Linear Programming to Determine Molecular Orientation at Surfaces through Vibrational Spectroscopy

by

Fei Chen
B.Sc., University of Victoria, 2017

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

© Fei Chen, 2017
University of Victoria

All rights reserved. This dissertation may not be reproduced in whole or in part, by photocopying or other means, without the permission of the author.

Linear Programming to Determine Molecular Orientation at Surfaces through Vibrational Spectroscopy

by

Fei Chen
B.Sc., University of Victoria, 2017

Supervisory Committee

Dr. Ulrike Stege, Co-Supervisor
(Department of Computer Science)

Dr. Dennis Hore, Co-Supervisor
(Department of Chemistry)

Supervisory Committee

Dr. Ulrike Stege, Co-Supervisor
(Department of Computer Science)

Dr. Dennis Hore, Co-Supervisor
(Department of Chemistry)

ABSTRACT

Applying linear programming to spectroscopy techniques, such as IR, Raman and SFG, is a new approach to extract the molecular orientation information at surfaces. Research has shown the computational gain when using the linear programming approach. However, linear programming approach does not always return the known molecular orientation distribution information when mock spectral information is applied to the linear programming model. The goal of my study is to figure out the reason that causes the failure. To achieve this goal, a simplified molecular model is designated to study the nature of the linear programming model. With the information gained, I further apply the linear programming approach to various cases in order to verify whether it can be systematically applied in different circumstances.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My husband, for supporting me in the low moments.

Dr. Ulrike Stege, for all the support, encouragement, inspiration and patience. I can only finish my thesis with all her help and courage.

Dr. Dennis Hore, for always giving me new ideas and wonderful discusses.

Kuo Kai Hung, for previous working and information sharing.

PITA and Dennis groups, for all the fun and knowledge we share in our weekly meeting.

Chapter 1

Introduction

1.1 Background and Motivation

A surface is what forms a common boundary between two phases of matter. The phases of matter can be of any form, i.e, solid, liquid, or gas. The behavior of a surface greatly affects the properties of a material. Examples for such behaviors are: oxidation, corrosion, chemical activity, deformation and fracture, surface energy and tension, adhesion, bonding, friction, lubrication, wear and contamination. Therefore, surface characterization identification remains an active area of research in the physics, chemistry, and biotechnology communities, as well as in modern electronics. It also plays a crucial role in surface science. Among various surface properties, molecular orientation is a key factor of all, because molecular orientation greatly affects molecules' surface properties in aspects such as: adhesion, lubrication, catalysis, and bio-membrane functions [12].

Many experimental techniques have been applied in the study of molecular orientation at surfaces. Among them the optical methods are preferable. Such methods include infrared (IR) absorption, Raman scattering and visible-infrared sum-frequency generation (SFG) spectroscopies. All these vibrational spectra carry quantitative structural information of molecules at surfaces. Although each of them has its own strengths and shortcomings, they all share the following advantages when compared with non-optical methods. First of all, they all can be applied to any surfaces accessible by light. Second, they are non-destructive. Third, they offer good spatial, temporal and spectral resolutions [2,12]. An important advantage of SFG techniques is that it can discriminate against bulk contributions. This means that its result will not take the effect from the bulk. In order to extract the quantitative structural information that molecules carry at surfaces, different spectroscopy techniques and analyse are required. Combining different spectroscopy techniques is a very effective way to study the goal of molecular orientation at surfaces. However, finding the most effective ways to combine these techniques are not known so far.

In order to analyze these vibrational spectra, various factors need to be considered. For example, a molecule's vibrational mode in the molecular frame, the orientation average of the molecules adsorbed onto the surface based on the mathematical orientation distribution function, and projecting the vibrational mode properties from the molecular frame to the laboratory frame. The main focus of our study is to apply Linear Programming (LP) using different spectral information to obtain molecular orientation distribution at surfaces. In this thesis, we will explore how LP can facilitate extracting quantitative structural information of molecules at surfaces.

In this thesis, we describe the problem at hand as LP problems. Our approach is to first study our LP model's properties by applying it to a simplified molecular model. After that, the LP model is applied to the representatives of realistic molecules, to further explore the possibilities of our LP model. The realistic molecules that we consider are six amino acids: methionine (Met), leucine (Leu), isoleucine (Ile), alanine (Ala), threonine (Thr) and valine (Val).

Before describing the LP basics and the molecule orientation studies, the basic theory of the IR, Raman and SFG spectra is introduced.

1.2 Experimental Probes: IR, Raman, SFG Vibrational Spectroscopy [9]

Vibrational spectra are produced by the changes of a molecule's dipole moment and polarizability. The dipole moment and polarizability are changing as the molecule's conformation is changing.

IR is the absorption of passing infrared light through a sample at each frequency, which can be expressed by Equation 1.1.

$$A_{\mathrm{IR}} = -log_{10}\left(\frac{I}{I_o}\right) \tag{1.1}$$

where A_{IR} is the measured IR absorbance. I is the light intensity after infrared light passes through the sample, and I_o is the initial light intensity.

The physical principle of IR spectra is the variation of the dipole moment μ (the first rank tensor) along the normal coordinates Q: $\partial\mu/\partial Q$. IR spectra can be further expanded by Equation 1.2.

$$A_{\mathrm{IR}} \approx \left| \frac{1}{\sqrt{2m_q w_q}} \frac{\partial\mu}{\partial Q} \right|^2 \tag{1.2}$$

where m_q is the reduced mass of the normal mode, and w_q is the resonance frequency. The dipole moment μ is a vector of x, y and z. The dipole moment derivatives can be expressed as Equation 1.3. The IR spectra can be obtained from three polarizations: x, y and z.

$$\frac{\partial\mu}{\partial Q} = \begin{bmatrix} \frac{\partial\mu_x}{\partial Q} \\ \frac{\partial\mu_y}{\partial Q} \\ \frac{\partial\mu_z}{\partial Q} \end{bmatrix} \tag{1.3}$$

In the Raman process, stocks-shifted light may be scattered from a molecule sample. Unlike IR, Raman spectra relate to the variation of the molecular polarizability α (the second rank tensor) along the normal coordinates Q: $\partial\alpha/\partial Q$.

$$I_{\mathrm{Raman}} \approx \left| \frac{1}{\sqrt{2m_q w_q}} \frac{\partial\alpha^{(1)}}{\partial Q} \right|^2 \tag{1.4}$$

where m_q and w_q are the same as defined in Equation 1.2. The polarizability is coupled with x, y, z components of the driving field and x, y, z components of the induced polarization. Therefore, there are 9 elements in the polarizability, which can be expressed as Equation 1.5. It results in 9 polarizations of Raman spectra: xx, yy, zz, xy, xz, yx, yz, zy and zx.

$$\frac{\partial \alpha^{(1)}}{\partial Q} = \begin{bmatrix} \frac{\partial \alpha^{(1)}_{xx}}{\partial Q} & \frac{\partial \alpha^{(1)}_{xy}}{\partial Q} & \frac{\partial \alpha^{(1)}_{xz}}{\partial Q} \\ \frac{\partial \alpha^{(1)}_{yx}}{\partial Q} & \frac{\partial \alpha^{(1)}_{yy}}{\partial Q} & \frac{\partial \alpha^{(1)}_{yz}}{\partial Q} \\ \frac{\partial \alpha^{(1)}_{zx}}{\partial Q} & \frac{\partial \alpha^{(1)}_{zy}}{\partial Q} & \frac{\partial \alpha^{(1)}_{zz}}{\partial Q} \end{bmatrix} \tag{1.5}$$

SFG stands for sum-frequency generation vibrational spectroscopy. SFG is a surface-specific technique. It is a non-linear optical process. SFG is the variation of the outer product of dipole moment and polarizability, $\alpha^{(2)}$ (the third rank tensor): $\frac{\partial \mu}{\partial Q} \otimes \frac{\partial \alpha}{\partial Q}$. Therefore, there are 27 elements for SFG spectra, which result in three unique polarizations of SFG spectra: xxz, xzx, and zzz.

$$I_{\text{SFG}} \approx \left| \alpha^{(2)}_{ijk} \right|^2 = \left| \frac{1}{2m_Q w_Q} \left(\frac{\partial \alpha^{(2)}_{ij}}{\partial Q} \otimes \frac{\partial \mu_k}{\partial Q} \right) \right|^2 \tag{1.6}$$

1.3 Linear Programming [7]

LP problems are optimization ones of a specific form. The standard form of LP is a minimization problem that has an objective function and a number of constraints as shown in Equation 1.7 [6].

$$\begin{array}{rrcl} \text{minimize} & c_1x_1 + c_2x_2 + \cdots + c_nx_n & & \\ \text{subject to} & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 \\ & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 \\ & \vdots & & \vdots \\ & a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = & b_m \\ & x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0 & & \end{array} \tag{1.7}$$

Food	Carrot	Cabbage	Cucumber	Required per dish
Vitamin A [mg/kg]	35	0.5	0.5	0.5mg
Vitamin C [mg/kg]	60	300	10	15mg
Dietary Fiber [g/kg]	30	20	10	4g
price[$/kg]	0.75	0.5	0.15	-

Table 1.1: Sample input of the diet problem.

where x_i are the decision variables, a_{ij} is a matrix of known coefficients, b_i and c_i are vectors of known coefficients. The expression to be minimized is called objective function. The equalities and the inequalities are the constraints that all the decision variables need to subject to. These constraints specify a convex polytope that the objective function needs to optimize over.

The diet problem is a popular example to illustrate the concept of LP. It is described as follows: a restaurant would like to satisfy certain minimal nutrition requirements with the lowest price over some food selections, as shown in Table 1.1. In this example, in each meal, the minimum requirements for vitamin A, vitamin C and dietary fiber are 0.5mg, 15mg and 4g. The restaurant has three food options: raw carrot, raw white cabbage and pickled cucumber. The table also displays the nutrition content and the price of each ingredient. With all the information, we want to know how much carrot, cabbage and cucumber is needed in each meal, so that the minimal nutrition requirements can be met with the lowest price. In summary, the goal is to minimize the price, and the constraints are the nutrition requirements. Therefore, the following LP problem is formulated as shown in Equations 1.8 to 1.14.

$$\begin{aligned}
\text{minimize} \quad & 0.75x_1 + 0.5x_2 + 0.15x_3 && (1.8)\\
\text{subject to} \quad & 35x_1 + 0.5x_2 + 0.5x_3 \geq 0.5 && (1.9)\\
& 60x_1 + 300x_2 + 10x_3 \geq 15 && (1.10)\\
& 30x_1 + 20x_2 + 10x_3 \geq 4 && (1.11)\\
& x_1 \geq 0 && (1.12)\\
& x_2 \geq 0 && (1.13)\\
& x_3 \geq 0 && (1.14)
\end{aligned}$$

where x_1, x_2 and x_3 are the decision variables. Each decision variable presents the amount of the corresponding ingredient. Equation 1.8 is the objective function to be minimized. Equations 1.9 to 1.11 describe the nutrition requirements. Equations 1.12 to 1.14 ensure the amount of each ingredient to be greater than 0. The coefficients in the objective function represent c_i vector in Equation 1.7. The coefficients of the decision variables in Equation 1.9, 1.10 and 1.11 represent a_{ij} matrix. b_i vector is composed by the right-hand side of Equation 1.9, 1.10 and 1.11.

The simplex method is an algorithm designed to solve LP problems. In order to apply simplex method, the above LP problem needs to transfer into its standard form. The inequalities of Equations from 1.9 to 1.11 need to transform to equalities. Therefore, a new variable, called a slack variable (SV) is introduced to change each inequality to equality [1]. The standard form of the above LP model is shown in Equation 1.15.

$$\begin{aligned} \text{minimize} \quad & 0.75x_1 + 0.5x_2 + 0.15x_3 \\ \text{subject to} \quad & 35x_1 + 0.5x_2 + 0.5x_3 - s_4 = 0.5 \\ & 60x_1 + 300x_2 + 10x_3 - s_5 = 15 \\ & 30x_1 + 20x_2 + 10x_3 - s_6 = 4 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \\ & x_3 \geq 0 \\ & s_4 \geq 0 \\ & s_5 \geq 0 \\ & s_6 \geq 0 \end{aligned} \tag{1.15}$$

where s_4, s_5 and s_6 are the introduced SVs.

With the existing LP solvers that implemented simplex method, the optimal solution can be obtained within a second.

It has been shown that for any LP problem, there are only three kinds of possible solutions: feasible and bounded, feasible and unbounded, and infeasible. If the solution space is feasible and bounded, then there is exactly one optimum solution. If it is feasible but unbounded, then there is a solution space with an infinite number of optimal solutions [3].

A general LP problem can be a minimization or maximization problem. Its constraints can be equalities or inequalities. For each non-standard LP problem, there are ways to convert it into its standard form. Furthermore, for a LP problem that contains n decision variables, its solution would be in the n-dimensional space R^n. Each constraint is a hyperplane. It divides R^n into two half-spaces. Therefore, all the constraints together cut this R^n space into a convex polyhedron when there are feasible solutions. This makes LP a convex problem. The benefit of a convex problem is that a local optimal solution is also the global optimum. LP solvers return the optimal solution. If a LP problem has a unique optimal solution, this solution is a vertex of the convex polyhedron. In other words, LP is a convex, deterministic process. It is guaranteed to converge to a single global optimum if there is a bounded solution space.

Another advantage of LP is that LP solvers can deal with tens or hundreds of thousands of variables, which makes it suitable for the study of a molecule's orientation distribution at surfaces. Furthermore, LP problems are intrinsically easier to solve than many non-linear problems.

Various algorithms are available in solving LP problems, such as: simplex, interior point, and path-following algorithms. Both interior point and simplex algorithms are common and mature ones that work well in practice. The simplex method is comparatively easier to understand and implement than interior point one. The simplex method takes the advantage of the geometric concept that it visits the vertices of the feasible set (convex polyhedron), and checks the optimal solution among each visited vertex. The converging approach is different for these two methods. If there are n decision variables, usually simplex method converges in $O(n)$ operations with $O(n)$ pivots. Interior point traverses the edges between vertices on a polyhedral set. Generally speaking, the interior point method is faster for larger problems that have a sparse matrix. However, when we were experimenting with these two methods, the speed of them was not much different from each other for our study. In our study, the simplex method has proved to be efficient and effective, and it is used for all the test cases described in this thesis.

The last but not the least advantage of LP is the speed of existing LP solvers. For any LP problem, if it has an optimal solution, this solution is always a vertex. Simplex method is based on this insight, namely that it starts at a vertex, then pivot from vertex to vertex, until it reaches the optimum. Although it has been shown that simplex method is not a polynomial algorithm, in practice it usually takes $2n \sim 3n$ steps to solve a problem (n is the number of the decision variables).

The LP solver we use is called "GNU linear programming tool kit" (GLPK). It has implemented both simplex and interior point methods in C programming language. It is open-source and intended to solve large scale LP problems.

Currently there are two main approaches in studying the orientation distribution of molecules at surface. One is comparing the experimental spectra with few predicted ones, and select the one that most matches the experimental one. Another one is running an exhaustive algorithm to explore the most possible solution space [9]. However, both approaches take a lot of time and computational resources. Therefore, applying LP will result a large gain in computation.

1.4 Conclusion and Open Questions from Previous Study [4]

Our research is based on Hung's study. In his study, he mentioned that generating model spectra that match the target experimental spectra from a list of known candidates is a way to extract molecular orientation information at surface. The traditionally exhaustive way of achieving this goal consumes too much computational effort, therefore, he introduced LP approach to vibrational spectroscopy study. The LP approach results in pseudo polynomial time $O(n)$, which is a great improvement compared to $O(n!)$.

However, depends on different test case settings, the LP approach may not always return the target composition of candidates that generates the mock target experimental spectra. When considering the candidates from one type of molecule at surfaces, the return solution of the LP approach does not match the known target composition. When considering the candidates from a mixture of molecules, the return solution of the LP approach does match the know target composition. The reason why the LP approach failed to return the target composition has not been thoroughly studied by Hung.

Moreover, when applying the LP approach, only SFG spectral information has been considered in his LP model. The possibility and applicability of using IR and Raman spectral information to the LP approach have not been considered. Meanwhile, the possibility of combining different spectral information to the LP approach has not been considered.

1.5 Aims and Scope

The goal of our study is to figure out the underlying properties of the LP approach, figure out what is the cause that the LP approach fails to obtain the target composition of candidates in some test cases. Based on the gained information, we further explore the applicability of the LP approach to different test cases. Our plan is applying the LP approach to a simplified molecular model first to study the limitations of basic problem instances of our LP model the properties of our LP model.

With the properties learnt from the first step, the LP approach is then applied to realistic molecules. There are two types of test cases, one is considering the candidates coming from one type of realistic molecule, to see if the LP approach can return the target composition of the mock spectra. Another one is considering the candidates coming from different types of realistic molecules. If the LP approach achieve in obtaining the target composition, then how the LP approach applied systematically will be studied.

The purpose is to check if LP approach will return the target composition of the spectra for one type of molecule at surfaces. If yes, whether the LP model can be applied generally to one type molecule will be studied. If not, what is the underlie reason will be explored. Similar study will also be applied to different molecules at surfaces. At last, the experimental spectral information is brought into consideration.

1.6 Overview of the Thesis

The reminder of this thesis is as follows. Chapter 2 explains the current approaches to extract the molecular orientation distribution at surfaces, as well as how to produce IR, Raman and SFG spectra. Chapter 3 introduces the LP model using a simplified molecular model, and studies the properties of our LP model for this simplified test case. Chapter 4 applies our LP approach to one type of molecule at surfaces. Chapter 5 studies the LP approach to a mixture of different molecules at surfaces. Chapter 6 studies the LP approach to experimental spectral data. Chapter 7 is the conclusion and future work.

Chapter 2

Methods

2.1 Description

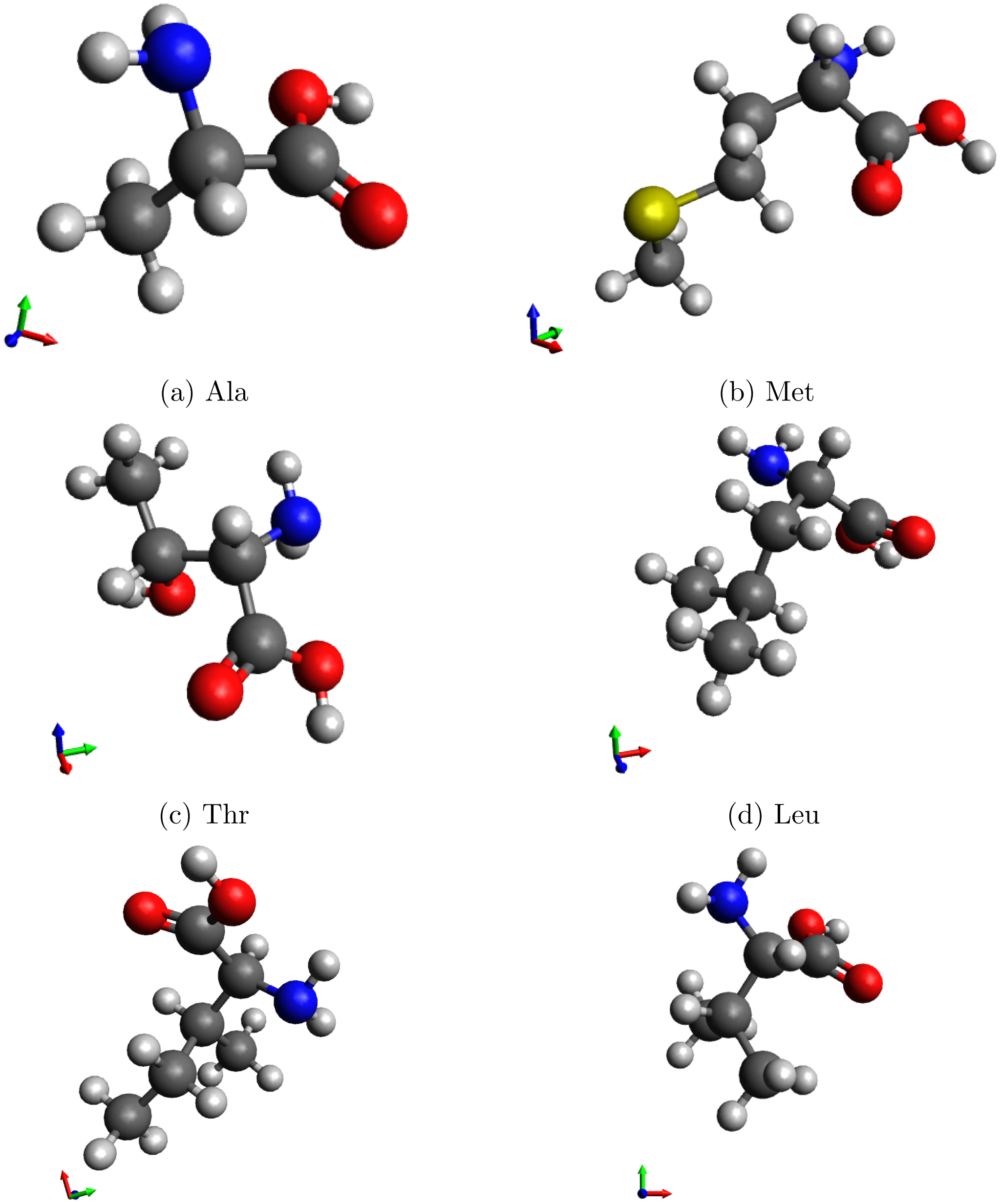
Before introducing and analyzing the LP model and applying it to the realistic molecules' vibrational spectra, there are a few factors to address. First of all, creating each amino acid's IR, Raman and SFG spectra is an essential step. This part research has been done thoroughly by Hung [4]. In this chapter, I introduce the content that is related to our study.

2.2 Structure of Realistic Molecules

Figure 2.1 illustrates the molecule structure of the six amino acids in the molecular frame. These amino acids are used in the test cases related to realistic molecules. The a, b and c are the molecular frame coordinates. When a molecule lays on a surface, we need to transfer the molecular frame to the lab frame where the surface exists.

2.3 Generating Model Spectra [5]

To generate these amino acids' vibrational spectra, a molecule's vibration modes need to be modelled in the molecular frame, and then transferred to the laboratory frame where surfaces exist. Chapter 2 in Hung's thesis [4] describes how to perform electronic structure calculations using a software package called The General Atomic and



(a) Ala

(b) Met

(c) Thr

(d) Leu

(e) Ile

(f) Val

Figure 2.1: Molecule structure of Ala, Met, Thr, Leu, Ile and Val in molecular frame. Blue axis is designated as c axis, red axis is designated as a axis, green is designated as b axis.

Molecular Electronic Structure System (GAMESS) [8] to obtain the derivatives of every dipole moment and polarizability. Then he introduced how to use Direction Cosine Matrix (DCM) to transfer these two derivatives from the molecular frame to the laboratory one. After that, Euler angles could be extracted from DCM. Euler angles are used to describe a molecule's orientation at surfaces. They are labelled by θ, φ and ψ as shown in Figure 2.2. They are referred to as *tilt*, *azimuthal* and *twist* angles, respectively. Let x, y and z be lab frame Cartesian coordinates, and let a, b and c be the molecular frame coordinates. *Tilt* angle θ is the angle between z and c. *Azimuthal* angle φ is the rotation about z. *Twist* angle ψ is a twist about c [9]. After three steps of successive rotations of Euler angles, molecule properties can be transferred from the molecular frame to the lab one.

In order to achieve the above steps, Hung first did a Hessian calculation using GAMESS. Secondly, seven snapshots of a molecule vibrating in different modes were taken. Thirdly, he did a force field calculation to obtain the derivatives of dipole moment and polarizability for each of the seven snapshot moment. Then the derivatives of dipole moment and polarizability are obtained by the interpolation of these seven snapshot moment. Because the two obtained derivatives are in the molecular frame, Hung used DCM to convert these two derivatives into the lab frame. Then he abstracted Euler angles from DCM. After these electronic structure calculations, the derivatives information, which is the molecular property information, is obtained.

In our study, those molecular property information is used to generate the amino acids' spectral information directly. Each molecule's property information contains the derivatives of dipole moment and polarizabilities of each vibrational mode. Depending on the number N of atoms in a molecule, there are $3N-6$ vibrational modes. Furthermore, Equations 2.2 to 2.5 are used to generate the amino acids' IR, Raman and SFG spectra.

All the test cases in our study are limited to only consider the *tilt* angle distribution of Euler angles, and assume isotropy on *twist* and *azimuthal* angular distributions. A uniform distribution is applied to *twist* and *azimuthal* angles. For angle φ, it requires the surfaces to be not striped, so that the molecule has no preference on the xy plane on the lab frame. There can be no anisotropy in the plane of the surface. Because of this, we can limit the candidate number by integrating angle φ. For angle

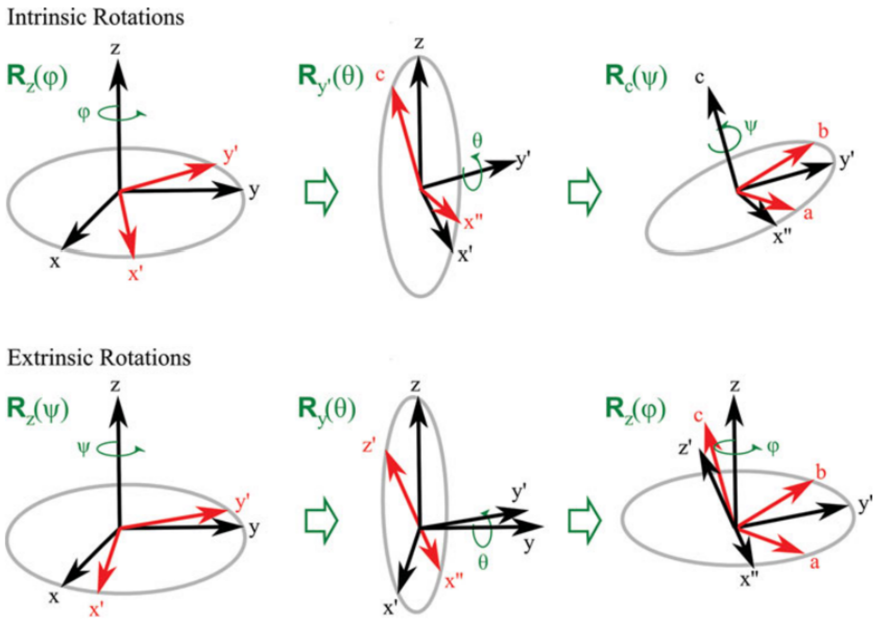


Figure 2.2: The Euler angles represented as the spherical polar angles θ, φ and ψ, and the illustration of the three successive rotations that transform the lab x, y, z coordinate system into the molecular a, b, c frame intrinsically and extrinsically. Reproduced from Ref. 9.

ψ, a uniform distribution implies that a molecule has cylindrical symmetry in its preference of surface. The molecule can be tilted, but has no '*twist*' preference. With the integration of these two Euler angles, the number of candidates for one molecule will be greatly reduced. Therefore, a candidate in our study is a specific molecule with specific θ value. However, the number of the candidates is still large when considering angle θ only. For example, from 0° till 180°, candidates are obtained in 10° steps, there are 18 candidates for just one molecule. For a mixture of six molecules, the number of possible combinations of all these molecules' candidates is $18^6 = 34012224$.

When molecules lay on a surface, the orientation of each single molecule varies. To simulate the vibrational spectra, a reasonable orientation distribution for the molecules needed to be studied. The orientation distribution requires either to do a molecular dynamic simulation in order to study the distribution of molecule orientations at surface, or come up with an analytic orientation distribution function. In our study, the LP approach is appropriate for the second method. Moreover, the δ-distribution function shown in Equation 2.1 is used to represent the molecule orientation distribution that models the spectrum signals. This means that all the molecules are tilted at the same angle at surface. This assumption is applied across the whole study.

$$f_{(\theta)} = \delta(\theta - \theta_o) \tag{2.1}$$

The absorption of an IR spectrum is proportional to the square of the lab-frame dipole moment derivative. For example, the x-polarized absorption spectrum is given by Equation 2.2:

$$A_x(\omega_{\mathrm{IR}}) = \sum_q \frac{1}{2m_q\omega_{\mathrm{q}}} \left\langle \left[\frac{\partial u_x}{\partial Q}\right]_q^2 \right\rangle \frac{\Gamma_q^2}{(\omega_{\mathrm{IR}} - \omega_{\mathrm{q}})^2 + \Gamma_q^2} \tag{2.2}$$

where A_x represents x-polarized IR obsorbance. ω_{IR} is the frequency of the probe radiation, μ is the dipole moment, m_q is the reduced mass, w_q is resonance frequency. Γ_q is the homogeneous line width, is set to 6 in all the test cases. Q_q is the normal mode coordinate of the qth vibrational mode. All values of ω_{IR}, μ, m_q, Q are obtained from the electronic structure calculations. Furthermore, because φ and ψ angles are integrated, the x-polarized spectrum is identical with the y-polarized one. Therefore, there are only two unique polarized IR spectra. For simplicity, IR spectra are referred to as x and z in future test cases. A_y and A_z are computed accordingly.

The intensity of Raman scattering is proportional to the square of lab frame transition polarizability. For example, Raman spectroscopy with an x-polarized excitation source collects the x-polarized component of the scattered radiation, which can be approximated using Equation 2.3.

$$I_{xx}(\Delta\omega) = \sum_q \frac{1}{2m_q\omega_{\mathrm{q}}} \left\langle \left[\frac{\partial \alpha_{xx}^{(1)}}{\partial Q}\right]_q^2 \right\rangle \frac{\Gamma_q^2}{(\Delta\omega - \omega_{\mathrm{q}})^2 + \Gamma_q^2} \tag{2.3}$$

where I_{xx} represents xx-polarized Raman intensity. Δw is the Stokes Raman shift. $\alpha_{xx}^{(1)}$ is one component of the nine-element polarizability tensor. m_q, w_q, Γ_q, and Q_q are the same as defined above for IR spectra. All the values of ω_{IR}, μ, m_q, Q are obtained from the electronic structure calculations. Similar to IR spectroscopy, because of the integration of φ and ψ angles, only four unique spectra are obtained from the following polarization: xx, xy, xz and zz. For simplicity, Raman spectra are referred to as xx, xy, xz and zz in future test cases.

SFG spectral signal is the imaginary part of the second-order susceptibility, $\left|\chi^{(2)}\right|$. $\chi^{(2)}$ is derived from the second-order polarizability $\alpha^{(2)}$ as shown in Equation 2.4. The imaginary part of $\left|\chi^{(2)}\right|$, which is the SFG spectral signal, is displayed as Equation 2.5.

$$\chi^{(2)}_{xxz}(\omega_{\mathrm{IR}}) = \sum_{q} \frac{1}{2m_q\omega_{\mathrm{q}}} \left\langle \left[\frac{\partial \alpha^{(1)}_{xx}}{\partial Q}\right]_q \left[\frac{\partial u_z}{\partial Q}\right]_q \right\rangle \frac{1}{\omega_{\mathrm{q}} - \omega_{\mathrm{IR}} + i\Gamma_q} \tag{2.4}$$

$$\mathrm{Im}\left[\chi^{(2)}_{\mathrm{xxz}}(\omega_{\mathrm{IR}})\right] = \sum_{\mathrm{q}} \frac{1}{2\mathrm{m_q}\omega_{\mathrm{q}}} \left\langle \left[\frac{\partial \alpha^{(1)}_{\mathrm{xx}}}{\partial \mathrm{Q}}\right]_{\mathrm{q}} \left[\frac{\partial \mathrm{u_z}}{\partial \mathrm{Q}}\right]_{\mathrm{q}} \right\rangle \frac{\Gamma_{\mathrm{q}}}{(\omega_{\mathrm{q}} - \omega_{\mathrm{IR}})^2 + \Gamma_{\mathrm{q}}^2} \tag{2.5}$$

where $\chi^{(2)}_{xxz}$ is the second-order susceptibility. It is probed by an x-polarized visible incoming beam at frequency ω_{vis} and a z-polarized infrared beam incoming with frequency ω_{IR}. Both incoming beams are incident to the sample. Then the x-component at frequency $\omega_{\mathrm{SFG}} = \omega_{\mathrm{vis}} + \omega_{\mathrm{IR}}$ is selected for detection. As $i = \sqrt{-1}$ is in the denominator, $\chi^{(2)}$ is a complex value [4]. The SFG response considered in this thesis is the imaginary component of the $\chi^{(2)}$. Same as IR and Raman spectroscopy, all the values of ω_{IR}, μ, m_q, Q are obtained from the electronic structure calculations. Because of the integration of φ and ψ angles, only three unique non-zero spectra are obtained from the following polarizations: xxz, xzx and zzz. For simplicity, SFG spectra are referred as xxz, xzx and zzz in future test cases.

With these equations and the electronic structure calculations, IR, Raman and SFG spectra can be generated for a candidate of a molecule. Taking Met as an example, Figure 2.3 displays x-polarized IR spectra of the following candidates: Met with θ equals 0°, 20°, 40° and 60°. Their spectra are prefixed with *candidate_* in the labels. *ir_x_* indicates the spectroscopy technique, "number" indicates the θ angle's value. The spectra labelled as *target_ir_x* is generated by combining 10% of *candidate_ir_x_*0, 50% *candidate_ir_x_*20 and 40% *candidate_ir_x_*40.

Similarly, Figures 2.4, 2.5 and 2.6 depict the spectra of the same candidates and targets for z-polarized IR, xx-polarized Raman and xxz-polarized SFG spectrum, respectively. In Figure 2.3, the biggest differences among the candidates exist at each vibrational mode. The valid range for the wavenumber is 1000 to 2000.

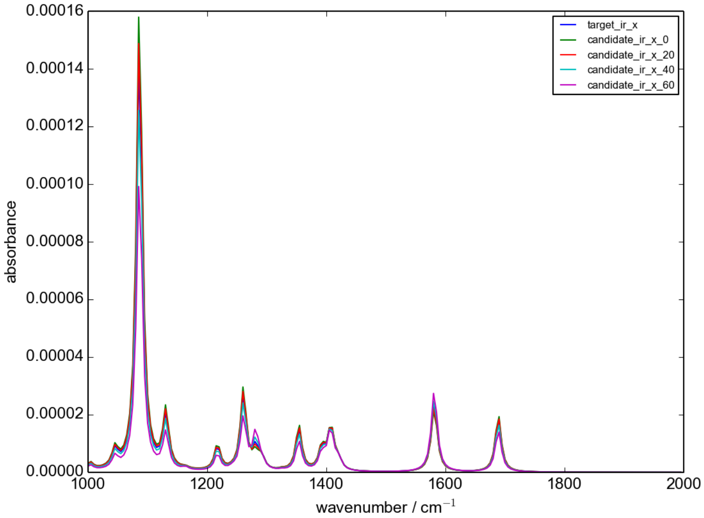


Figure 2.3: IR x-polarized spectra of methionine's four candidates and target. The candidates are with θ of 0°, 20°, 40° and 60°. The target is produced by combining 10% of $candidate\_ir\_x\_0$, 50% $candidate\_ir\_x\_20$ and 40% $candidate\_ir\_x\_40$.

2.4 Conclusion

Chapter 2 briefly explains what the current approaches are to extract molecular orientation distribution at surfaces, the molecular structures of six amino acids, and how to produce IR, Raman and SFG spectra theoretically. In Chapter 3, our LP model is described and its properties are studied. It is conducted by using a simplified molecular model to gain an insight of our approach. The motivation of creating a simplified molecular model is to create a molecule as simple as possible that will allow us to study the properties of the LP model for this basic case. Information gained in Chapter 3 allows us to then study the approach using molecules in Chapters 4, 5 and 6.

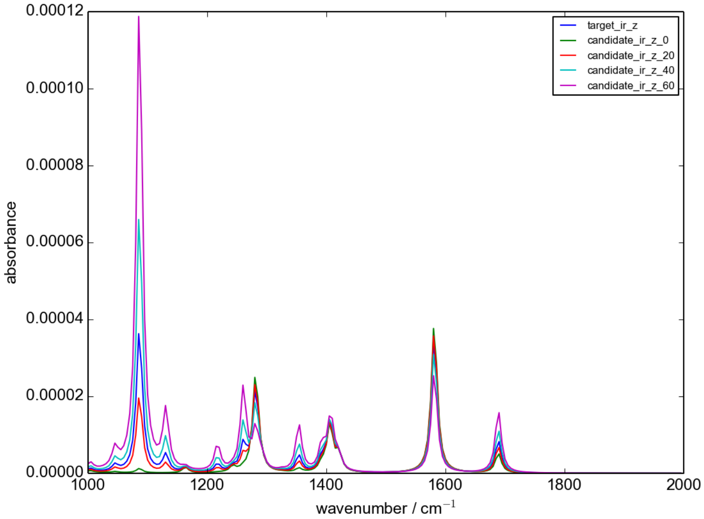


Figure 2.4: IR z-polarized spectra of methionine's four candidates and target. The candidates are with θ of 0°, 20°, 40° and 60°. The target is produced by combining 10% of *candidate_ir_x_0*, 50% *candidate_ir_x_20* and 40% *candidate_ir_x_40*.

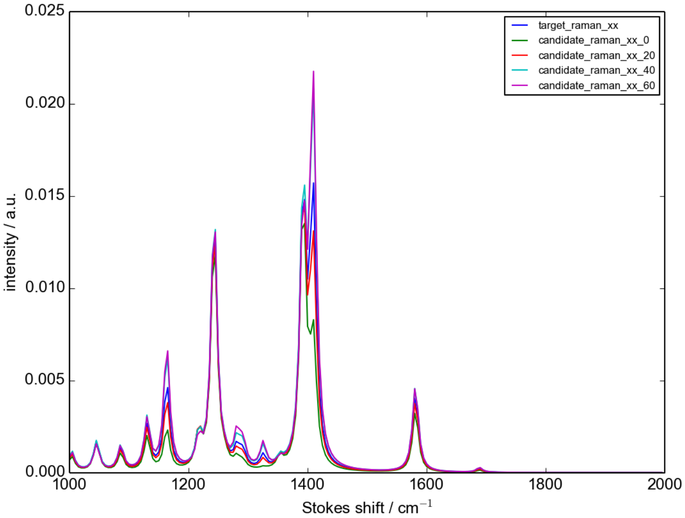


Figure 2.5: Raman xx-polarized spectra of methionine's four candidates and target. The candidates are with θ of 0°, 20°, 40° and 60°. The target is produced by combining 10% of *candidate_ir_x_0*, 50% of *candidate_ir_x_20* and 40% of *candidate_ir_x_40*.

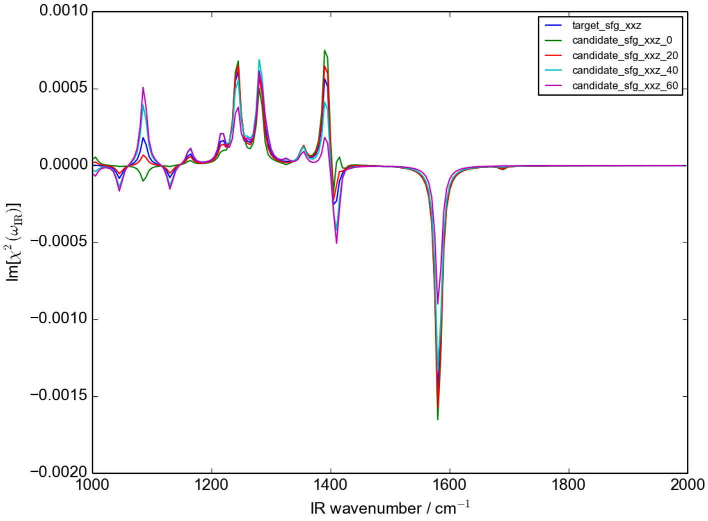


Figure 2.6: SFG xxz-polarized spectra of methionine's four candidates and target. The candidates are with θ of 0°, 20°, 40° and 60°. The target is produced by combining 10% of $candidate\_ir\_x\_0$, 50% of $candidate\_ir\_x\_20$ and 40% of $candidate\_ir\_x\_40$.

Chapter 3

Simplified Molecular Model

3.1 Description

The goal of this chapter is to introduce our LP model, as well as exploring its properties by using a simplified molecular model. The purpose of introducing the simplified molecular model is to limit the complexity that comes from the parameters needed to describe the realistic models, so that the analysis of the nature of the LP model can be focused.

Our simplified molecular model contains four vibration modes. Theses vibrational peaks are at frequencies of 2850, 2960, 3050 and 3200 cm^{-1}. The widths of the peaks are 5, 10, 5 and 15 cm^{-1}, respectively. The amplitudes of the peaks are 1, 0.7, -0.2 and 0.5 cm^{-1}, respectively. The comparing angles of the peaks are $15°, 90°, 0°$ and $60°$.

Only IR spectroscopy is considered for the simplified molecular model. Equation 3.1 is used to generate the z-polarized IR spectrum. Moreover, both Euler angles φ and ψ are integrated into uniform distribution, only the difference on angle θ is considered.

$$f_\theta(\omega_{\mathrm{IR}}) = \sum_{q=1}^{4} A_q^2 cos^2(\theta - \theta_q) \frac{\Gamma^2}{(\omega_{\mathrm{IR}} - \omega_{\mathrm{q}})^2 + \Gamma^2} \tag{3.1}$$

where A_q is the amplitude, θ_q is the comparing angle, Γ is the width, and ω_{q} is the frequency. Ten candidates are produced with ten different θ values as follows: 0°, 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, and 90°. Their spectra are shown in Figure 3.1. The 10 candidates have peaks at the same frequencies.

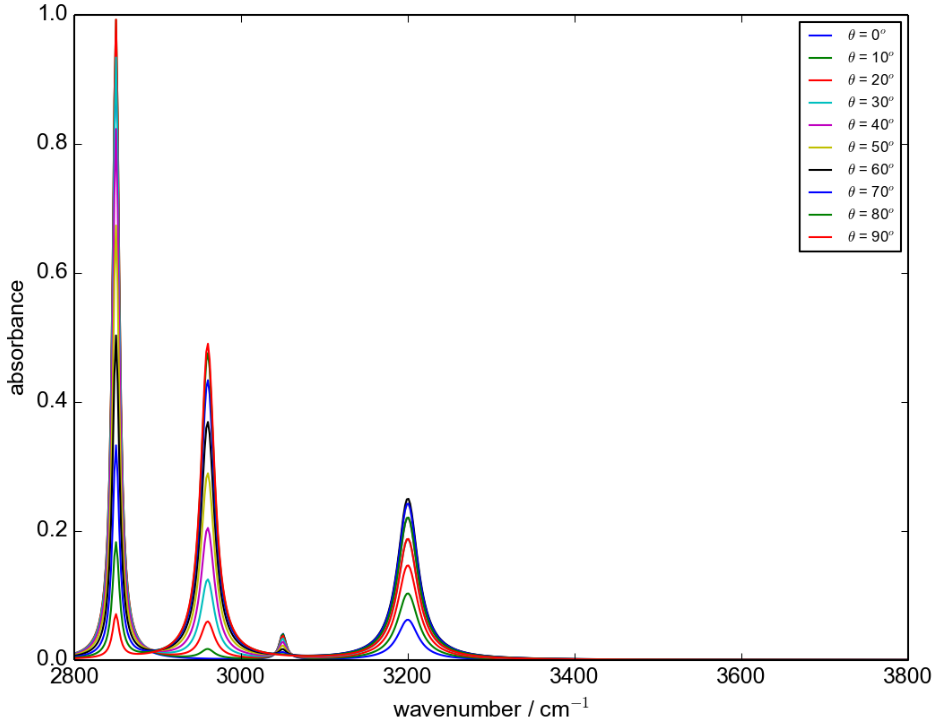


Figure 3.1: z-polarized IR spectra of candidates of simplified molecular model.

3.2 Linear Programming Model for Spectral Study

Equation 3.2 describes the objective function that build the basis of our LP model, as well as one constraint that limits the sum of all the candidates' percentage to 100%.

$$\begin{aligned}\underset{p_c}{\text{minimize}} \sum_{n=1}^{N_p} \left| \text{Target} - \sum_{c=1}^{N_c} p_c f_\theta(x) \right| \\ \sum_{c=1}^{N_c} p_c = 1 \end{aligned} \tag{3.2}$$

where p_c are the unknown percentages of the candidate, which are the decision variables. These percentages are returned by LP solver, and called return composition in future test cases. N_p is the number of points selected along the wavenumber, both for candidates and target spectra. Target refers to the corresponding data points selected in target spectra. N_c is the number of candidates. For each data point, the absolute residual between the target spectrum and the one composed by the decision variables is calculated. The objective function minimizes the sum of the absolute residuals over all the data points.

The optimal solution returned by the LP solver is then compared with the target composition to see if they match each other. This equation has also been used to study the composition of Ribonucleic acid (RNA) with ultraviolet (UV) spectra [10] and other UV spectroscopy studies [11] back in the 60s.

However, because of the absolute signs in the objective function, Equation 3.2 is not an LP problem. We transform Equation 3.2 by getting rid of the absolute signs. We introduce one more variable X and two more constraints to each data point as shown in Equation 3.3. The previous model in Equation 3.2 is then converted into the one in Equation 3.4, and it can be solved by LP solvers.

$$\begin{aligned} X &= \left| \text{Target} - \sum_{\mathrm{c=1}}^{\mathrm{N_p}} \mathrm{p_c f_\theta(x)} \right| \\ X &\geq \text{Target} - \sum_{\mathrm{c=1}}^{\mathrm{N_c}} \mathrm{p_c f_\theta(x)} \\ X &\geq -\text{Target} + \sum_{\mathrm{c=1}}^{\mathrm{N_c}} \mathrm{p_c f_\theta(x)} \end{aligned} \tag{3.3}$$

$$
\begin{aligned}
minimize \sum_{n=1}^{N_p} X_p \\
\mathrm{X_1 - Target_1 + \sum_{c=1}^{N_c} p_c f_\theta(x_1) \geq 0} \\
\mathrm{X_1 + Target_1 - \sum_{c=1}^{N_c} p_c f_\theta(x_1) \geq 0} \\
\vdots \\
\mathrm{X_n - Target_n + \sum_{c=1}^{N_c} p_c f_\theta(x_n) \geq 0} \\
\mathrm{X_n + Target_n - \sum_{c=1}^{N_c} p_c f_\theta(x_n) \geq 0} \\
\sum_{c=1}^{N_c} p_c = 1
\end{aligned}
\tag{3.4}
$$

Note that the LP model exactly describes our problem to be solved. Assuming that we can obtain sufficiently precise data, solving the LP will yield the target composition. Recall that if the solution space of an LP instance is feasible and bounded, then there is a unique optimum solution.

3.3 Linear Programming Model Implementation

Next, I describe how to solve instances of our LP model described in Equation 3.4. Code is written to generate a file that contains all the candidates' spectral information needed for the test cases. In this step, the molecular property information that generated from the electronic structure calculations are used. For a specific candidate, given the molecular property information and a value θ, the candidate's spectral information is obtained. To further illustrate, a candidate class is written. This class defines how to use the molecular property information to generate the needed spectral information. Given a candidate's molecular property information and a value θ, a instance of this specific candidate is created. For the simplified molecular model, this class only contains IR spectral information.

In the second step, additional code is written to generate a target composition of a list of candidates. Then the target composition is used to generate the target spectra. The probe range, which is the range of the wavenumber, is from 2800 to 3300 cm^{-1} for the simplified molecular model. It is from 2000 to 3000 cm^{-1} wavenumber for realistic molecules. The target spectral information is generated in the same text file as candidate's spectral information. Depend on the test case, code can be used to generate text files that contain selected types of spectral information.

In the third step, the LP model is constructed by using the spectral information text file generated in the second step. This part of the code was written by Hung [4]. It reads all the candidates and target spectral information, and builds the LP model as shown in Equation 3.4. It then outputs our LP input file for LP solver.

In the fourth step, we use as LP solver the "GNU linear progarmming tool kit" (GLPK) which will return the optimum solution for our input file.

3.4 Test Cases

In Cases 1 and 2, four candidates are selected. The detail is shown in Table 3.1. In Case 1, there are four candidates with θ of $0°, 10°, 20°$, and $30°$. In Case 2, the four candidates are of θ values $0°, 5°, 10°$, and $15°$. Instead of having ten degree variance in θ, a five degree difference is applied on θ in Case 2. This means that in Case 2 the candidates are more similar to each other than the ones in Case 1 as their spectra are more similar. In both cases, 100 data points are selected evenly along the wavenumber from z-polarized IR spectra. The target composition of the candidates is the same for both cases. In Case 1, the return composition is the same as the target one, however, the return composition for Case 2 does not match its target.

In order to figure out why the return composition in Case 2 is different from the target one, the spectra generated by the return composition is plotted together with the target spectra as shown in Figure 3.2. Note that the result spectra is identical to the target one. Note that their residual is 0. In order to see whether this observation can be generalized, Case 3 is set up in Table 3.2. Case 3 contains more candidates

Test Case index	1	2
Number of Candidates	4	4
Candidates	[0, 10, 20, 30]	[0, 5, 10, 15]
Target Composition	[0.1, 0.5, 0.4, 0]	[0.1, 0.5, 0.4, 0]
Number of Data Points	100, z	100, z
Return Composition	[0.1, 0.5, 0.4, 0]	[0, 0.80, 0.10, 0.1]

Table 3.1: Test cases 1 and 2 for the simplified molecular model.

than Cases 1 and 2. Ten candidates are included with θ values ranging from 0° to 90°.

Table 3.2 indicates that the return composition of Case 3 is different from the target one. Figure 3.3 shows that the spectrum produced by the return composition is almost identical to the one generated by the target composition in Case 3. The residual is negligible as well. This observation is comparable to Case 2.

Among Cases 1, 2 and 3, only the return composition of Case 1 matches its target one. However, in Case 2, the difference in value θ among the candidates is smaller than Case 1. In Case 3, the number of the candidates is larger than Case 1. Both effects increase the complexity of the test cases. In both Cases 2 and 3, the spectrum constructed by the return composition matches the one built by the target composition.

The above observation demonstrates that there are multiple compositions can achieve in constructing the spectrum that are close to the target one. The numerical limitation helps the LP solver to converge to a unique optimum solution. The reason for Case 1 to return a composition that matches the target one, is that the spectral information applied to the LP model is competent. The constraints constructed in the LP model for Case 1 eventually converge to the target composition.

In order to add necessary information to construct the constraints in our LP model, IR's second polarization is introduced to the simplified molecular model: the x polarization. Figure 3.4 describes how the x-polarized spectra presented for 10

Test Case index	3
Number of Candidates	10
Candidates	[0, 10, 20, 30, 40, 50, 60, 70, 80, 90]
Target Composition	[0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0]
Number of Data Points	100, z
Return Composition	[0, 0, 0.73, 0, 0.21,0, 0, 0.057, 0, 0]

Table 3.2: Test case 3 for the simplified molecular model.

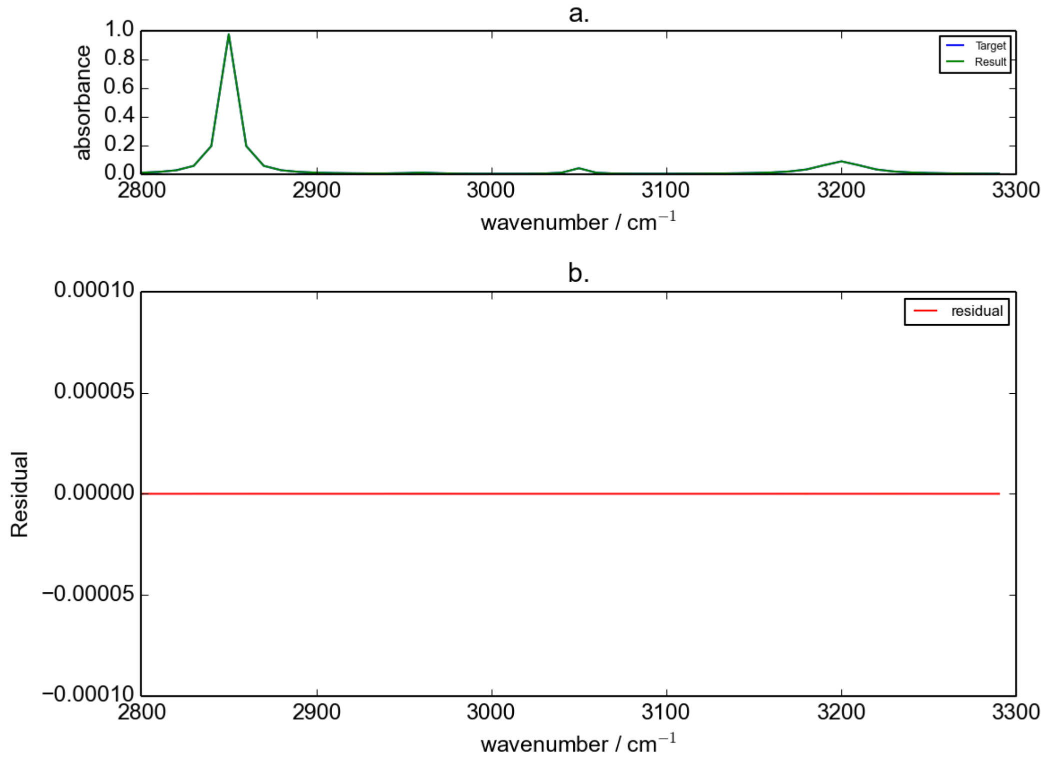


Figure 3.2: a. z-polarized IR target spectrum plotted with the one constructed by the return composition in Case 2 of simplified molecular model; b. The residual plot between the spectra.

candidates. Case 4 and 5 include both polarizations' spectral information in the LP model. In Table 3.3, Case 4 is based on Case 2, with x-polarized IR spectral information added. 100 data points are selected from this additional spectrum, then converted to additional decision variables and constraints in the LP model. Case 5 is based on Case 3, with x-polarized IR spectral information added. In both Case 4 and 5, the return composition matches the target one. This further demonstrates that as long as we have sufficing information for the LP model, the LP solver returns a composition matches the target one.

3.5 Constraint Study Based on Test Case 4

From Cases 1 to 5 for our simplified molecular model, we know that having an instance with sufficient information as input to our LP model is the key to obtain the

Test Case index	4	5
Number of Candidates	4	10
Candidates	[0, 5, 10, 15]	[0, 10, 20, 30, 40, 50, 60, 70, 80, 90]
Target Composition	[0.1, 0.5, 0.4, 0]	[0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0]
Number of Data Points	100, z 100, x	100, z 100, x
Return Composition	[0.1, 0.5, 0.4, 0]	[0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0]

Table 3.3: Test cases 4 and 5 for the simplified molecular model.

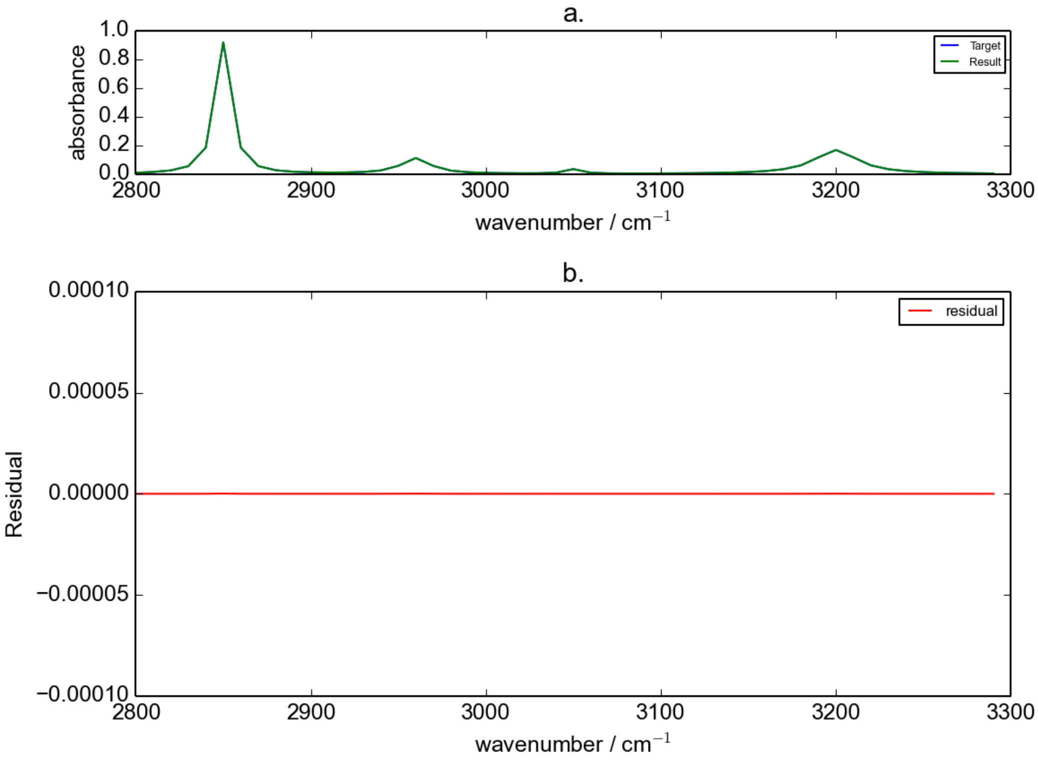


Figure 3.3: a. z-polarized IR target spectrum plotted with the one constructed by the return composition in Case 3 of simplified molecular model; b. The residual plot between the two spectra.

target composition. Having sufficient information means having enough constraints to help to converge to the desired target composition. The information stems from the data points selected along the spectra. This leads us to do a more detailed study on the constraints in order to see how many data points are enough to get the target composition.

Based on Case 4, test cases about creating different LP instances using different spectral information are designed in Table 3.4. The number of data points indicates how many data points are selected. Points Selection shows how data points are selected. For example, [2800, 3300, 50] means along wavenumber from 2500 to 3300, every 50 wavenumber a data point is selected along a spectrum. z and x indicate the corresponding polarization of IR spectrum.

As Table 3.4 indicates, the return compositions in Cases 6 to 14 do not return

Test Case #	# Data Points	Points Selection	Return Composition
6	10	[2800, 3300, 50], z	[0, 0.8, 0.10, 0.1]
7	20	[2800, 3300, 25], z	[0, 0.8, 0.10, 0.1
8	25	[2800, 3300, 20], z	[0, 0.8, 0.10, 0.1]
9	32	[2800, 3300, 15], z	[0, 0.8, 0.10, 0.1]
10	50	[2800, 3300, 10], z	[0, 0.8, 0.10, 0.1]
11	100	[2800, 3300, 5], z	[0, 0.8, 0.10, 0.1]
12	$100 + 1$	[2800, 3300, 5], z [2800, 3300, 500], x	[0, 0.8, 0.10, 0.1]
13	$100 + 5$	[2800, 3300, 20], z [2800, 3300, 100], x	[0, 0.8, 0.10, 0.1]
14	$100 + 10$	[2800, 3300, 20], z [2800, 3300, 50], x	[0, 0.8, 0.10, 0.1]
15	$100 + 50$	[2800, 3300, 20], z [2800, 3300, 10], x	[0.1, 0.5, 0.4, 0]
16	$100 + 100$	[2800, 3300, 20], z [2800, 3300, 5], x	[0.1, 0.5, 0.4, 0]

Table 3.4: Constraint study based on Case 4 for the simplified molecular model. For more detailed result data, refer to Table A.3.

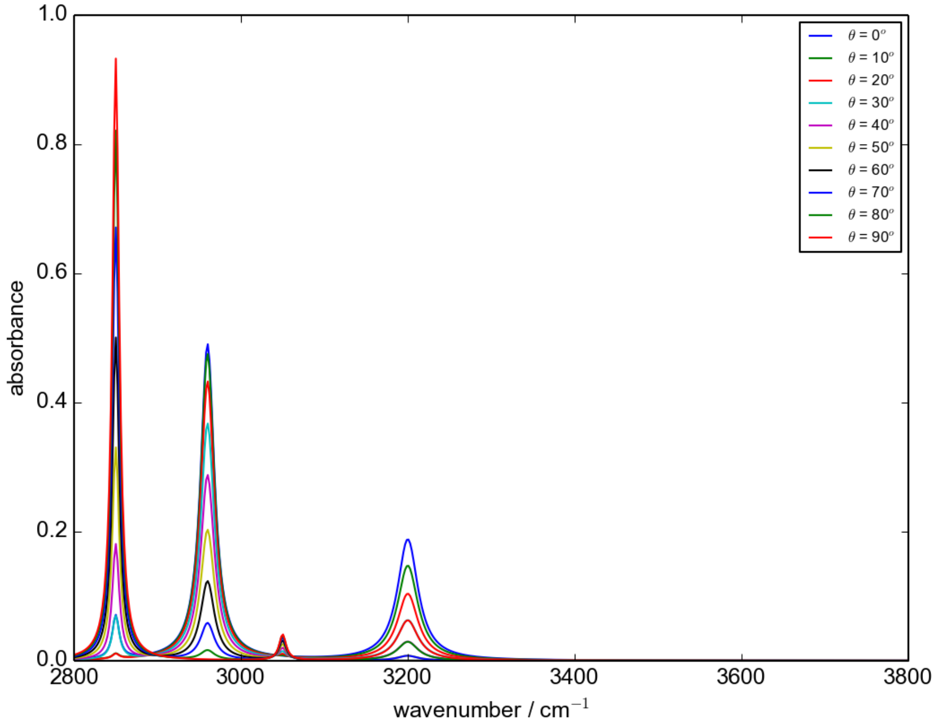


Figure 3.4: x-polaried IR spectra of candidates of simplified molecular model with θ value expanded from 0° to 90°.

the target compostion. To the contrary, in Cases 15 to 16, the return composition matches the target one. Figure 3.5 displays the spectra conducted by $[0, 0.80, 0.1, 0.1]$ and $[0.1, 0.5, 0.4, 0]$, both x- and z-polarized IR spectra generated by these two compositions are identical.

3.6 Constraint Study Based on Test Case 5

Based on Case 5, similar constraint study is conducted as displayed in Table 3.5, and the same observation is obtained as the test cases in Table 3.4. When the result composition $[0, 0, 0.73, 0, 0.21, 0, 0, 0.057, 0, 0]$ and target one are used to plot the spectra, the produced spectra are identical, as shown in Figure 3.6. Although these two constraint studies do not give a clear answer about how many data points are enough to get the target composition, it confirms that as long as the spectral information is sufficient, the LP solver will return the target composition.

Test Case #	# of Data Points	Point Selection	Return Composition
17	10	[2800, 3300, 50], z	[0.16, 0, 0, 0.83, 0, 0, 0, 0, 0, 0.017]
18	25	[2800, 3300, 20], z	[0, 0, 0.73, 0, 0.21, 0, 0, 0.057, 0, 0, 0]
19	50	[2800, 3300, 10], z	[0, 0, 0.73, 0, 0.21, 0, 0, 0.057, 0, 0, 0]
20	100	[2800, 3300, 5], z	[0, 0, 0.73, 0, 0.21, 0, 0, 0.057, 0, 0, 0]
21	500	[2800, 3300, 1], z	[0, 0, 0.73, 0, 0.21, 0, 0, 0.057, 0, 0, 0]
22	$100 + 1$	[2800, 3300, 5], z [2800, 3300, 500], x	[0, 0, 0.73, 0, 0.21, 0, 0, 0.057, 0, 0, 0]
23	$100 + 10$	[2800, 3300, 5], z [2800, 3300, 50], x	[0.36, 0, 0.31, 0.33, 0, 0, 0, 0, 0]
24	$100 + 20$	[2800, 3300, 5], z [2800, 3300, 25], x	[0.17, 0, 0, 0.79, 0, 0, 0.035, 0, 0, 0]
25	$100 + 25$	[2800, 3300, 20], z [2800, 3300, 20], x	[0.17, 0, 0, 0.79, 0, 0, 0.035, 0, 0, 0]
26	$100 + 50$	[2800, 3300, 5], z [2800, 3300, 10], x	[0, 0, 0.75, 0, 0.15, 0, 0.1, 0, 0, 0]
27	$100 + 84$	[2800, 3300, 5], z [2800, 3300, 6], x	[0.17, 0, 0, 0.79, 0, 0, 0.035, 0, 0, 0]
28	$100 + 100$	[2800, 3300, 5], z [2800, 3300, 5], x	[0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0]

Table 3.5: Constraint study based on Case 5 of simplified molecular model. For more detailed result data, refer to Table A.4.

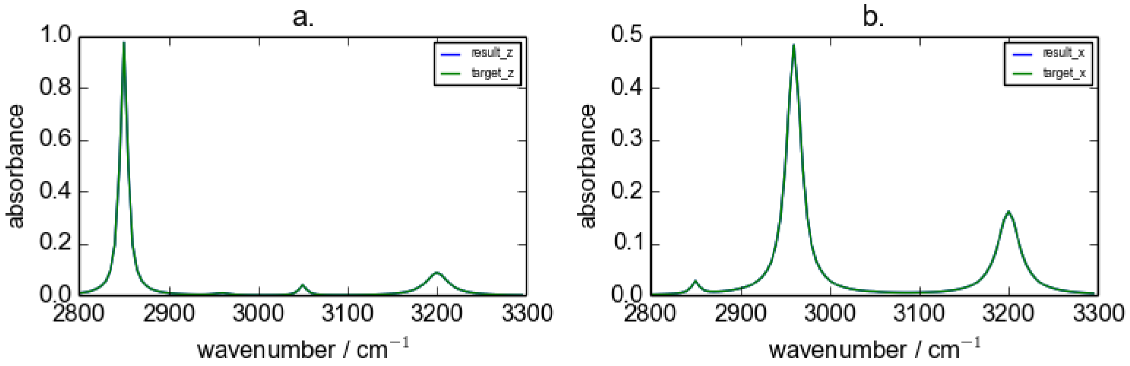


Figure 3.5: IR spectra plotted by the return compositions from the constraint study based on Case 4 of simplified molecular model. a. z-polarized IR spectra; b. x-polarized IR spectra.

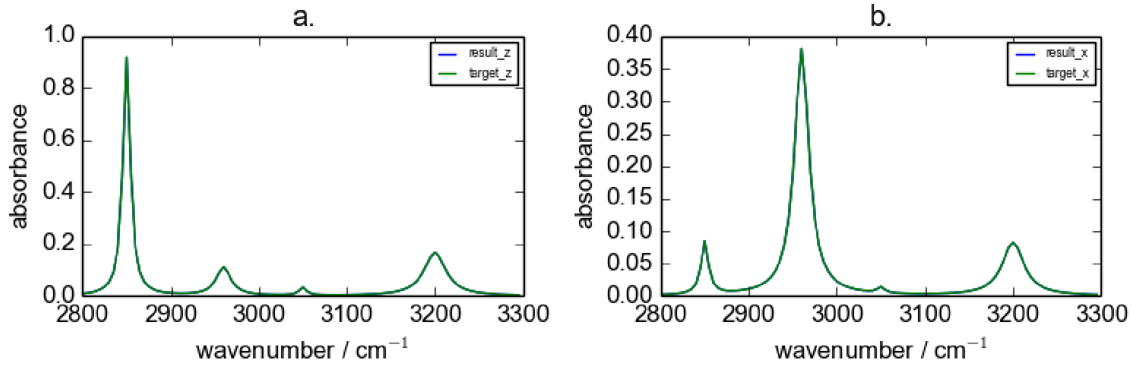


Figure 3.6: IR spectra plotted by the return compositions from the constraint study based on Case 5 of simplified molecular model. a. z-polarized IR spectra; b. x-polarized IR spectra.

3.7 Discussion and Conclusion

Recall that our LP model, for sufficient data sets are expected to return the target composition. We can conclude that, if the target composition is not returned correctly, then the data we collect is not sufficient to describe the test cases to the LP model.

However, when the target composition is not returned correctly, the return composition does build spectra that are identical to the target ones. This means that there is more than one composition that can build the spectra that are identical to the target ones.

With the help of the simplified molecular model, we know the reason why the LP model cannot return the target composition in some test cases. In the next step, we want to figure out with all the spectral information available for a realistic molecular model, can the instances of our LP model return the target composition?

Chapter 4

Realistic Molecular Model

4.1 Description

From experimenting with the simplified molecular model, we learnt that lacking sufficient spectral information is the key cause for the failure of obtaining the target composition. First of all, in the simplified molecular model, there are only four vibrational modes, and thus the spectral information is limited. Secondly, the similarity among the candidates is high, as all the candidates are coming from one same molecule. Third, only IR spectra is considered.

In test cases discussed this chapter are conducted using realistic molecules. In addition to IR, both Raman and SFG spectra are calculated for these molecules, which makes the study one step closer to the overall goal and scope. The realistic molecule focused on this chapter is Met amino acid.

Same as with the simplified molecular model, in order to limit the possible candidate space of Met, *twist* and *azimuthal* angular distributions are assumed to be isotropic, which are integrated into uniform distribution. Only Euler angle θ is considered in Met's surface orientation distribution function. In Section 2.3, we explained how a molecule's IR, Raman and SFG spectra are generated. Two unique IR spectra can be obtained from x-, and z-polarizations. Four unique Raman spectra can be obtained from xx-, xy-, xz- and zz-polarizations. Three unique SFG spectra can be obtained from xxz-, xzx- and zzz-polarizations.

Test Case #	1	2
# Candidates	4	4
Candidates	[0, 20, 40, 60]	[0, 20, 40, 60]
Target Composition	[0.1, 0.5, 0.4, 0]	[0.1, 0.5, 0.4, 0]
# Data Points	200, x	200, z
Return Composition	[0.70, 0, 0, 0.30]	[0.70, 0, 0, 0.30]

Table 4.1: Test Case 1 and 2 for Met candidates.

Test Case #	3	4
# Candidates	4	4
Candidates	[0, 20, 40, 60]	[0, 20, 40, 60]
Target Composition	[0.1, 0.5, 0.4, 0]	[0.1, 0.5, 0.4, 0]
# Data Points	200, x 200, z	200, x 200, xx
Return Composition	[0.70, 0, 0, 0.30]	[0.1, 0.5, 0.4, 0]

Table 4.2: Test Case 3 and 4 for Met candidates.

The goal is to check with all these spectral information available, is it sufficient for the LP solver to help the corresponding LP instances to return the target composition of the candidates of Met at a surface. If yes, we need to figure out which spectral information is sufficient. If no, we need to check if the cause of the failure is the same as in the case of the simplified molecular model.

4.2 Test Cases

In Table 4.1 and 4.2, four test cases are set up with four candidates and one same target composition. These four candidates have the following θ values: 0°, 20°, 40° and 60°. The only difference among these four test cases is the spectroscopy information we select to build the LP instances, and this is indicated by the Number of Data Points. In Case 1, only x-polarized IR spectral information is used. This means that only data points from x-polarized IR are selected as input to the LP model. Accordingly for Case 2, data points are obtained from spectra of IR's z-polarized IR. In Test Case 3, the spectral information of x- and z-polarized IR is combined. At last, in Case 4, spectral information of x-polarized IR and xx-polarized Raman are combined. Case 4 contains the most abundant spectral information, as its return composition matches the target one.

When merely using IR information, the return composition is the same in Case 1, 2 and 3. Figure 4.1 displays the resulting spectra generated by using the return composition obtained from the first three test cases. The resulting spectra is almost identical to the target ones. It indicates that with only IR spectral information is not sufficient to get the target composition. However, the spectra built by the return composition matches the target spectra. This means that further information is needed to build the constraints of the LP model. The more valid constraints are introduced, the more accurate the return composition will be.

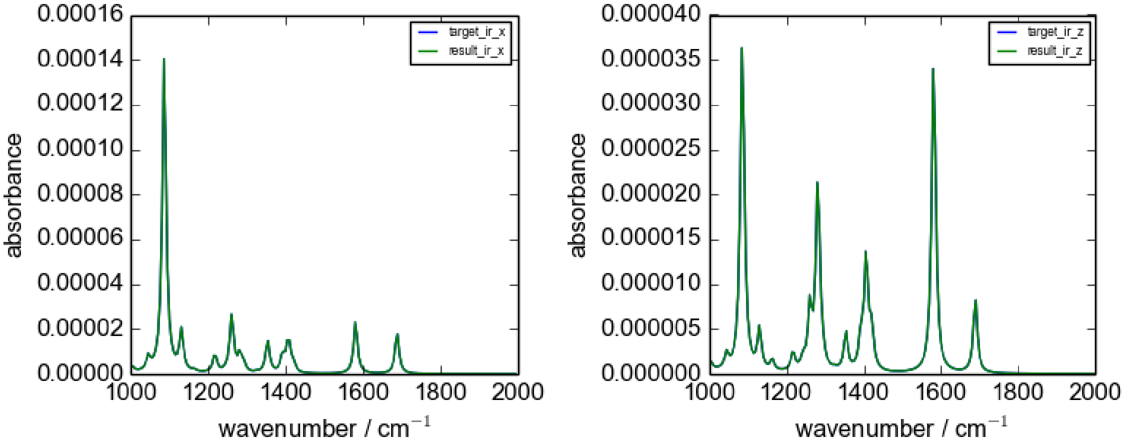


Figure 4.1: Comparing target IR spectra with the ones generated by the return composition of Cases 1, 2 and 3.

In Case 4, combining the spectral information of IR and Raman is sufficient to obtain the target composition. When the difference in degree θ for candidates decreases from 20° to 10°, understanding if Raman and IR together is still sufficient to derive the target composition is desired. Therefore, the following test cases shown in Table 4.3 are conducted.

Case 5 shows that the LP model with instance built by merely using IR spectral information is not sufficient to derive the target composition. Case 6 indicates that combining IR and Raman spectral information helps to derive the target composi-

# Candidates	4	
Candidates	[0, 10, 20, 30]	
Target Composition	[0.1, 0.5, 0.4, 0]	
Test Case index	# Data Points	Result Composition
5	200, x 200, z	[0.75, 0, 0, 0.23]
6	200, x 200, z 200, xx	[0.1, 0.5, 0.4, 0]
7	200, xx 200, xy 200, xz	[0.1, 0.5, 0.4, 0]
8	200, xx 200, xy 200, zz	[0.1, 0.5, 0.4, 0]
9	200, xx 200, xy 200, xz 200, zz	[0.1, 0.5, 0.4, 0]

Table 4.3: Test case 5 to 9 for Met candidates.

tion. Case 7, 8 and 9, illustrate that Raman spectral information itself is sufficient to obtain the target composition.

For test cases in Table 4.1, 4.2 and 4.3, combining IR and Raman spectral information to build an LP instance is sufficient enough to obtain the target composition. In order to further study the limitation of the LP model, the complexity of the test case needs to be increased. Therefore, another group of test cases is designed as shown in Table 4.4. There are five candidates included in the test cases. Each candidate has θ with the following degree: 0°, 10°, 20°, 30° and 40°. The target composition is more complex than previous test cases, each candidate takes 20% in the mixture.

Case 10 uses only IR spectral information to build the LP instance, and the return composition does not match the target one. Case 11 uses only Raman spectral information, and the return composition does not match the target neither. Same for Case 12 that uses only SFG spectral information. From Case 13, different kinds of spectral information are combined. In Case 13, IR and Raman spectral information is used to produce the LP model, still the return composition is different from the target one. Case 14 combines Raman and SFG, Case 15 uses IR and SFG, Case 16 cooperates all the three spectral information, however, none of them returns a composition that matches the target one.

The results of Cases 10 to 16 indicate that despite combining all the spectral information of IR, Raman and SFG, it is still not sufficient to attain the target composition for the test cases set up in Table 4.4. The spectral information we apply to the LP model is showing its limitation in these test cases. In order to confirm the reason causing the LP solver to return the target composition due to insufficient information, further test cases are conducted in Table 4.5.

Number of Candidates	5	
Candidates	[0, 10, 20, 30, 40]	
Target Composition	[0.2, 0.2, 0.2, 0.2, 0.2]	
Test case index	Constraints	Result
10	200, x 200, z	[0.61, 0, 0, 0, 0.40]
11	200, xx 200, xy 200, xz 200, zz	[0.25, 0, 0.50, 0, 0.25]
12	200, xxz 200, xzx 200, zzz	[0.32, 0, 0.31, 0.16, 0.21]
13	200, x 200, z 200, xx 200, xy 200, xz 200, zz	[0.25, 0, 0.50, 0, 0.25]
14	200, xx 200, xy 200, xz 200, zz 200, xxz 200, xzx 200, zzz	[0.32, 0, 0.31, 0.16, 0.21]
15	200, x 200, z 200, xxz 200, xzx 200, zzz	[0.32, 0, 0.31, 0.16, 0.21]
16	200, x 200, z 200, xx 200, xy 200, xz 200, zz 200, xxz 200, xzx 200, zzz	[0.32, 0, 0.31, 0.16, 0.2]

Table 4.4: Test Case 10 to 16 for Met candidates. For more detailed result data refer to Table A.1.

4.3 Test Cases to Explain the Limitation of our LP Model for instances obtained for the Met Molecule

To further explore the reasons when our LP model reaches its limitation for the realistic molecule, Cases 17 and 18 are conducted. To make the est case more general than Cases 1 to 16, candidates' θ values are expanded from 0° to 80°. In total, there are nine candidates. Because the SFG spectra for θ of 90° is a straight line, it is excluded from all the test cases related to realistic molecules. As target compositions, five candidates are randomly selected. The difference between Case 17 and 18 is that different amount of data points are selected to build the instances of our LP model. From all three spectroscopy techniques' spectral information, every 5^{th} wavenumber a data point is selected for Case 17. Every 500^{th} wavenumber a data point is selected for Case 18. As a result, Case 17 and 18 each returns a different composition. Both compositions do not match the target one.

However, in both Case 17 and 18, when the return composition is used to generate the IR, Raman and SFG spectra, these spectra are plotted together with the spectra created by the target composition. Note that all spectral data are identical for IR, Raman and SFG. Figures 4.2, 4.3 and 4.4 display the spectra plotted by using the return composition and the target one of Case 17. All spectra is almost identical to each other as shown in the figures. The same is true for Case 18, as shown in Figures 4.5, 4.6 and 4.7. These figures show that there is more than one composition that can perfectly construct the target spectra. The data information used to construct the instances of our LP model is not sufficient to converge to the return composition that exactly matches the target one. This conclusion exactly fits the result obtained from the test cases we have done with the simplified molecular model.

4.4 Conclusion

With all the test cases we have run with Met, we figure out that even combine all the available spectral information to the LP model, it is not guaranteed to return the target composition. The reason is the same as applying spectral information of

# Candidates	9	
Candidates	[0, 10, 20, 30, 40, 50, 60, 70, 80]	
Target Composition	[0.22, 0.29, 0.052, 0.083, 0.36, 0, 0, 0, 0]	
Test Case #	# of Data Points	Result Composition
17	each 5 wavenumber of IR, Raman and SFG spectra	[0.16, 0.39, 0.0, 0.099, 0.35, 0.0, 0.0, 0.0, 0.0]
18	each 500 wavenumber of IR, Raman and SFG spectra	[0.40, 0.0, 0.20, 0.036, 0.36, 0.0, 0.0, 0.0, 0.0]

Table 4.5: Test case 17 and 18 to explain the limitation of our LP model for Met molecule. For more detailed result data refer to Table A.2.

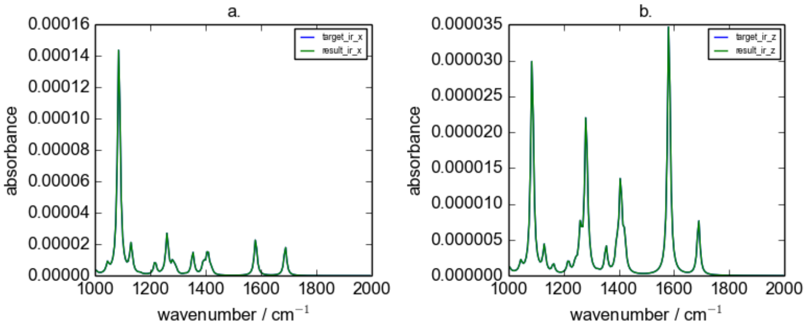


Figure 4.2: IR spectra plotted by using target composition and return composition of Case 17. a. x-polarized IR spectra; b. z-polarized IR spectra.

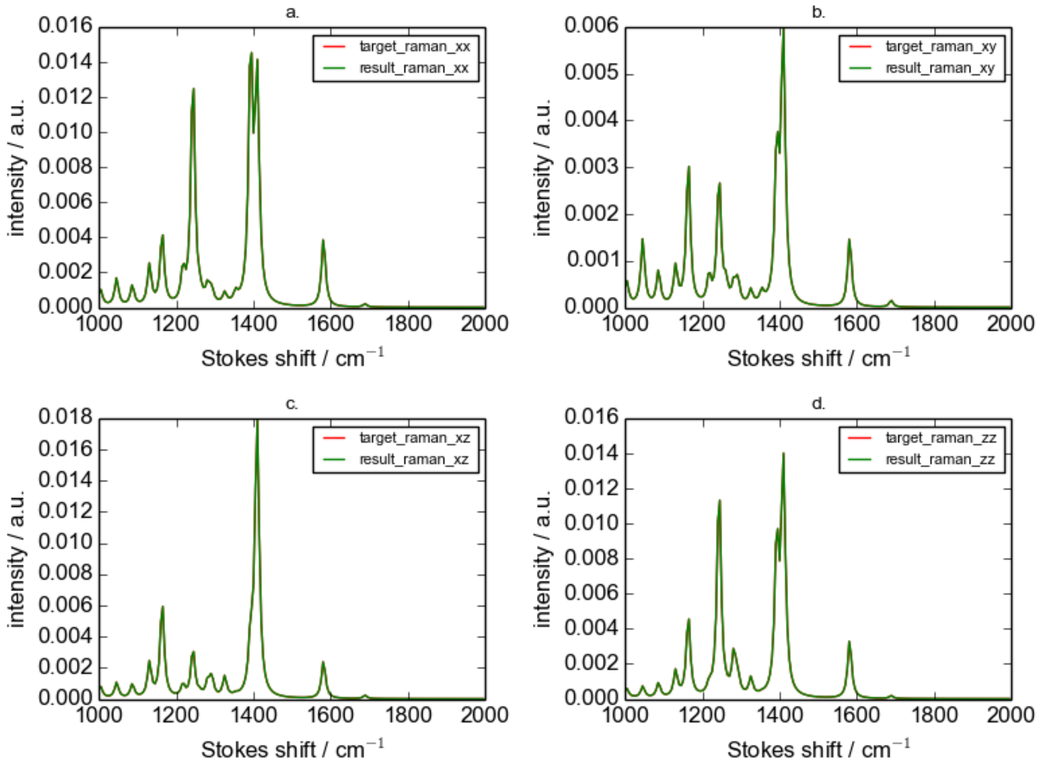


Figure 4.3: Raman spectra plotted by using the target composition and the return composition of Case 17. a. xx-polarized Raman spectra; b. xy-polarized Raman spectra; c. xz-polarized Raman spectra; b. zz-polarized Raman spectra.

the simplified molecular model to the LP model. The spectral information is not sufficient for the LP model in order to obtain the desired target composition. The spectra constructed by the return composition is identical to the target spectra.

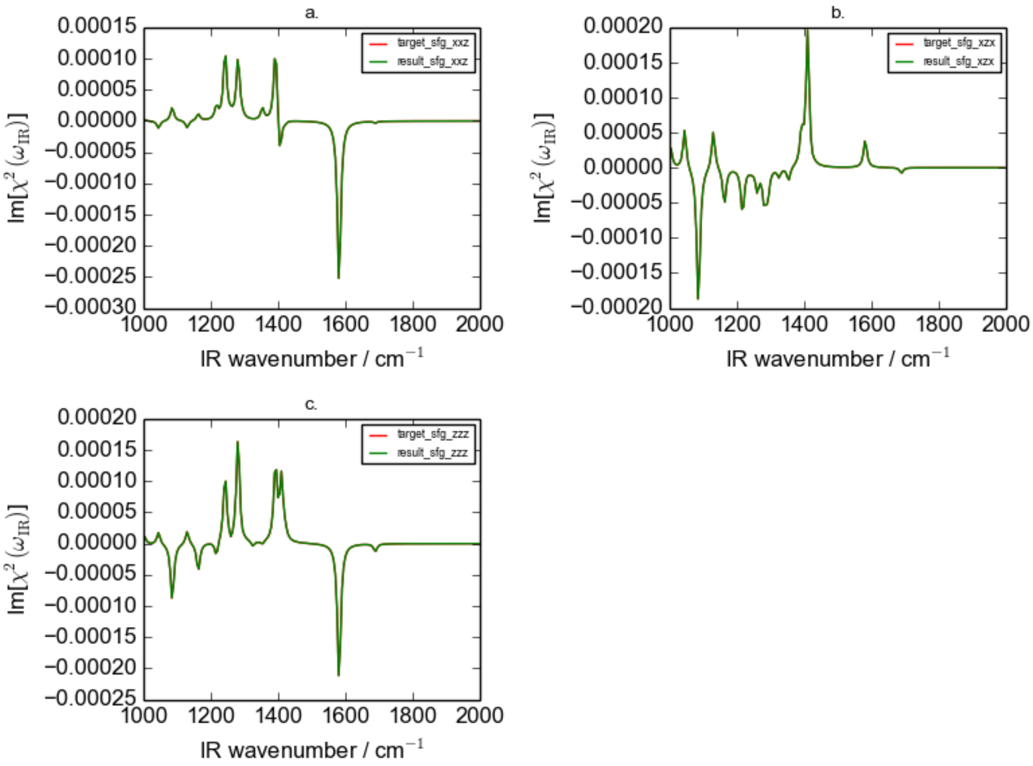


Figure 4.4: SFG spectra plotted by using the target composition and the return composition of Case 17. a. xxz-polarized SFG spectra; b. xzx-polarized SFG spectra; c. zzz-polarized SFG spectra.

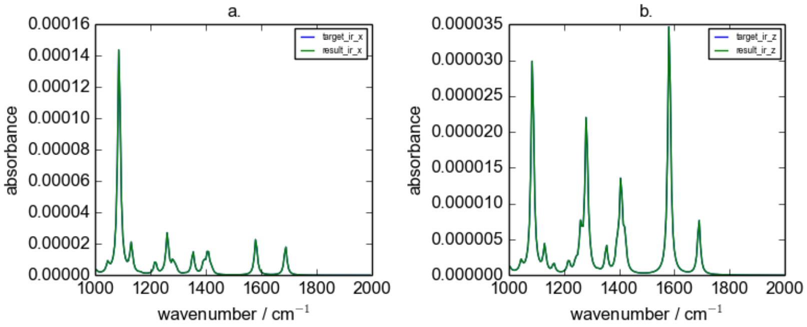


Figure 4.5: IR spectra plotted by using the target composition and the return composition of Case 18. a. x-polarized IR spectra; b. z-polarized IR spectra.

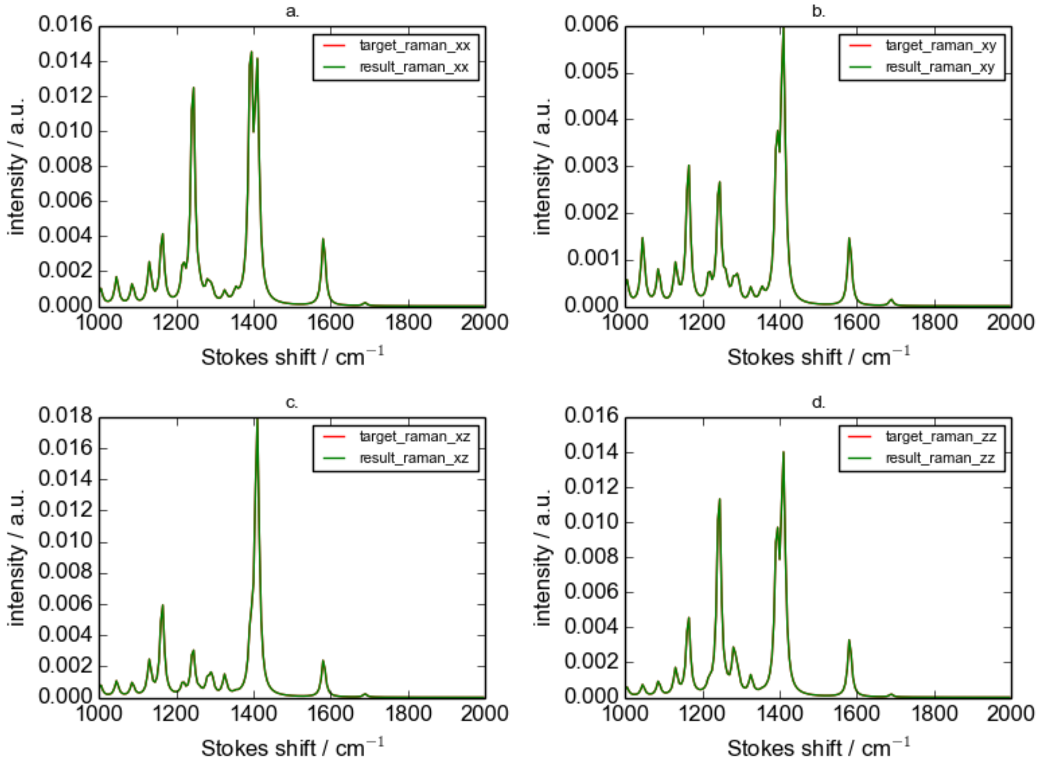


Figure 4.6: Raman spectra plotted by using the target composition and the return composition of Case 18. a. xx-polarized Raman spectra; b. xy-polarized Raman spectra; c. xz-polarized Raman spectra; b. zz-polarized Raman spectra.

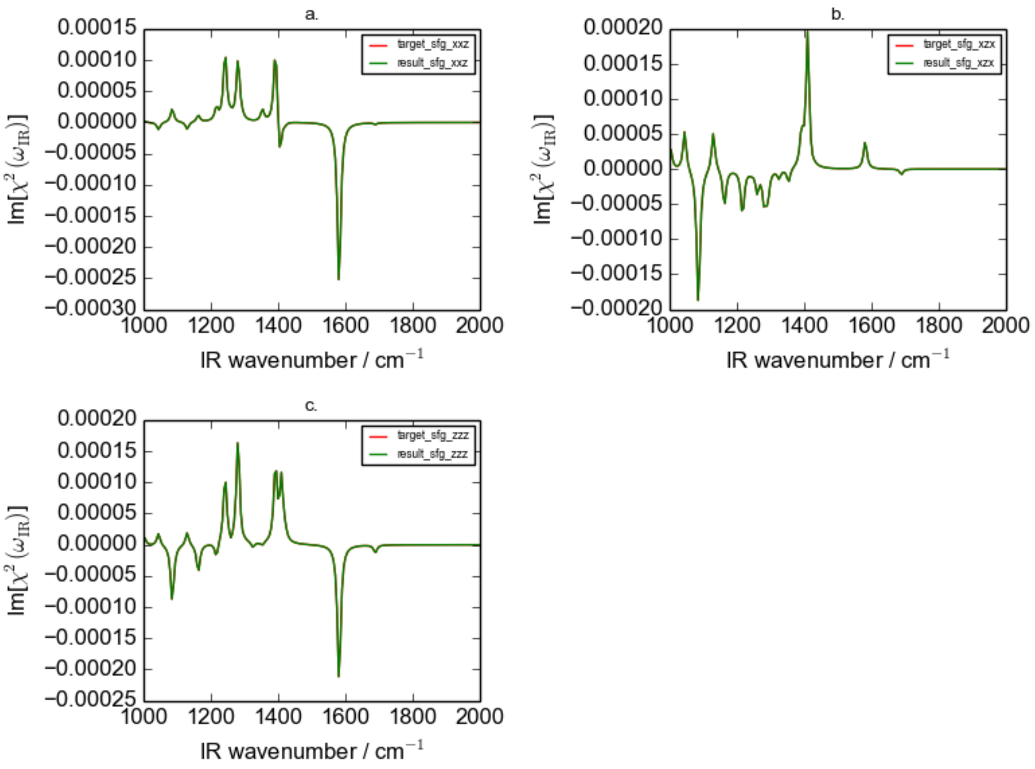


Figure 4.7: SFG spectra plotted by using the target composition and the return composition of Case 18. a. xxz-polarized SFG spectra; b. xzx-polarized SFG spectra; c. zzz-polarized SFG spectra.

Chapter 5

Mixture of Realistic Molecules

5.1 Description

In Chapter 4, test cases indicate that for one type of molecule at surfaces, even when combining the information of all the three spectral information, the built LP instances are still not sufficient to obtain the target composition in most test cases. In another word, the existing spectral information is not adequate to obtain the target composition of one type of molecule at surfaces. Multiple return compositions can build the target spectra. Besides one type of molecule at surfaces, we are also interested in the case where candidates coming from different molecules. For a mixture of different molecules at surfaces, we want to figure out with available spectral information, can the LP model help to return the target composition. In the case where the LP model is sufficient to compute the target composition, we are interested in which the specific combination of spectroscopy techniques is sufficient. Moreover, we want to know the accuracy of this specific combination in obtaining the target composition.

5.2 Test Cases

The first part of this section, we study the cases where each molecule's candidates expanded from 0° to 80° on θ to see which spectral information is sufficient in obtaining the target composition. Then in the second part, we study the cases where we study the cases where each molecule's candidates expanded from 0° to 180° on θ.

5.2.1 Test Cases Considering Each Amino Acid Candidates from θ 0° to 80° in the Mixture of Realistic Molecules

To study the molecular orientation distribution of various molecules at surfaces, further test cases are constructed. These test cases have the following common settings.

First, as mentioned in Chapter 1, there are six different amino acids in the mixture: Met, Leu, Ile, Ala, Thr and Val. For each amino acid, only the difference of θ is considered. Each amino acid has nine candidates in the mixture. They have the following θ values: 0°, 10°, 20°, 30°, 40°, 50°, 60°, 70° and 80°. When θ equals 90°, the SFG spectra is a straight line. Therefore the corresponding candidate is excluded from all the test cases. As a result, there are 54 candidates in the mixture.

Second, the target composition needs to be generated. The operation includes two steps: randomly pick one candidate from each of the amino acid's nine candidates, then randomly generate a percentage for the selected candidate. The target composition is made of six randomly selected candidates with assigned percentage coming from the amino acids. The remaining 48 candidates have 0 percentage in the target composition. That is six selected candidate makes 100% component of the mixture.

Third, the IR, Raman and SFG spectra need to be generated for all 54 candidates and the target.

Table 5.1 displays a set of test cases. Each test case contains different spectral information. In Case 1, candidates x- and z-polarized IR spectra are obtained. The target's IR spectra are generated by the dot product of the target composition and all the candidates' spectral data. Then the corresponding LP instance is conducted using Equation 3.4. Therefore, we conclude that in Case 1, only IR information is used to build the LP instance. Similarly, in Case 2, only Raman spectral information is used to build the LP instance. In Case 3, only SFG spectral information is used to build the LP instance.

Starting from Case 4, spectral information of different spectroscopy techniques are combined to build the LP instance. In Case 4, IR spectral information is combined with Raman. In Case 5, IR spectral information is combined with SFG. In Case

Test Case Index	Spectral Information
Case 1	x and z polarized IR spectra
Case 2	xx, xy, xz and zz polarized Raman spectra
Case 3	xxz, xzx and zzz polarized SFG spectra
Case 4	x and z polarized IR spectra xx, xy, xz and zz polarized Raman spectra
Case 5	x and z polarized IR spectra xxz, xzx and zzz polarized SFG spectra
Case 6	xx, xy, xz and zz polarized Raman spectra xxz, xzx and zzz polarized SFG spectra
Case 7	x and z polarized IR spectra xx, xy, xz and zz polarized Raman spectra xx, xzx and zzz polarized SFG spectra

Table 5.1: Detailed test cases set for the mixture of amino acids.

6, Raman and SFG spectral information are incorporated. At the end, in Case 7, information of all three spectral is put together: IR, Raman and SFG.

Finally, this set of test cases is run 100 times in order to see which test case in the set returns the target composition with the highest accuracy. This accuracy is measured by the time of the LP solver returns the target composition. The scoring mechanism that measures whether or not a return composition matches the target one is described in the next section.

5.2.2 Scoring method

At the first glance, it may appear a useful approach that the sum of residuals between the spectra composed by the return composition and the target one can be used to measure the accuracy of the return composition. However, recall that in most test cases conducted earlier, the spectra generated by the return composition are identical to the target ones. The sum of residuals between these spectra is negligible, which makes it appropriate to use it as a scoring criterion.

Another way to measure the accuracy of the return composition is to compare it directly with the target composition. Calculating the sum of the residuals between a target composition and a return one directly can be a fast approach to evaluate the accuracy of each test case. The shortage of this approach is that it cannot be used to measure in realistic test cases where the target composition is unknown. However, in the current test cases, this approach can be a way to evaluate the return composition for all the test cases where the target compositions are known in advance.

The return composition of each test case in the set is obtained for each run. Each return composition is compared with the target one to calculate the sum of the residuals. If the sum is smaller than a certain threshold, which is 10^{-7}, then the return composition is considered to be the same as the target one.

The test case set is run 100 times, based on the scoring method, the result is shown in Figure 5.1. Case 2, 3, 4, 5, 6 and 7 return the target composition in all 100 runs. This result indicates that Raman or SFG alone is sufficient to obtain the target

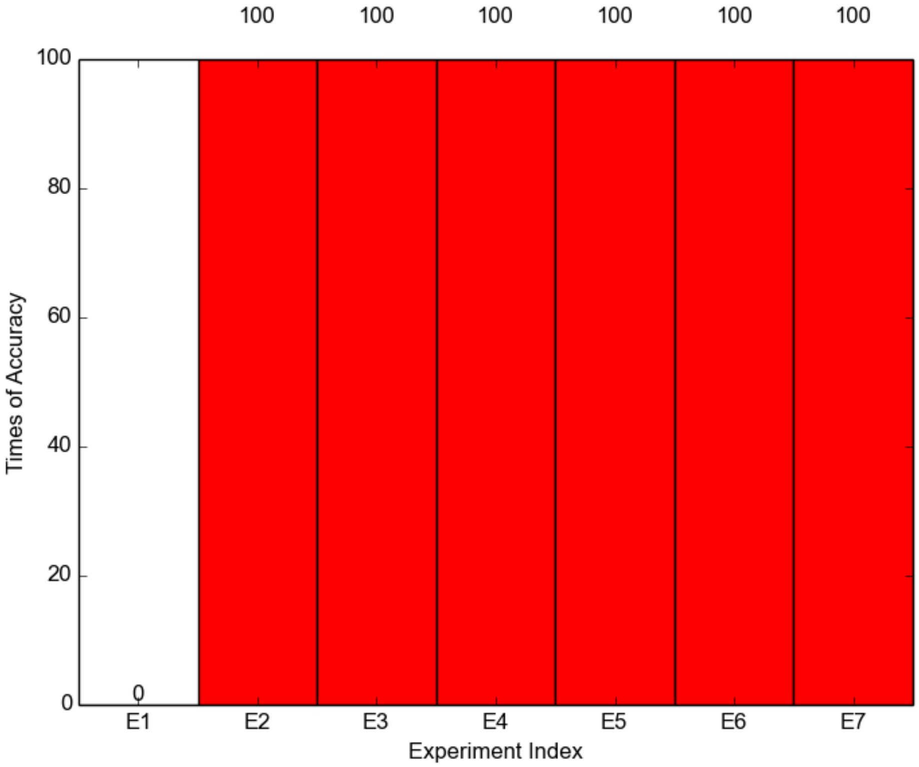


Figure 5.1: Accuracy analysis for test cases considering a mixture of amino acids with candidates from 0° to 80° on θ for each amino acid. Accuracy indicates how many times each test case in the set return a composition that matches the target one.

composition, for a mixture of amino acids with candidates from 0° to 80° on θ for each amino acid. Any test cases that contain Raman and SFG spectral information result in the same accuracy.

The only exception is Case 1. The accuracy is fairly low, which indicates that IR spectra alone do not contain sufficient information in order to obtain the target composition.

To gain more understanding of the return composition of Case 1, the test case set is re-run 100 times. In each run, IR x- and z-polarized spectra are plotted both by the returned composition and the target one. The result is that these spectra conducted by the two different compositions are identical in each run. Let us randomly take one run as an example. Figure 5.2 displays the plotted spectra. Note that they are identical to each other. The residual is very small for the data points where these two spectra are not overlapped. This further indicates that IR spectral information is not sufficient to obtain the target composition.

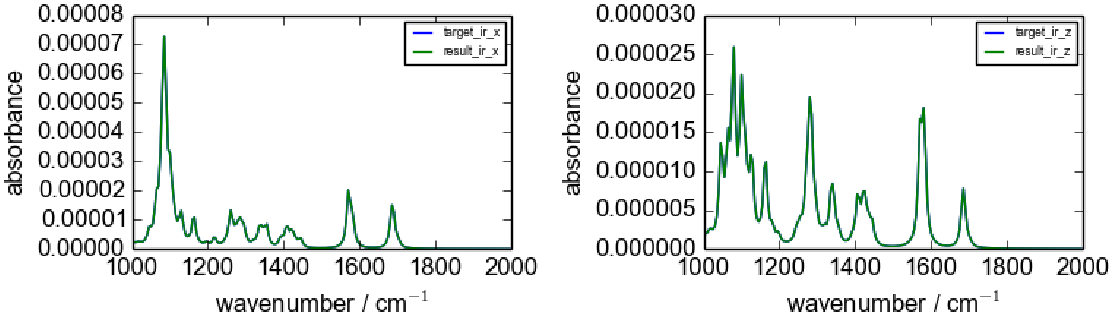


Figure 5.2: IR spectra plotted by the result composition and the target composition of one ramdon run when considering each amino acid candidates from 0° to 80° on θ in the Mixture of realistic molecules.

5.2.3 Test Cases Considering Each Amino Acid Candidates from θ 0° to 180° in the Mixture of Realistic Molecules

To further study the capacity of the LP models, the candidate pool is expanded from 0° to 180° in terms of the θ value. Therefore, each amino acid has 18 candidates. In total, there are 108 candidates in the mixture. The same set of test cases as in Table 5.1 is used. The only difference is that instead of randomly selecting one candidate from nine candidates, it is selected from eighteen. All 108 candidates' IR, Raman and SFG spectra need to be generated. Figure 5.3 illustrates the results obtained in 100 runs. The accuracy in Case 1 is still low. This is not surprising as the complexity of the candidates has increased. Moreover, IR spectra for candidate of θ is identical to the one of θ complement, as shown in Figure A.1. This also increases the difficulty for the LP model to return the target composition.

However, it should be noted that the accuracy for Case 2 has dramatically dropped. This can be caused by the fact that the Raman spectra for one candidate with a θ is identical to the one of θ complement as displayed in Figure A.2.

Figure 5.3 shows that also the accuracy for Case 3 is no longer high. After increasing the number of amino acid candidates from nine to eighteen, the complexity

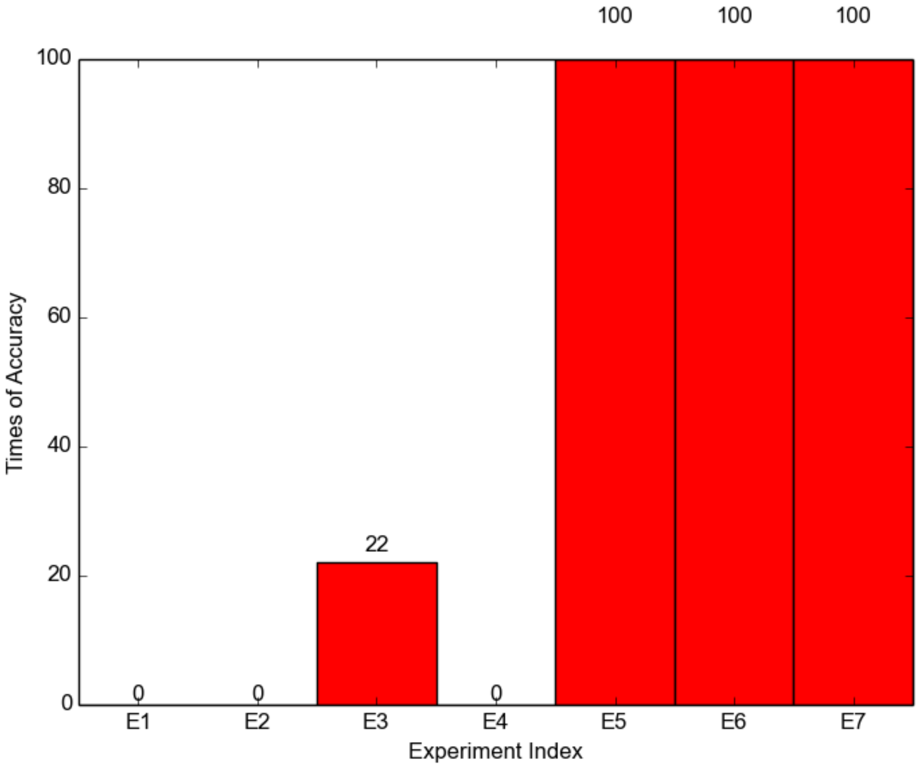


Figure 5.3: Accuracy analysis for test cases considering a mixture of amino acids with candidates from 0° to 180° on θ for each amino acid. Accuracy indicates how many times each test case in the set return a composition matches the target one.

of candidates has increased as well. SFG spectra from candidate of θ is symmetric to the one of θ complement, which has increased the difference of the candidates as shown in Figure A.3. However, the SFG spectral information is still not sufficient for obtaining the target composition.

The good result starts to emerge when using the combinations of IR and SFG or Raman and SFG. Figure 5.3 shows that Cases 5, 6, and 7 all have 100% accuracies. SFG spectra is needed as it is the only information that distinguish θ from its complement. Once combining SFG spectral information with other technique to obtain extra spectral information, it is sufficient to obtain the target composition.

Although the accuracy in Case 2 is low, there is still some noticable result in the return composition: for each amino acid, the percentage assigned is correct; however, the candidate presented may not always be correct. We observe that it is either the correct θ, or its complementary. We randomly select one run of the test case set as an example, Figure 5.4 displays the target composition. Figure 5.5 displays the return composition of Case 2. Figure 5.6 is the return composition of Case 6. Figure 5.4 and 5.6 are identical, which means the return composition of Case 6 is the same as the target one. The values in Figure 5.5 are the same as Figure 5.4. However, the position of each value is not the same in two the figures. For example, the percentage value 0.30 of Met is for $\theta = 120°$ in Figure 5.4, but is for $\theta = 60°$ in Figure 5.5. These two angles are complementary. This observation is the same for Ile, Ala, Thr, and Val in the figures. This observation is a general case across all the runs of the test case set. The return composition of Case 6 matches the target one. However, the return composition of Case 2 fails to pick the correct candidate of each amino acid from this candidate's θ complementary. This can be explained as the Raman spectra for a specific value of θ is the same as its complement.

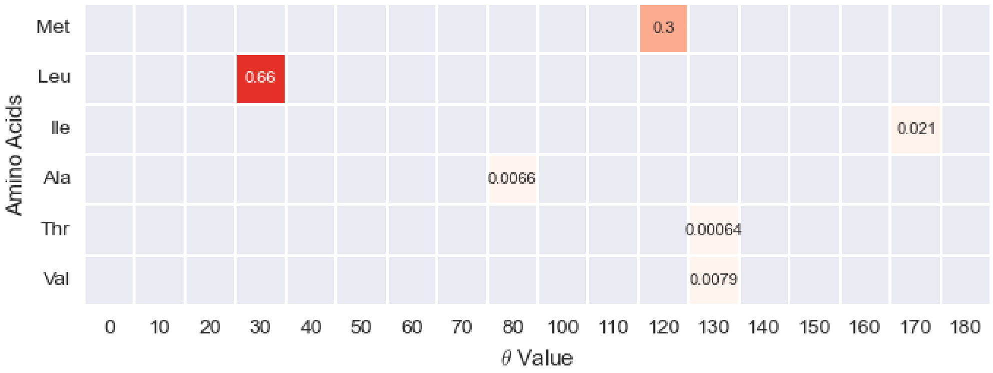


Figure 5.4: Target composition of one random run of six mixed amino acids with candidates expanded from 0° to 180° on θ for each amino acid. More detailed data of this target composition can be found in Appendix A.1.

5.3 Conclusion

Raman and SFG spectral information each alone is sufficient to obtain the target composition, when considering a mixture of molecules with candidates expanded from 0° to 80° on θ for each amino acid. When the candidates of each molecule are expanded from 0° to 180° on θ, SFG spectral information needs to be combined with IR or Raman in order to obtain the target composition. SFG spectral information is needed,

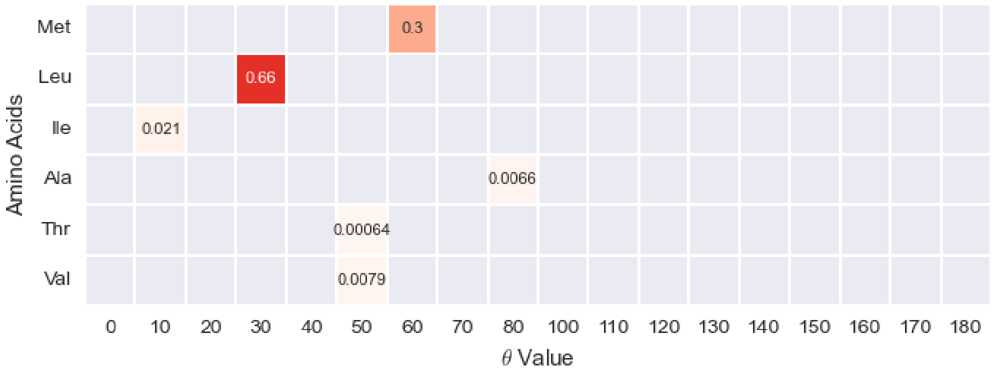


Figure 5.5: Return composition of Case 2 for one random run of six mixed amino acids with candidates expanded from 0° to 180° on θ. More detailed data of this return composition can be found in Appendix A.2.

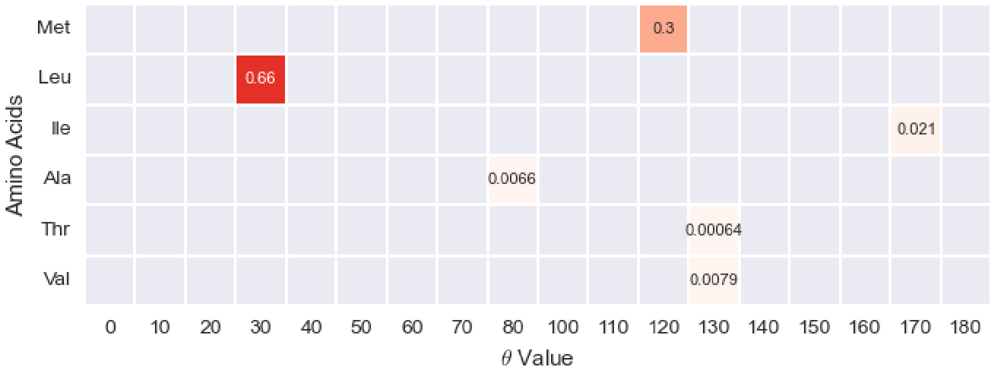


Figure 5.6: Return composition of Case 6 for one random run of six mixed amino acids with candidates expanded from 0° to 180° on θ. More detailed data of this return composition can be found in Appendix A.3.

as it is the only information to distinguish candidate of θ from its complement.

Chapter 6

Possibilities for Treating Experimental Data

6.1 Description

The experimental spectra obtained from IR, Raman or SFG techniques have an amplitude scaling factor when compared to the candidate spectra generated mathematically. This means that between candidates' theoretical spectra and the experimental one, there is an unknown scaling factor. Within one particular spectroscopy technique, the scaling factor is the same for all the polarizations. Take IR as an example. The scaling factor for the spectrum of x-polarization is the same as the one for the spectrum of z-polarization. Therefore, it is necessary to consider this scaling factor when applying the spectral information to the LP model.

6.2 Test Case

The first part of this section, we study the cases where each molecule's candidates expanded from $0°$ to $80°$ on θ to see which spectral information is sufficient in obtaining the target composition when scaling factor is considered. Then in the second part, we study the cases where we study the cases where each molecule's candidates expanded from $0°$ to $180°$ on θ.

6.2.1 Test cases with Scaling Factor Considering Each Amino Acid Candidates from $0°$ to $80°$ on θ in the Mixture

In Chapter 5, the instances of our LP model constructed by Cases 2 to 7 in Table 5.1 for θ ranged from 0° to 80° do well in retrieving the target composition for the mixed amino acids. Therefore, based on these test cases, we investigate whether the LP instances built by using the real experimental data return the target composition.

Therefore, the same test case settings in Table 5.1 are used for the following test cases. The goal is the same, that is to figure out which spectral information is sufficient to retrieve the target composition for the mixture of six amino acids' candidates. The only difference is that, in each run of the test case set, an arbitrary scaling factor is generated for IR, Raman and SFG, respectively. Therefore, the target spectra are not only composed by the target composition of all candidates, but also need to multiple by the randomly generated scaling factors of each spectroscopy technique.

To start with, we limit the scaling factors to be smaller than 1.

After a few runs of the test case set, it is observed that the returned compositions always contains one extra variable in every test case. For Cases 2, 4, 6 and 7, the returned composition contains the correct selected candidates. However, the percentage values of the selected candidates are different from the target composition. The ratio between the returned percentage and the target percentage is the same for all the selected candidates. Furthermore, we add this ratio to the extra variable which equals 1. We randomly select one test case run as an example. Figure 6.2 displays the target composition, only the selected candidates are annotated with assigned percentage. Figure 6.2 displays the return composition of Case 2. The selected candidates in the return composition are correct. However, each percentage value is different from the one in the target composition. There is one extra value in Figure 6.2 with a value of 0.4.

Moreover, Equation 6.1 shows the ratio between the percentage of the selected candidates in the return composition and the target one is the same for all the amino acids (more detailed calculated can be found in Appendix A.9). The value of this ratio is 0.6. When this ratio is added up with the extra variable (referred to as slack

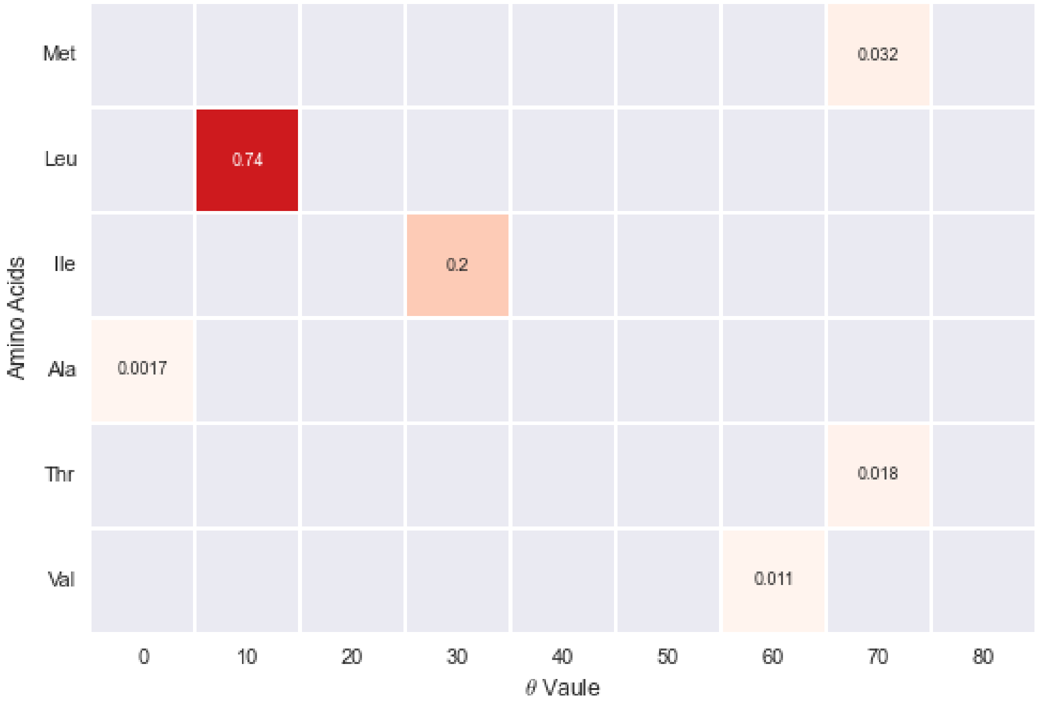


Figure 6.1: Target composition for one random run of the test case set with scaling factor for mixed amino acids, with θ expanded from 0° to 80°. More detailed data of this target composition can be found in Appendix A.4.

variable (SV) in LP) 0.4, the total is 1. As the scaling factors are pre-generated in the test case set, the value is known, which is 0.6 for Raman spectra. In conclusion, the SV is returned by LP. Then the scaling factor (SF) equals to $1 - SV$. From the scaling factor, the ratio between the return composition and the target one is known. At the end, the target composition can be re-built from the ratio and the return composition. The re-constructed target composition matches the original one.

$$\frac{0.019}{0.032} = \frac{0.44}{0.74} = \frac{0.12}{0.2} = \frac{0.001}{0.0017} = \frac{0.011}{0.018} = \frac{0.0067}{0.011} = 0.6 \tag{6.1}$$

To verify if the above observation can be generalized, the test case set in Table 5.1 is run 100 times with randomly generated scaling factors in each run. Figure 6.3 indicates the test case result. Cases 2, 4, 6 and 7 hit the above observation with almost 100% frequency. This indicates that even with the scaling factor, Raman spectral information alone is sufficient to study the mixed molecules' orientation dis-

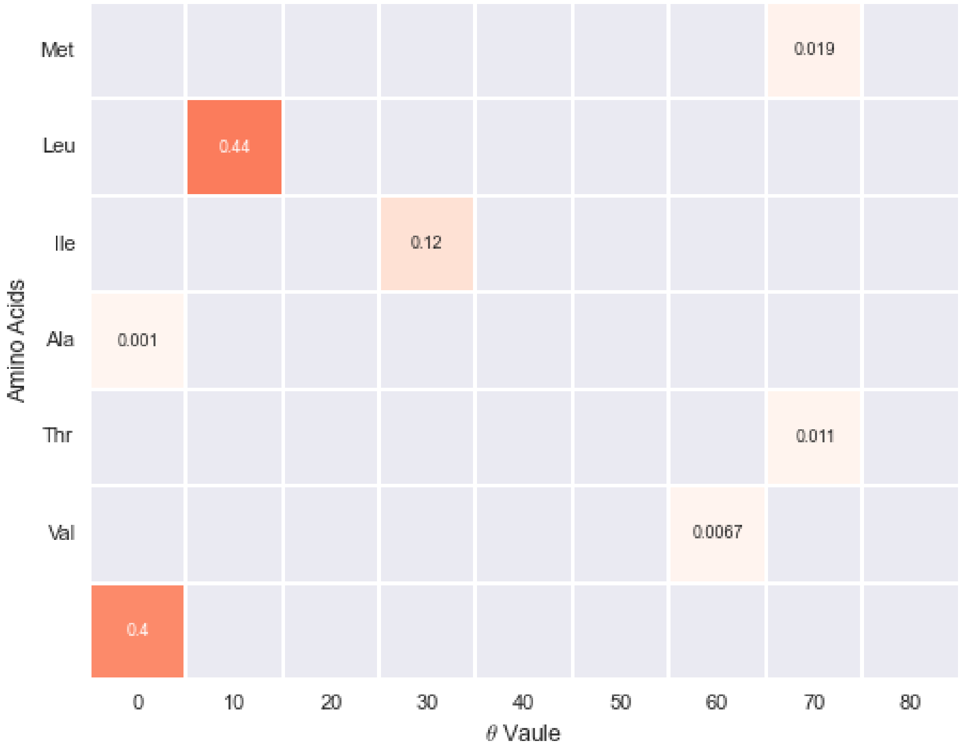


Figure 6.2: Return composition of Case 2 for one random run of the test case set with scaling foctor for mixed amino acids, with θ expanded from 0° to 80°. More detailed data of this target composition can be found in Appendix A.5.

tribution at interfaces when each amino acid's candidates expanded from 0° to 80° on θ. The target composition can be re-constructed correctly from the return slack variable and the return composition. Figure 6.3 also illustrates that Case 3 does not hit the above observation with high frequency. With the scaling factor as the addition, SFG spectral information is not sufficient to obtain the target composition. Case 5 indicates that even combining IR and SFG spectral information, the constructed LP model cannot help to reconstruct the target composition. This can be caused by the different scaling factors of these two spectroscopy techniques.

6.2.2 Test Cases with Scaling Factor Considering Each Amino Acid Candidates from 0° to 180° on θ

When each amino acid's candidates are expanded from 0° to 180° for θ, the same test case set is applied 100 times with randomly generated scaling factors in each run.

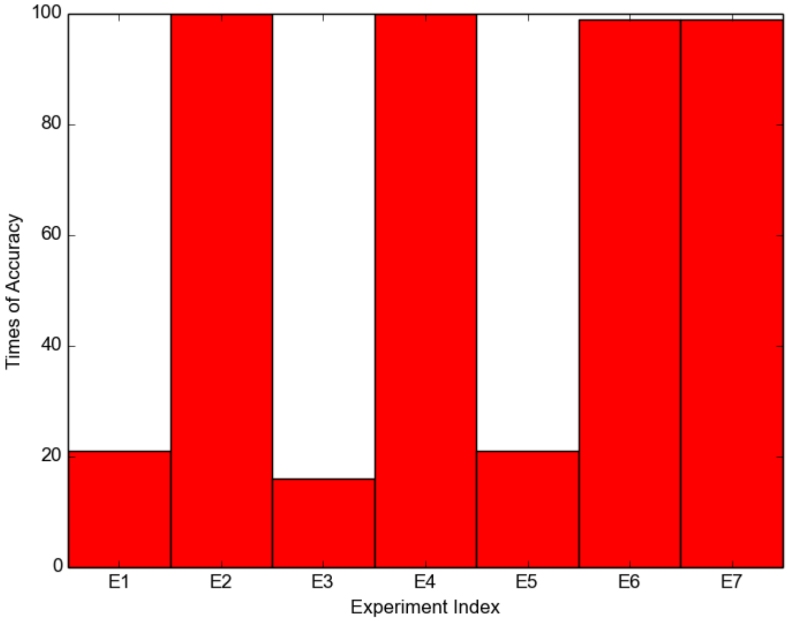


Figure 6.3: Test case accuracy analysis for test cases using experimental spectra data that contains scaling factor that is smaller than 1 and candidates with θ from 0° to 80°.

The test case result from the 100 run illustrates that all test cases in the set meets the above observation with zero frequency.

However, when we further analyze the return compositions of Cases 2 and 6, there are few other observations to be noted. To facilitate the explanation, one random run is picked as an explicit example. Figure 6.4 is the target composition. Figure 6.5 and Figure 6.6 are the return compositions of Case 2 and Case 6. The generated scaling factor for IR, Raman and SFG are 0.863411, 0.770505 and 0.239947.

In Figure 6.5, in the return composition of Case 2, the slack variable equals $1 - SF = 1 - 0.77 = 0.23$. For each amino acid, the selected candidate in the return composition may not be the exact one as shown in the target composition. However, this selected candidate is always either the correct one, or the correct one's θ complimentary. Moreover, the ratios between the percentage of each selected candidate in Figure 6.5 and Figure 6.4 are the same as shown in Equation 6.2 (more detailed calculated can be found in Appendix A.10). These ratios all equal to the scaling factor of Raman.

In Figure 6.5, for each amino acid, there are two selected candidates in the return composition. These two selected candidates are the correct one and its θ complimentary. When the percentages of these two selected candidates are added, it equals to the percentage returned for the amino acid in Figure 6.4. $0.27+0.14 = 0.41$. Between these two selected candidates, the correct one's percentage is always larger' than its θ complement. $0.27 > 0.14$. In conclusion, we observe Case 2 returns the slack variable, the scaling factor, and we can obtain the ratio between the returned candidates and the target ones. However, in order to distinguish the exact candidate of each amino acid, the extra information from Case 6 is required. Case 6 tells the correct candidate from its complement on θ. Together with the return information from Case 2 and 6, the target composition can be obtained. These observations can be applied to every run of the test case set.

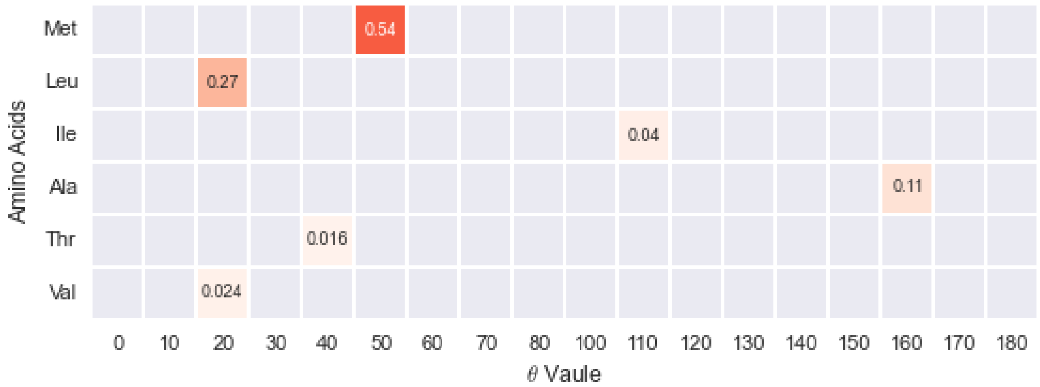


Figure 6.4: Target composition of one random run of test cases containing scaling factor and the mixed amino acids' candidates with θ expended from 0° to 180°. More detailed data of this target composition can be found in Appendix A.6.

$$\frac{0.41}{0.54} = \frac{0.21}{0.27} = \frac{0.03}{0.04} = \frac{0.088}{0.11} = \frac{0.012}{0.016} = \frac{0.019}{0.024} = 0.77 \tag{6.2}$$

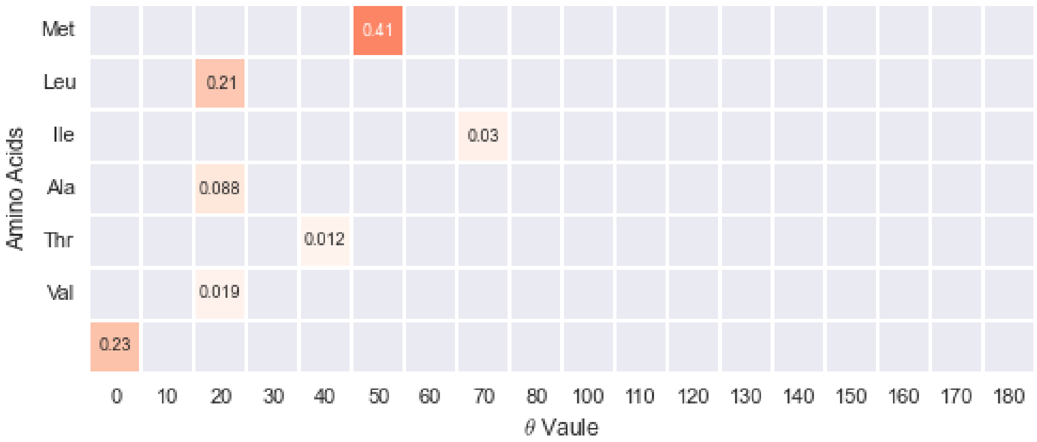


Figure 6.5: Return composition of Case 2 for one random run of test cases containing scaling factor and the mixed amino acids' candidates with θ expended from 0° to 180°. More detailed data of this target composition can be found in Appendix A.7.

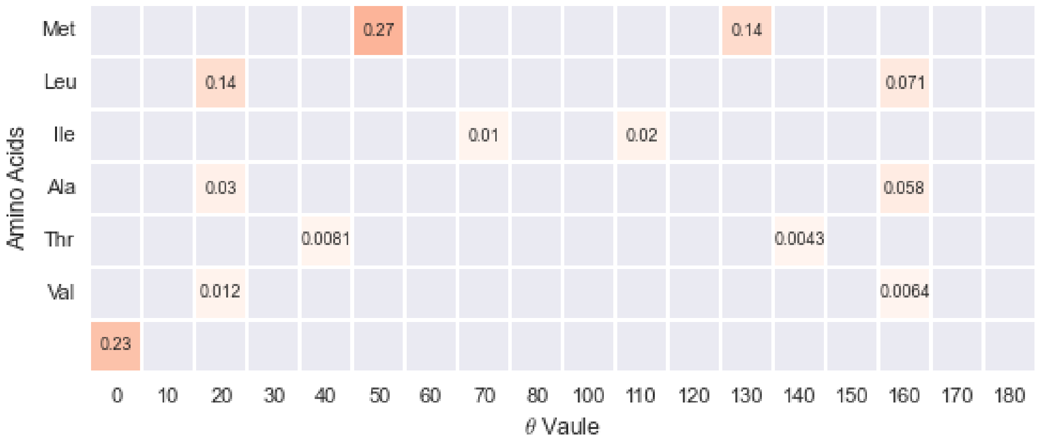


Figure 6.6: Return composition of Case 6 for one random run of test cases containing scaling factor and the mixed amino acids' candidates with θ expended from 0° to 180°. More detailed data of this target composition can be found in Appendix A.8.

6.3 Conclusion

With the introduction of the scaling factor to different spectroscopy techniques, Raman spectral information alone is sufficient to obtain the target composition, when considering a mixture of amino acids with candidates expanded from 0° to 80° on θ. The target composition can be re-constructed from the return SV and composition. The SF equals 1 minus SV.

When each amino acid's candidates are expanded from 0° to 180°, both return compositions from Case 2 and 6 are needed to obtain the target composition.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In addition to existing two common approaches in studying the possible composition of candidates of model spectra, the use of LP has been explored by Hung [4]. It has been shown that LP can solve this problem in pseudo polynomial time $O(n)$, which is much better in computational gain than the traditionally exhaustive way of extracting molecular orientation information. However, the reason why the LP model does not always return the target composition of mock spectra was unknown. The first goal of this study is to figure out this reason.

The achieve the first goal, a simplified molecular model is designed to analyze the nature of the LP model. It has shown that as long as there is the right data set, the target composition is obtained. If the target composition is not returned correctly, then there is not sufficient spectral information to describe the test cases to the LP model.

Furthermore, when we use all the spectral information of a realistic molecule (Met) to build the LP instances, it is not guaranteed to return the target composition all the time. The spectral information we collect is not sufficient in describing the test cases to the LP model most of the time.

Following the scenario of having one type of realistic molecule at surfaces, the test cases of having multiple types of realistic molecules at surface are also explored.

When each molecule's candidates expanded from 0° to 80° for θ, Raman or SFG spectral information alone is sufficient to obtain the target composition. When the candidates are expanded from 0° to 180° on θ, SFG spectral information needs to combine with IR or Raman in order to obtain the target composition.

At last, instead of generating the target spectra by combining different candidates directly, they are obtained from real experimental data. Therefore, for each spectroscopy technique, there is a scaling factor between the candidate spectra generated theoretically and the real experimental target spectra. When consider a mixture of realistic molecules with candidates expanded from 0° to 80° on θ, Raman spectral information alone is sufficient to obtain the target composition. Because the target composition can be re-constructed from the return SV and composition. The SF equals 1 minus SV. When each realistic molecule's candidates are expanded from 0° to 180°, both return compositions of using only Raman spectral information and using Raman and SFG spectral information are needed to obtain the target composition.

7.2 Contributions

- By studying the LP instances built by using spectral information of simplified molecular model, lack of sufficient spectral information int the instances is the reason that the LP solver does not return the target composition in some test cases.

- In the case of studying one type of realistic molecular model at surfaces, even combining all three spectral information of IR, Raman and SFG to build the LP instances, it is not sufficient to obtain the target composition for most test cases.

- In the case of different types of realistic molecular models at surfaces, Raman or SFG spectral information alone is sufficient to obtain the target composition when candidates of each molecular model expanded from 0° to 80° on θ value. When candidates of each molecular model expanded from 0° to 180° on θ value, SFG spectral information needs to be combined with IR or Raman to obtain the target composition.

- When the slack variable is introduced to each spectral technique, the case of different types of realistic molecular models at surfaces is considered. When each molecular model's candidates expanded from 0° to 80° on θ value, Raman spectral information alone is sufficient to obtain the target composition. When each molecular model's candidates expanded from 0° to 80° on θ value, the return compositions, of the LP instances using only Raman spectral information and using Raman and SFG spectral information, are both needed to obtain the target composition.

7.3 Future Work

Our LP model has proven its efficiency in studying molecular orientation distribution at surfaces when different molecules are considered. However, when considering one type of molecule at the surface, there is not enough spectral information for the LP model to obtain the target composition. One of the most important direction is to collect more spectral information to the LP model. Another direction is to combine LP technique with other computation model to further constraint the solution space of the target composition.

Appendix A

Additional Information

Detailed data value for Figure 5.4

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.299586 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.664053 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.021196 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00662804 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.000642609 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00789 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \tag{A.1}$$

Detailed data value for Figure 5.5

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0.299586 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.664053 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.021196 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00662804 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.000642609 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.00789 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \tag{A.2}$$

Detailed data value for Figure 5.6

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.299586 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.664053 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.021196 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00662804 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.000642609 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00789 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \tag{A.3}$$

Detailed data value for Figure 6.1

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.03218 & 0 \\ 0 & 0.73929 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.19745 & 0 & 0 & 0 & 0 & 0 \\ 0.00173 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01819 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01116 & 0 & 0 \end{bmatrix} \tag{A.4}$$

Detailed data value for Figure 6.2

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.019308 & 0 \\ 0 & 0.443574 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.11847 & 0 & 0 & 0 & 0 & 0 \\ 0.001038 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.010914 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.006696 & 0 & 0 \\ 0.4 & & & & & & & & \end{bmatrix} \tag{A.5}$$

Detailed data value for Figure 6.4. Target composition

$$\begin{bmatrix} & 0 & 0 & 0 & 0 & 0.53762 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.26894 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.03951 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.11382 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.01604 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.02407 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{A.6}$$

Detailed data value for Figure 6.5. Return composition of Result 2

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.414239 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.20722 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0304427 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0876989 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0123589 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0185461 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.229495 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{A.7}$$

Detailed data value for Figure 6.6. Return composition of Result 6

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.27162 & 0 & 0 & 0 & 0 & 0 & 0 & 0.142619 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.135875 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0713442 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0104812 & 0 & 0 & 0.0199615 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0301941 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0575048 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.00810383 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00425508 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0121608 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00638527 & 0 & 0 \\ 0.229495 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{A.8}$$

$$\frac{0.019308}{0.03218} = \frac{0.443574}{0.73929} = \frac{0.11847}{0.19745} = \frac{0.001038}{0.00173} = \frac{0.010914}{0.01819} = \frac{0.006696}{0.01116} = 0.6 \tag{A.9}$$

$$\frac{0.414239}{0.53762} = \frac{0.20722}{0.26894} = \frac{0.0304427}{0.03951} = \frac{0.0876989}{0.11382} = \frac{0.0123589}{0.01604} = \frac{0.0185461}{0.02407} = 0.770505 \tag{A.10}$$

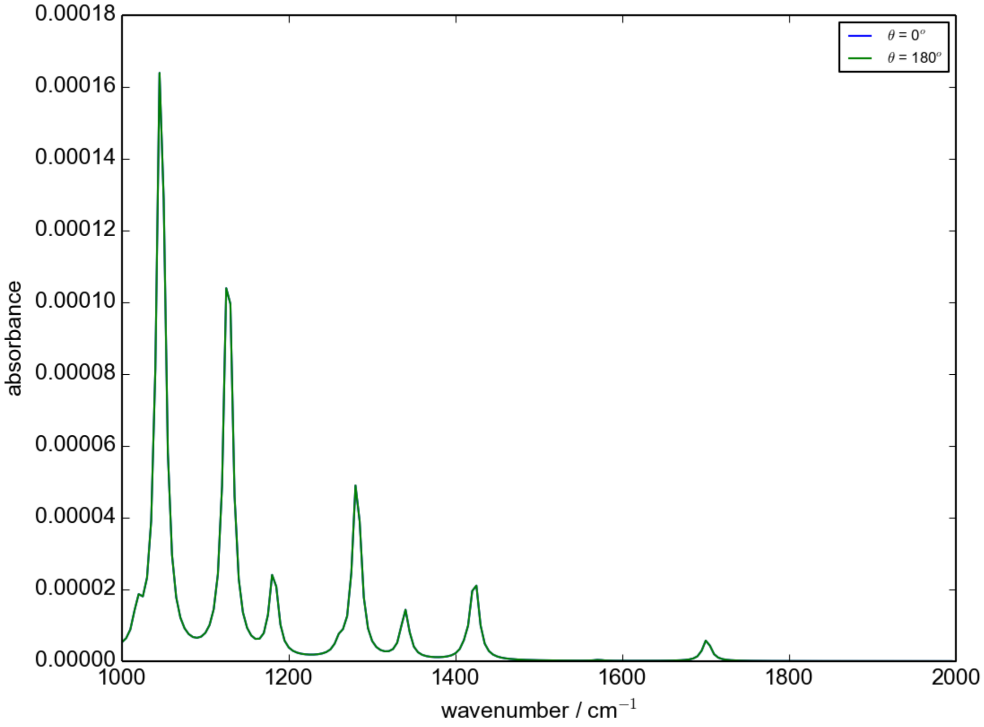


Figure A.1: IR z projection spectrum for alanine candidate with θ of 0° is identical to alanine candidate with θ of 180°.

Number of Candidates	5	
Candidates	[0, 10, 20, 30, 40]	
Target Composition	[0.2, 0.2, 0.2, 0.2, 0.2]	
Test case index	Constraints	Result
10	200, x 200, z	[0.607766, 0, 0, 0, 0.392234]
11	200, xx 200, xy 200, xz 200, zz	[0.247792, 0, 0.502139, 0, 0.250069]
12	200, xxz 200, xzx 200, zzz	[0.321014, 0, 0.31018, 0.163041, 0.205764]
13	200, x 200, z 200, xx 200, xy 200, xz 200, zz	[0.247792, 0, 0.502139, 0, 0.250069]
14	200, xx 200, xy 200, xz 200, zz 200, xxz 200, xzx 200, zzz	[0.321014, 0, 0.31018, 0.163041, 0.205764]
15	200, x 200, z 200, xxz 200, xzx 200, zzz	[0.321014, 0, 0.31018, 0.163041, 0.205764]
16	200, x 200, z 200, xx 200, xy 200, xz 200, zz 200, xxz 200, xzx 200, zzz	[0.321014, 0, 0.31018, 0.163041, 0.205764]

Table A.1: More detailed result data of Test Case 10 to 16 for methionine candidates.

# Candidates	9	
Candidates	[0, 10, 20, 30, 40, 50, 60, 70, 80]	
Target Composition	[0.2201, 0.28905, 0.05201, 0.08251, 0.35633, 0, 0, 0, 0]	
Test Case #	# of Data Points	Result Composition
17	each 5 wavenumber of IR, Raman and SFG spectra	[0.158921, 0.388434, 0.0, 0.0985466, 0.354099, 0.0, 0.0, 0.0, 0.0]
18	each 500 wavenumber of IR, Raman and SFG spectra	[0.397991, 0.0, 0.203394, 0.0357663, 0.362848, 0.0, 0.0, 0.0, 0.0]

Table A.2: More detailed result data of Test case 17 and 18 to explain the limitation of our LP model for methionine molecule.

Test Case #	# Data Points	Points Selection	Return Composition
6	10	[2800, 3300, 50], z	[0, 0.796962, 0.103038, 0.1]
7	20	[2800, 3300, 25], z	[0, 0.796962, 0.103038, 0.1]
8	25	[2800, 3300, 20], z	[0, 0.796962, 0.103038, 0.1]
9	32	[2800, 3300, 15], z	[0, 0.796962, 0.103038, 0.1]
10	50	[2800, 3300, 10], z	[0, 0.796962, 0.103038, 0.1]
11	100	[2800, 3300, 5], z	[0, 0.796962, 0.103038, 0.1]
12	$100 + 1$	[2800, 3300, 5], z [2800, 3300, 500], x	[0, 0.796962, 0.103038, 0.1]
13	$100 + 5$	[2800, 3300, 20], z [2800, 3300, 100], x	[0, 0.796962, 0.103038, 0.1]
14	$100 + 10$	[2800, 3300, 20], z [2800, 3300, 50], x	[0, 0.796962, 0.103038, 0.1]
15	$100 + 50$	[2800, 3300, 20], z [2800, 3300, 10], x	[0.1, 0.5, 0.4, 0]
16	$100 + 100$	[2800, 3300, 20], z [2800, 3300, 5], x	[0.1, 0.5, 0.4, 0]

Table A.3: Constraint study based on Case 4 of simplified molecular model.

Test Case #	# of Data Points	Point Selection	Return Composition
17	10	[2800, 3300, 50], z	[0.156758, 0, 0, 0.825977, 0, 0, 0, 0, 0, 0.017265]
18	25	[2800, 3300, 20], z	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
19	50	[2800, 3300, 10], z	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
20	100	[2800, 3300, 5], z	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
21	500	[2800, 3300, 1], z	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
22	100 + 1	[2800, 3300, 5], z [2800, 3300, 500], x	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
23	100 + 10	[2800, 3300, 5], z [2800, 3300, 50], x	[0.361587, 0, 0.312061, 0.326352, 0, 0, 0, 0, 0]
24	100 + 20	[2800, 3300, 5], z [2800, 3300, 25], x	[0.174023, 0, 0, 0.791447, 0, 0, 0.0345301, 0, 0, 0]
25	100 + 25	[2800, 3300, 20], z [2800, 3300, 20], x	[0.174023, 0, 0, 0.791447, 0, 0, 0.0345301, 0, 0, 0]
26	100 + 50	[2800, 3300, 5], z [2800, 3300, 10], x	[0, 0, 0.753209, 0, 0.146791, 0, 0.1, 0, 0, 0]
27	100 + 84	[2800, 3300, 5], z [2800, 3300, 6], x	[0.174023, 0, 0, 0.791447, 0, 0, 0.0345301, 0, 0, 0]
28	100 + 100	[2800, 3300, 5], z [2800, 3300, 5], x	[0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0]

Table A.4: Constraint study based on Case 5 of simplified molecular model.

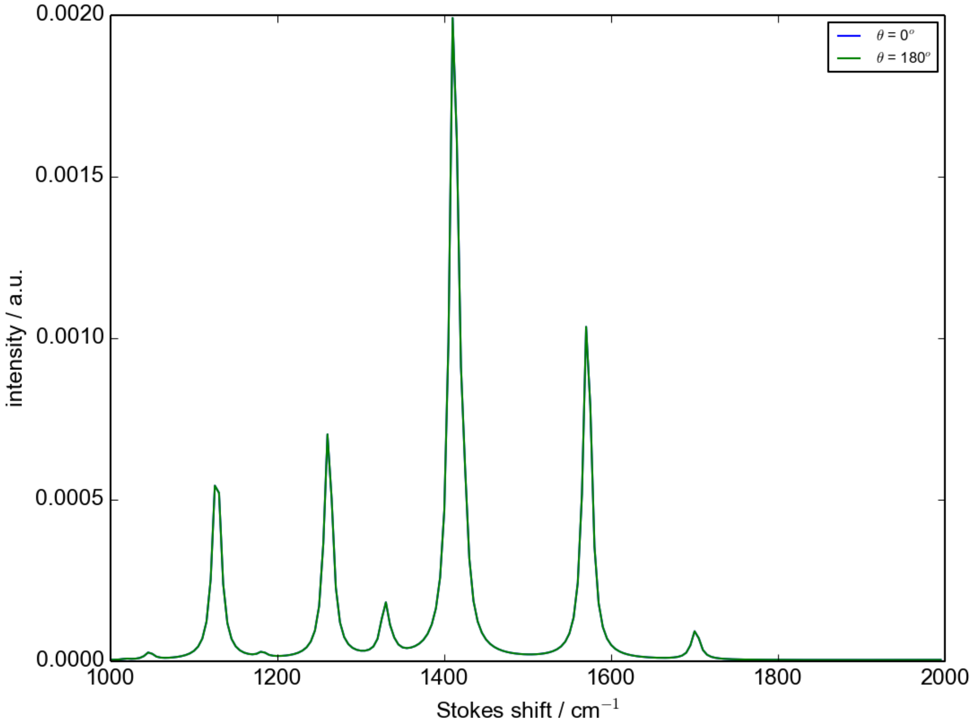


Figure A.2: Raman zz projection spectrum for alanine candidate with θ of 0° is identical to alanine candidate with θ of 180°.

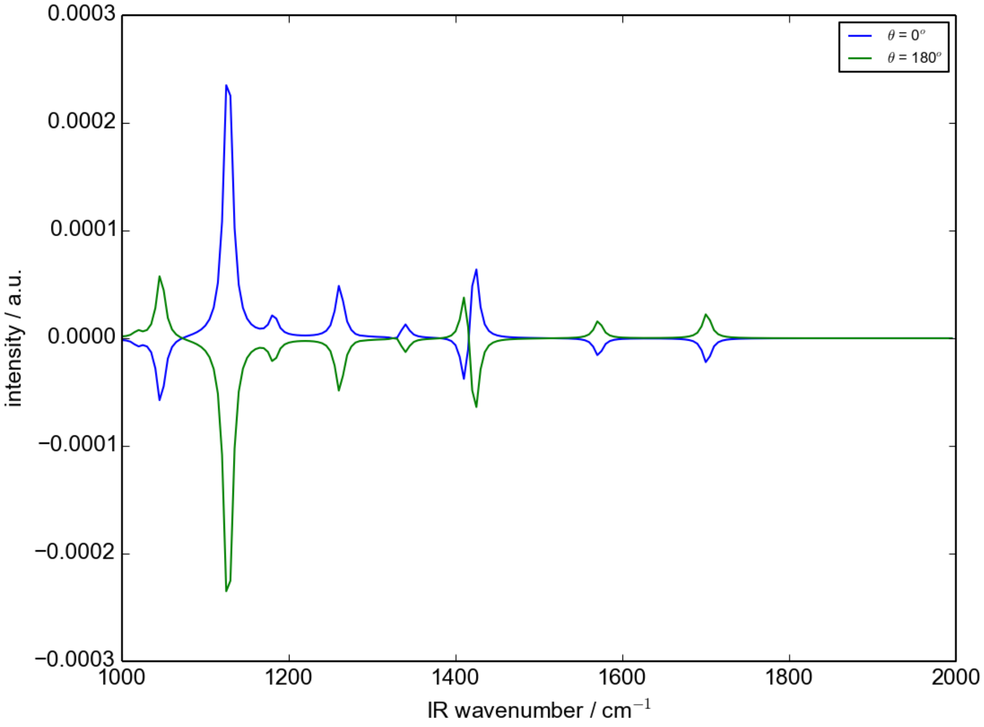


Figure A.3: SFG *zzz* projection spectrum for alanine candidate with θ of 0° is not identical to alanine candidate with θ of 180°, but symmetric along wavelength.

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press New York, NY, USA, 2004.

[2] Sophie Brasselet. Polarization-resolved nonlinear microscopy: application to structural molecular and biological imaging. *Adv. Opt. Photon.*, 3(3):205, Sep 2011.

[3] Vasek Chvatal. *Linear Programming.* 1983.

[4] Kuo Kai Hung. Extracting surface structural information from vibrational spectra with linear programming. Master's thesis, University of Victoria, 2015.

[5] Kuo-Kai Hung, Ulrike Stege, and Dennis K. Hore. Ir absorption, raman scattering, and ir-vis sum-frequency generation spectroscopy as quantitative probes of surface structure. *Applied Spectroscopy Reviews*, 50(4):351–376, 2015.

[6] Catherine Lewis. *Linear Programming: Theory and Applications.* Whitman College Mathematics Department, 2008.

[7] Jiri Maousek and Bernd Cartner. *Understanding and Using Linear Programming.* Springer, 2007.

[8] M.W.Schmidt, K.K.Baldridge, J.A.Boatz, S.T.Elbert, M.S.Gordon, J.H.Jensen, S.Koseki, N.Matsunaga, K.A.Nguyen, S.J.Su, T.L.Windus, M.Dupuis, and J.A.Montgomery. *General Atomic and Molecular Electronic Structure System.* Department of Chemistry Iowa State University, July 2016.

[9] Roy Sandra, Hung Kuo-Kai, Stege Ulrike, and K.Hore Dennis. Rotations, projections, direction cosines, and vibrational spectra. *Applied Spectroscopy Reviews*, 49:233–248, May 1999.

[10] Pratt Arnold W, Toal J. Nicolet, and Rushizky George W. Computer assisted analysis of oligonucleotides. *Annals of the New York Academy of Sciences*, 128(3):900–913, 1966.

[11] William C. Whiten, Marvin B. Shapiro, and Arnold W. Pratt. Linear programming applied to ultraviolet absorption spectroscopy. *Communications of the ACM*, 6:66–67, 1963.

[12] Zhuang X., Miranda P. B., Kim D., and Shen Y. R. Mapping molecular orientation and conformation at interfaces by surface nonlinear optics. *Phys. Rev. B*, 59:12632–12640, May 1999.