theory (Møller and Plesset, 1934), another electron correlation method, applies perturbations to the Hamiltonian, by starting from a related problem where the solution is known, and treating what is left from the actual problem to be solved as perturbations. There are several orders of MP theory, with first order equal to HF theory, and successive orders becoming increasingly computationally expensive.

In 1964 Hohenberg and Kohn developed the Hohenberg-Kohn theorem which stated that "the ground state energy, and all other ground state electronic properties are uniquely determined by electron density." (Hohenberg and Kohn, 1964). Kohn and Sham then took this and developed Density Functional Theory. (Kohn and Sham, 1965).

In Density Functional Theory, the ground state electron density is a sum of densities constructed from one electron orbitals, which are found by iteratively solving the Kohn-Sham equations, which are conceptually like the Hartree-Fock equations, so DFT is considerably faster than the previous methods.

Appendix 3: Table of Regular Expressions

Table 1: Table of regular expressions used in extracting information from log files

Information Extracted	Regular Expression
Temperature	"Temperature(?:\s|=)*(\d*.\d*)"
Frequencies	"Frequencies --"
Isotope	"([A-Za-z]*)\(Iso=(\d*)\)"
Item Convergence Table	"(^.*?Converged\?(?:\n.*(YES)){4})"
Atomic Coordinate Matrices	"^.*?Z(?:\n.*(\d\.\d+)){1,}"
Individual Coordinate Lines	"\s+\d*\s+(\d+)((?:(?:\s*)(?:\s|-)\d\.\d\d){1,})"
Low Frequencies	"Low frequencies ---"