Spectroscopy Sensitivity Study by Linear Programmin

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

Spectroscopy Sensitivity Study by Linear Programmin

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

This document is a possible Latex framework for a thesis or dissertation at UVic. It should work in the Windows, Mac and Unix environments. The content is based on the experience of one supervisor and graduate advisor. It explains the organization that can help write a thesis, especially in a scientific environment where the research contains experimental results as well. There is no claim that this is the *best* or *only* way to structure such a document. Yet in the majority of cases it serves extremely well as a sound basis which can be customized according to the requirements of the members of the supervisory committee and the topic of research. Additionally some examples on using LaTeXare included as a bonus for beginners.

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 her all 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.

I believe I know the only cure, which is to make one's centre of life inside of one's self, not selfishly or excludingly, but with a kind of unassailable serenity-to decorate one's inner house so richly that one is content there, glad to welcome any one who wants to come and stay, but happy all the same in the hours when one is inevitably alone.

Edith Wharton

DEDICATION

Just hoping this is useful!

Chapter 1

Introduction

1.1 Background and Motivation

An interface is what forms a common boundary between two phases of matter. The phases of matter can be of any form, i.e, solid, liquid, and gas. The behavior of a surface greatly affects the properties of a material, such as 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 electronic technology. 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, bio-membrane functions and so on. [**?**]

Many experimental techniques have been applied in the study of molecular orientation at interfaces. Among them the optical methods are more preferable. Such methods include infrared (IR) absorption, Raman scattering and visible-infrared sum-frequency generation (SFG) spectroscopy. All these vibrational spectra carry quantitative structural information of molecules at interfaces. Although each of them has its own strengths and shortcomings, they all share the following advantages when compared with other non-optical methods. First of all, they all can be applied to any interfaces accessible by light. Second, they are non-destructive. Third, they are highly sensitive to good spatial, temporal and spectral resolutions [**?**], [**?**]. An

important advantage of SFG techniques is: it can discriminate against bulk contributions. This means that its result will not take the effect from the bulk (TODO: double check with Dennis). In order to extract the quantitative structural information that molecules carry at interfaces, different spectroscopy techniques and analyse are required. Combining different spectroscopy techniques is a very effective way to achieve the goal of molecular study. However, finding the most effective ways to combine these techniques may not be clear sometimes.

In order to analyse 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 interface based on the mathematical distribution function(TODO: further explained) and projecting the vibrational mode properties from molecular frame to laboratory frame. The main focus of our study is to analyze the data of these three spectroscopy techniques using Linear Programming (LP). In the following, we will explore how LP can facilitate extracting quantitative structural information of molecules at interfaces.

Our approach is to first study a model of a small molecule using LP. We then apply our finding and method to real molecules which are six amino acids: methionine, leucine, isoleucine, alanine, threonine and valine.

Before we explain the LP technique and LP model, as well as describe the molecules studied, we introduce the basic theory of the IR, Raman and SFG spectra.

1.2 Experimental Probes: IR, Raman, SFG

Vibrational spectra (IR, Raman and SFG) 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.

For IR, its absorption is the absorption-transmission-reflection mode (resonant). The physical principle is the variation of the static dipole moment μ (the first rank tensor) along the normal coordinates Q: $^{\partial\mu}/_{\partial Q}$.

$$I_{IR} \approx \left| \frac{1}{\sqrt{2m_Q w_Q}} \frac{\partial \mu}{\partial Q} \right|^2 \tag{1.1}$$

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 1.2. The IR spectra can be obtained from 3 polarizations: x, y, z.

$$\frac{\partial \mu}{\partial Q} = \begin{bmatrix} {}^{\partial \mu_x}/_{\partial Q} \\ {}^{\partial \mu_y}/_{\partial Q} \\ {}^{\partial \mu_z}/_{\partial Q} \end{bmatrix} \tag{1.2}$$

Raman is scattered from the molecule sample. Unlike IR, Raman spectra relate to the variation of the molecular polarizability α (the second rank tensor) along the normal coordinates Q: ${}^{\partial \alpha}/_{\partial Q}$.

$$I_{Raman} \approx \left| \frac{1}{\sqrt{2m_Q w_Q}} \frac{\partial \alpha^{(1)}}{\partial Q} \right|^2 \tag{1.3}$$

where $|0\rangle$, $\langle 1|$, m are same as defined above. w is the resonance frequency. The polarizability is coupled with (x, y, z) components of the driving field and x, y, z components of the induced dipole. Therefore, there are 9 elements in the polarizability, which can be expressed as Equation 1.4. Furthermore, it results in 9 polarizations of Raman spectra: xx, yy, zz, xy, xz, yx, yz, zy and zx.

$$\frac{\partial \alpha^{(1)}}{\partial Q} = \begin{bmatrix} \frac{\partial \alpha_{xx}^{(1)}}{\partial Q} & \frac{\partial \alpha_{xy}^{(1)}}{\partial Q} & \frac{\partial \alpha_{xz}^{(1)}}{\partial Q} \\ \frac{\partial \mu_{yx}}{\partial Q} & \frac{\partial \alpha_{yy}^{(1)}}{\partial Q} & \frac{\partial \alpha_{yz}^{(1)}}{\partial Q} \\ \frac{\partial \mu_{zx}}{\partial Q} & \frac{\partial \alpha_{zy}^{(1)}}{\partial Q} & \frac{\partial \alpha_{zz}^{(1)}}{\partial Q} \end{bmatrix} \tag{1.4}$$

SFG stands for sum frequency generation vibrational spectroscopy. It is a surface-specific technique. SFG is a non-linear optical process. It is sensitive to the molecular orientation in odd orders. Comparing to linear optical spectroscopy, the biggest advantage of SFG is that it is surface specific. The spectroscopy signal only comes from the surface, not the bulk. SFG is the variation of the outer product of dipole moment

and polarizability, $\chi^{(2)}$ (the third rank tensor): $^{\partial\mu}/_{\partial Q} \otimes {}^{\partial\alpha}/_{\partial Q}$. Therefore, there are 27 elements for SFG spectra, which result in 27 polarizations of SFG spectra.

$$I_{SFG} \approx \left| \frac{1}{2m_Q w_Q} \left(\frac{\partial \alpha^{(1)}}{\partial Q} \otimes \frac{\partial \mu}{\partial Q} \right) \right|^2 \tag{1.5}$$

1.3 Linear programming

Linear Programming (LP) problems are an optimization problems of a specific form. The standard form of LP is a minimization problem that has an objective function and constraints as shown in Equation 1.6 [**?**]:

$$\begin{aligned} \text{minimize} \quad & c_1x_1 + c_2x_2 + ... + c_nx_n \\ \text{subject to} \quad & a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = \quad b_1 \\ & a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = \quad b_2 \\ & \qquad\qquad . \qquad\qquad\quad . \\ & \qquad\qquad . \qquad\qquad\quad . \\ & \qquad\qquad . \qquad\qquad\quad . \\ & a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = \quad b_m \\ & x_1 \geq 0, x_2 \geq 0, ..., x_n \geq 0, \end{aligned} \tag{1.6}$$

where x_i are the decision variables, a_{ij} is a matrix of know 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 constraints are the conditions that the decision variables need to subject to. They specify a convex polytope over which the objective function is to be optimized.

The diet problem is a popular example to illustrate the concept of LP. It can be described as follows: a restaurant would like to achieve the minimal nutrition requirement with the lowest price of the food selection as shown in Table **??**. For each meal, the minimum requirements for vitamin A, vitamin C and dietary fiber are 0.5 mg, 15 mg and 4 g. The restaurant has three options: raw carrot, raw white cabbage and pickled cucumber. The table also displays the nutrition contains and

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

the price of each ingredient. With all this information, we need to know how much carrot, cabbage and cucumber to add to 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, we come up with a model as shown Equations 1.7 to 1.13.

$$\begin{aligned}
\text{minimize} \quad & 0.75x_1 + 0.5x_2 + 0.15x_3 && (1.7)\\
\text{subject to} \quad & 35x_1 + 0.5x_2 + 0.5x_3 \geq 0.5 && (1.8)\\
& 60x_1 + 300x_2 + 10x_3 \geq 15 && (1.9)\\
& 30x_1 + 20x_2 + 10x_3 \geq 4 && (1.10)\\
& x_1 \geq 0 && (1.11)\\
& x_2 \geq 0 && (1.12)\\
& x_3 \geq 0 && (1.13)
\end{aligned}$$

In this LP model, x_1, x_2 and x_3 are decision variables, each presents the amount of the ingredients. Equation 1.7 is the objective function to be minimized. Equation 1.8 to Equation 1.10 describe the nutrition requirements. Equation 1.11 to Equation 1.13 ensure the amount of each ingredient to be greater than 0. With the existing LP solvers that implemented Simplex Method, the optimal solution can be obtained within a second.

For a LP problem, there exist only three kinds of solutions: feasible and bounded solutions, feasible and unbounded solutions, and infeasible solutions. If the solution space is feasible and bounded, then there is one optimal solution. If it is feasible but unbounded, then there is a solution space with an infinite number of optimal solutions [**?**].

A general LP problem can be a minimization or a 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 n-dimensional space that is called R^n, each constraint is a hyperplane that divides this R^n space into two half-spaces. Therefore, all the constraints together cut this R^n space into a convex polyhedron if there are feasible solutions. This makes LP a convex problem. The benefit of a convex problem is that the local optimal solution is also the global optimum. LP solvers return is the optimal solution. If a LP problem has a unique optimal solution, then this solution is a vertex of the convex polyhedron.

LP is a convex, a deterministic process. It is guaranteed to converge to a single global optimum if there is a solution space. LP problems are intrinsically easier to solve than many non-linear problems.

Another advantage of LP is that it can deal with thousands of variables, which makes it suitable for the study of a molecule's coordination composition at interfaces.

Various algorithms are available in solving LP problems, such as: Simplex algorithm, Interior point, and Path-following algorithms. Both Interior Point and Simplex are common and mature algorithms that work well in practice. Simplex is comparatively easier to understand and implement than Interior Point. Simplex method takes the advantage of the geometric concept that it visits the vertices of the feasible set (convex polyhedron), and check the optimal solution among each visited vertex. The converging approach is also different for these two methods. If there are n decision variables, usually Simplex will converge in $O(n)$ operations with $O(n)$ pivots. Interior point traverses the edges between vertices on a polyhedral set. Generally speaking, Interior point method is faster for larger problems with sparse matrix. However, when experimenting with these two methods, the speed of them is not much different from each other for the current study. For our study, Simplex method has proved to be efficient and effective, it will be used for all the experiments.

Last but not the least advantage of LP is its speed. 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 $2n - 3n$ steps to solve a problem (n is the number of 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 ACNSI C. It is open-source and intended to solve large scale LP problems.

(TODO: compare linear programming with other tools, like quadratic programming and linear regression) Compare linear programming with quadratic programming, why linear programming is a better approach to the problem? (Having problems finding related work or how to prove it myself)

Is there only computational gain? Also consider the model itself and solution space (The problem is defined as "Candidate ratio problem" in Kai's thesis, same here???) to determine the level of similarity between spectra is not an easy task

1.4 Aims and scope

Given some target experimental spectra and a set of candidates spectra, to find out the right combination of candidates for the target spectra is the goal in this study. The approach is to build our LP model, and check if the optimal solutions returned by the solver match the target composition that was used to generate the target spectra. The LP models are built using spectra resulting from different techniques. Therefore, there are different LP models. for each of these models, we then analyze which model is best to reach the goal with the highest accuracy. From this, we can decide which spectroscopy technique(s) is sufficiently sensitive in finding the right combination of candidates. Furthermore, we will consider various study focuses, and for each focus, what spectroscopy techniques combined with LP modelling should be applied in order to obtain the accurate composition of the target spectra.

Chapter 2

Methods

2.1 Current approaches to molecular structure elucidation

Currently, there are two main approaches in studying the orientation distribution of molecules at interface. One is comparing the experimental spectra with few predicted ones, and select the one that most matches to the experimental one. Another one is running an exhaustive algorithm to explore the most possible solution space. [**?**] However, both approaches take a lot of time and computational resources. In Hung's study [**?**], a new approach is introduced by applying Linear Programming to vibrational spectra to extract the molecular structure elucidation. This LP approach helped to return the target orientation distribution information when the mock experimental spectrum consisted of different amino acids. However, when candidates are coming from the same amino acid, LP approach failed to return the target orientation distribution information. The reason why LP failed to return the target composition has not been thoroughly studied in Hung's study. My study is to figure out the underlying reason causing the return composition of the LP model does not match to the target composition. After this reason is resolved, we want to explore the limitation of the LP model. For the cases that LP model helps to return the target composition, we want to know if the LP models can be applied systematically to the similar cases.

2.2 Structure of molecules adsorbed to interfaces

(TODO: check with Dennis, how to expand this part.) A picture to display molecules adsorbed to interfaces

2.3 Generating model spectra

As mentioned in Chapter 1, before analyzing the vibrational spectra of amino acids, there are a few factors to address first. First of all, creating candidate spectra is an essential step. This part of research has been done thoroughly by Hung [**?**].

To generate these amino acids' vibrational spectra, a molecule's vibration modes need to be modelled in the molecular frame, then transferred to the laboratory frame to work with the systems where interfaces exist. Chapter 2 in Hung's thesis [**?**] describes how to perform electronic structure calculations using GAMESS [**?**] to obtain every dipole moment and polarizability derivatives. Then he introduced how to use Direction Cosine Matrix (DCM) to transfer these two derivatives from the molecular coordinate system to the laboratory one. After that, Euler angles could be extracted from DCM. Euler angles are used to describe a molecule's coordination at interfaces. They are labelled by θ, ϕ and ψ as shown in Figure **??**. They are referred as *tilt*, *azimuthal* and *twist* angles respectively. Let x, y and z be lab frame Cartesian coordinates, and 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 [**?**]. After three steps of successive rotations of Euler angles, molecule properties can be transferred from the molecular frame to the lab frame.

In order to achieve the above steps, Hung first did a Hessian calculation. Secondly, 7 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 7 snapshot moment. Then the derivatives of dipole moment and polarizability are obtained by the interpolation of these 7 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 abstracted Euler angles from DCM. After this transformation, he restored the derivatives information into some molecular property files for any further usage. (TODO: double check the accuracy with Dennis)

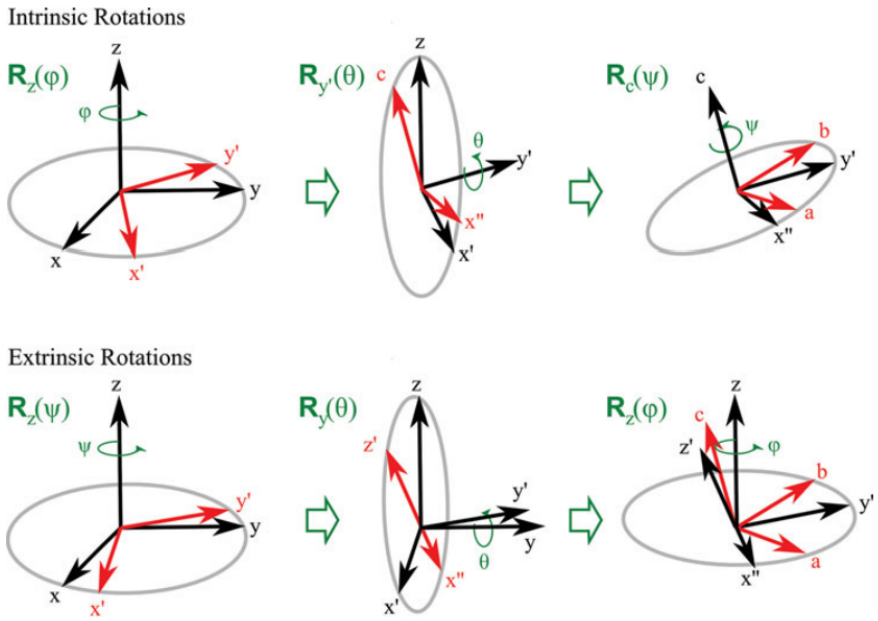

Figure 2.1: The Euler angles represented as the spherical polar angles θ, ϕ and ψ, and the illustration of the three successive rotations that transform the lab x, y, z coordinate system into the molecular a, b, c frame [**?**].

In my study, those molecular property files are used to generate the following amino acids' spectroscopy information: methionine, leucine, isoleucine, alanine, threonine and valine. Each molecular property file contains the derivatives of dipole moment and polarizabilities for vibrational mode of $3N - 6$. N is the number of atoms of a molecule. Furthermore, the following equations from Equation 2.2 to Equation 2.4 are used to generate each amino acid's IR, Raman and SFG spectra.

All the experiments are limited to only consider the *tilt* angle distribution among Euler angles, assume isotropy in *twist* and *azimuthal* angular distributions. Therefore, *twist* and *azimuthal* angles are integrated to create a uniform distribution. For angle ψ, it requires the surfaces to be not striped. However, there can be no anisotropy in the plane of the surface. Because of this, we can limit the candidate number by integrating angle ψ. On the other hand, for angle ϕ, a uniform distribution implies that the molecule has cylindrical symmetry in its preference of surface. This means that 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. However, the number of the amino acid candidates is still large when only considering θ angle. The possible combinations of all these amino acid candidates are still considered to be excessive (TODO: put these into an approximated number).

Furthermore, when molecules lay on an interface, the orientation for each molecule varies. Therefore, to simulate vibrational spectrum, a reasonable orientation distribution for the molecules needed to be studied. The orientation distribution requires either do a molecular dynamic simulation to study the distribution of molecule orientations at the interface, or come up with a analytic orientation distribution function. In my study, the second method is preferred. Moreover, Delta 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 one angle at the interface.

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

Infrared (IR) absorption spectroscopy is a harmonic approximation, its intensity 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.

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

where I_x represents x-polarized intensity. The same equation applies to I_y and I_z. 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 experiments. 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 molecular property files. Furthermore, because ϕ and ψ angles are integrated, the y-polarized spectrum is identical with the z-polarized one. Therefore, in current experiments, only two unique polarized IR spectra are obtained for one molecule. For simplicity, IR spectra are referred as y and z in future experiments.(TODO: need to double check the accuracy with Dennis)

The intensity of Raman scattering is proportional to the square of laboratory-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 from 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 Δw is the Stokes Raman shift. $\alpha_{xx}^{(1)}$ is one component of the 9-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 molecular property files. Similar to IR spectroscopy, because of the integration of ϕ and ψ angles, only 4 unique spectra are obtained from the following polarization: xx, xy, xz and zz. For simplicity, Raman spectra are referred as xx, xy, xz and zz in future experiments (TODO: double check the accuracy of the content with Dennis).

The intensity of SFG spectroscopy is proportional to the squared magnitude of the second-order susceptibility, $\left|\chi^{(2)}\right|^2$. $\chi^{(2)}$ is derived from the second-order polarizability, α^2. Equation 2.4 shows the response intensity of I_{xxx}.

$$I_{xxx}(\omega_{\mathrm{IR}}) = \sum_q \frac{1}{2m_q\omega_{\mathrm{q}}} \left\langle \left[\frac{\partial\alpha_{xx}^{(1)}}{\partial Q}\right]_q \left[\frac{\partial u_x}{\partial Q}\right]_q \right\rangle \frac{1}{\omega_{\mathrm{q}} - \omega_{\mathrm{IR}} - i\Gamma_q} \tag{2.4}$$

where I_{xxx} is the second-order susceptibility tensor. It is probed by an x-polarized visible incoming beam at frequency ω_{vis} and a x-polarized infrared beam incoming with frequency ω_{IR} are incident to the sample. Then the x-component of SFG at frequency $\omega_{SFG} = \omega_{\mathrm{vis}} + \omega_{\mathrm{IR}}$ is selected for detection. As $i = \sqrt{-1}$ is in the denominator, $X^{(2)}$ is a complex value [**?**]. The SFG response is the imaginary component of the second-order susceptibility. As in IR and Raman spectroscopy, all the values of ω_{IR}, μ, m_q, Q are obtained from the pickle files. Because of the integration of ϕ and ψ angles, only 3 unique non-zero spectra are obtained from the following polarization: yyz, yzy and zzz. For simplicity, SFG spectra are referred as yyz, yzy and zzz in future experiments. (TODO: double check the accuracy of the content with Dennis).

With these equations and the molecular property files, IR, Raman and SFG spectra can be generated for a candidate. A candidate in my study is a specific amino acid with specific θ value. (TODO: check accuracy with Dennis). Taking Methionine

as an example, Figure 2.2 displays x-polarized IR spectra of the following candidates: Methionine with θ of 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, Figure 2.3, 2.4 and 2.5 depict the spectra of the same candidates and targets for z-polarized IR, xx-projection Raman and yyz-projection SFG spectrum respectively. In Figure 2.2, the biggest differences among the candidates exist at each vibrational mode. The valid range for wavenumber isfrom 1000 to 2000. Each projection of IR, Raman or SFG, there are 200 data points can be extracted in the interval of 5 wavenumber. With these data points, the corresponding LP model is constructed as described in Chapter 3.

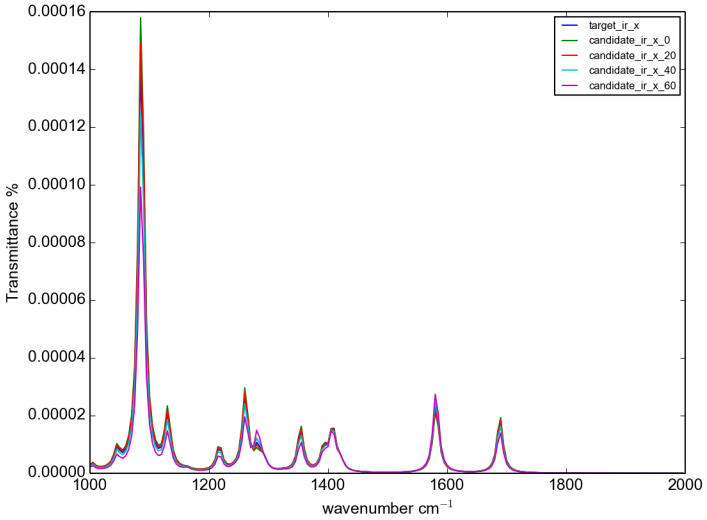

Figure 2.2: IR x projection spectra for methionine four candidates and target

2.4 The properties of the LP models

In Chapter 3, the properties of the LP models is studied. It is important to study the maximum capacity of the constructed LP models, and figure out what are the reasons that cause the limitation of them. Meanwhile, comparing the sensitivity of different

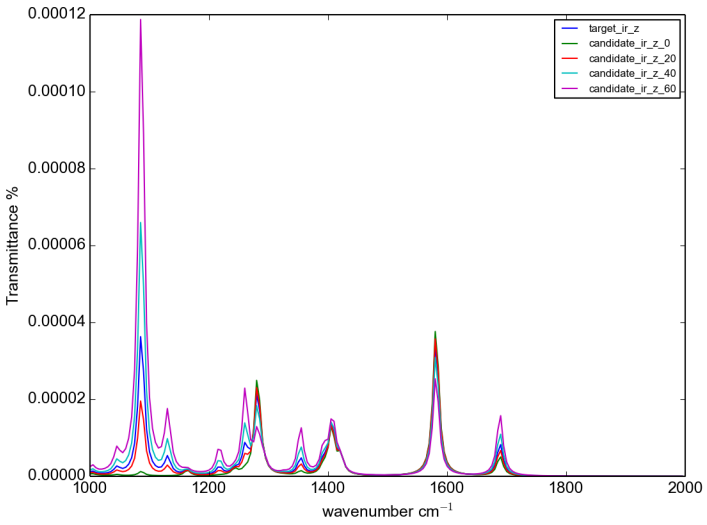

Figure 2.3: IR z projection spectra for methionine four candidates and target

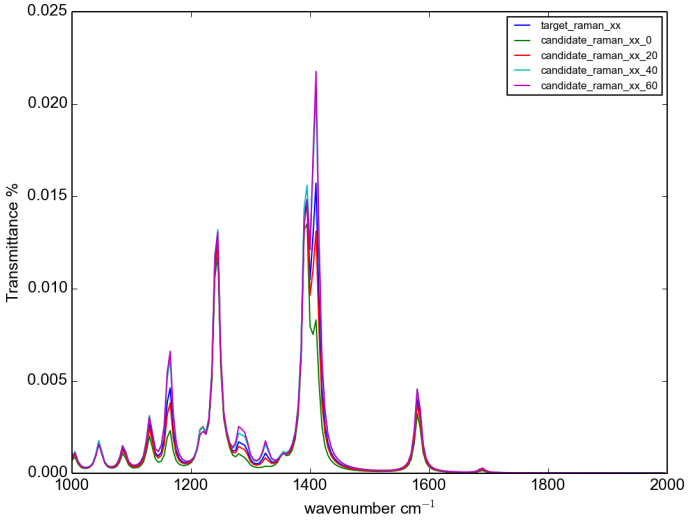

Figure 2.4: Raman xx projection spectra for methionine four candidates and target

spectroscopy techniques in the term of studying molecular orientation distribution at interfaces is also desired.

In Chapter 3, the study of the properties of the LP models is conducted by using a toy model to gain an insight into the limitation of our LP approach. With the further information gained in Chapter 3, further experiences will be conducted accordingly in Chapter 4, 5 and 6.

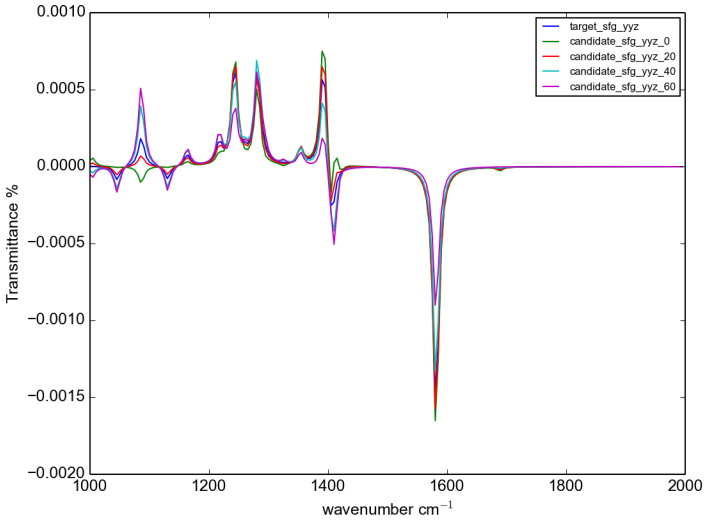

Figure 2.5: SFG *yyz* projection spectra for methionine four candidates and target

Chapter 3

Simplified Molecular Model

3.1 Description

The goal of Chapter 3 is to further expand this research focus and introduce the formulas used to generate our LP models. Before studying real life molecules, a toy molecule with limited vibration modes is first studied. By doing so, the nature of the LP formulas used to gain some further insights of the models we need is carefully analyzed. Our goal is to figure out with the all the spectral information available, could LP models we build output any valuable information.

A toy molecule with 4 vibration modes is constructed. Theses vibrational peaks are at frequencies of $2850, 2960, 3050$ and 3200. The widths of the peaks are $5, 10, 5$ and $15\ cm^{-1}$, respectively. The amplitude of the peak are $1, 0.7, -0.2$ and $0.5\ cm^{-1}$, respectively. The comparing angles of the peaks are $15, 90, 0$ and 60. (TODO: check with Dennis, how to explain those comparing angles?)

For this toy model, only IR spectroscopy is considered. Equation 3.1 is used to generate the cosine projection IR spectroscopy. Moreover, both ϕ and ψ Euler angles are integrated, only the difference on angle θ is considered for the toy model as well.

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

where A is the amplitude, θ_q is the *tilt* angle of the candidate, γ is the width, and ω_q is the frequency. (TODO: Double check the correct meaning of each symbol) Ten different candidates with 10 different θ values as follows: 0°, 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 90°. Their spectra are generated as shown in Figure **??**. The 10 candidates have peaks at the same frequencies. The spectral signal for candidates is comparatively strong at each peak.

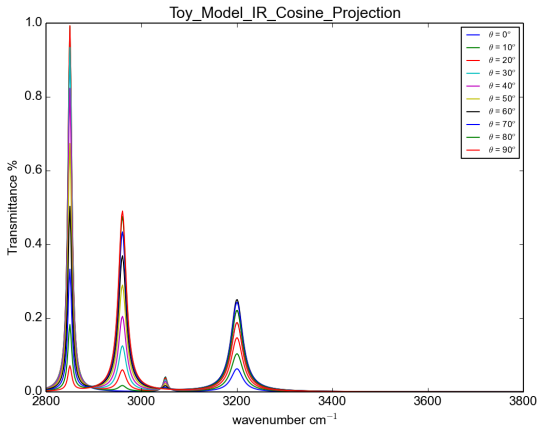

Figure 3.1: Toy model IR candidates cosine polarization

3.2 Linear Programming Model for Spectra Study

The LP model constructed to check if the optimal solution returned by the LP solver actually matches the target composition, is shown in Equation 3.2. This model has also been used to study the composition of Ribonucleic acid (RNA) with ultraviolet (UV) spectra [**?**] and other UV spectroscopy studies [**?**] back in the 60s.

$$minimize \sum_{p=1}^{number of points} \left| Target - \sum_{c=1}^{number of candidates} p_c f_\theta(x) \right| \tag{3.2}$$

where p_c are the unknown percentages for each candidate, which are the decision variables. 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. For each data point, the absolute residual between 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.

However, in order to use a LP approach, getting rid of the absolute signs in the objective function is needed. Because Equation 3.2 is subject to no restrictions, meanwhile, the objection function is not in standard form. To eliminate the absolute sign is achieved by introducing one more variable X and two more constraints for each data point as shown in Equation 3.3. Then the previous model in Equation 3.2 is converted into the one in Equation 3.4, which can be solved by an LP solver. At last, one more constraint is introduced to restrict the sum of the percentages to be 1, as shown in Equation 3.4.

For each point in the range of valid wavenumbers:

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

$$
\begin{aligned}
&minimize \sum_{p=1}^{points} X_p \\
X_1 - Target_1 + &\sum_{c=1}^{candidates} p_c f_\theta(x_1) \geq 0 \\
X_1 + Target_1 - &\sum_{c=1}^{candidates} p_c f_\theta(x_1) \geq 0 \\
&\cdots \\
X_n - Target_n + &\sum_{c=1}^{candidates} p_c f_\theta(x_n) \geq 0 \\
X_n + Target_n - &\sum_{c=1}^{candidates} p_c f_\theta(x_n) \geq 0 \\
&\sum_{c=1}^{candidates} p_c = 1
\end{aligned} \tag{3.4}
$$

3.3 Linear programming model implementation

To start, code is written to generate a file that contains all the candidates' spectral information needed for the experiments. For this step, the molecular properties files are used. For a specific candidate, given a molecular properties file and a θ value, the candidate's spectral information is obtained. For toy model, only the value of θ is needed, then Equation 3.1 is used to synthesize the spectral information.

In the first step, a candidate class is written. This class defines candidate's x- and z- polarized IR spectra; xx-, xy-, xz-, and zz- polarized Raman spectra; yyz-, yzy-, zzz- polarized SFG spectra. Given candidate's molecular properties and θ value, a instance of this specific candidate is created. For the toy model, it only contains IR spectral information. Therefore, one candidate only contains *cosine*- and *sine*-polarized IR spectra.

In the second step, more code is written to generate a target composition of a list of needed 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. For further experiments in Chapter 4, 5 and 6, the probe arrange is from

2000 wavenumber to 3000 wavenumber. For toy model, the probe range is from 2800 wavenumber to 3800 wavenumber. Then all the spectral information of candidates and target is created in a text file. The code can also be used to run experiments that contain different spectroscopy information. For example, one file can contain only candidates and target's IR spectral information, or contain all three spectroscopy information.

In the third step, the LP model is constructed by using the spectral information text file. This part of the code was written by Hung [**?**]. It reads all the candidates and target spectral information, and builds the LP model as shown in Equation 3.4, then creates CPLEX LP input file.

In the fourth step, we use LP solver "GNU linear progarmming tool kit" (GLPK) to obtain the result.

3.4 Experiments

In the first experiment set, 4 candidates are selected as shown in Table **??**. The table illustrates the detailed settings for Experiment 1 and 2. In Experiment 1, there are 4 candidates with θ of $0°, 10°, 20°$, and $30°$. For Experiment 2, the θ values are changed to $0°, 5°, 10°$, and $15°$. Instead of having a 10 degree variance in θ, 5 degree difference is applied on θ for Experiment 2. This means that when the candidates become more similar to each other as their spectra are more similar. In both experiments, 100 data points are selected evenly along the wavenumber from the spectra of *cosine*-polarized IR. The target composition of the candidates are the same for both experiments. In Experiment 1, the return composition is the same as the target one, however, the return composition for Experiment 2 does not match to the target one.

In order to figure out why the return composition in Experiment 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 **??**. Note that the result spectra is almost identical to the target one. The residual between them is almost 0. In order to see whether this observation is a general case, another experiment Experiment 3 is set up in Table **??**. Experiment 3 contains more candidates than Experiments 1 and 2. 10 candidates are included with θ values ranging from $0°$ to $90°$.

Experiment 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(cos)	100(cos)
Return Composition	[0.1, 0.5, 0.4, 0]	[0, 0.796962, 0.103038, 0.1]

Table 3.1: Experiment 1 and 2 Setting

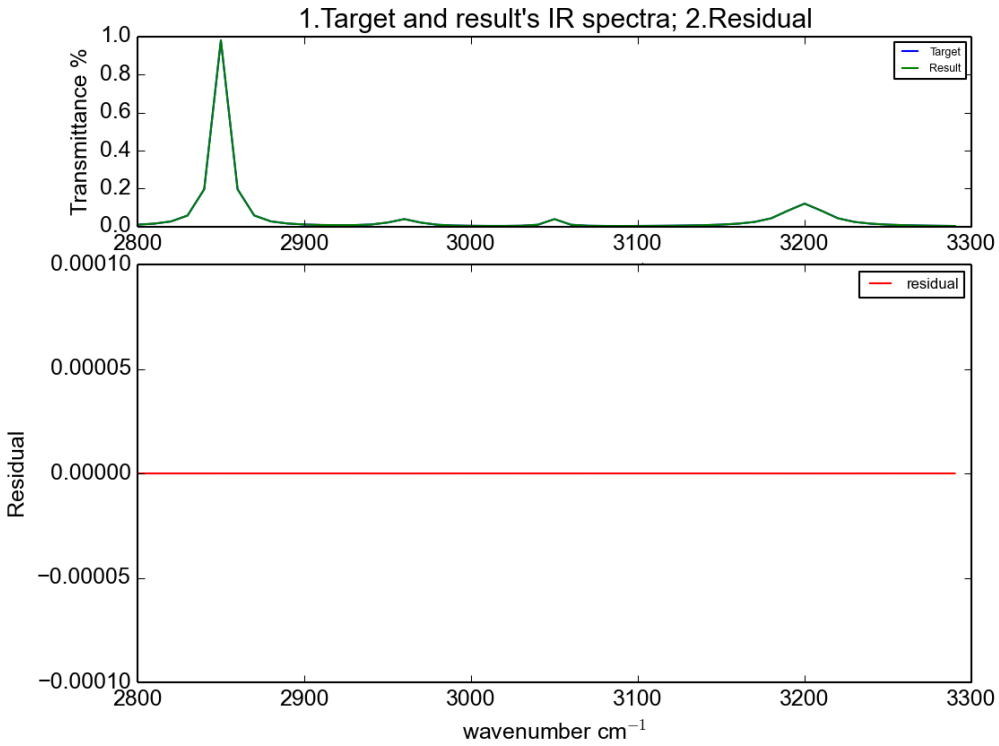

Figure 3.2: Toy Model Result Plotting for 4 Candidates on IR Cosine Projection

Table **??** indicates the return composition for Experiment 3 is different from the target one. Figure **??** shows that the spectrum produced by the return composition is almost identical to the one generated by the target composition. The residual is negligible. The result is the same as Experiment 2.

Among Experiment 1, 2 and 3, only Experiment 1 return composition matches its target one. For Experiment 2 and 3, the return composition is totally different from the target one. However, for Experiment 2, the difference in θ among the candidates is smaller than Experiment 1. For Experiment 3, the number of the candidates is larger than Experiment 1. Both effects increase the complexity of the experiments. Therefore, further increase the difficulty for LP model to return the target composition. for both Experiment 2 and 3, the spectra constructed by the return composition matches to the one built by the target composition.

Experiment 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(cos)
Return Composition	[0, 0, 0.730541, 0, 0.212061,0, 0, 0.0573978, 0, 0]

Table 3.2: Experiment 3 Setting

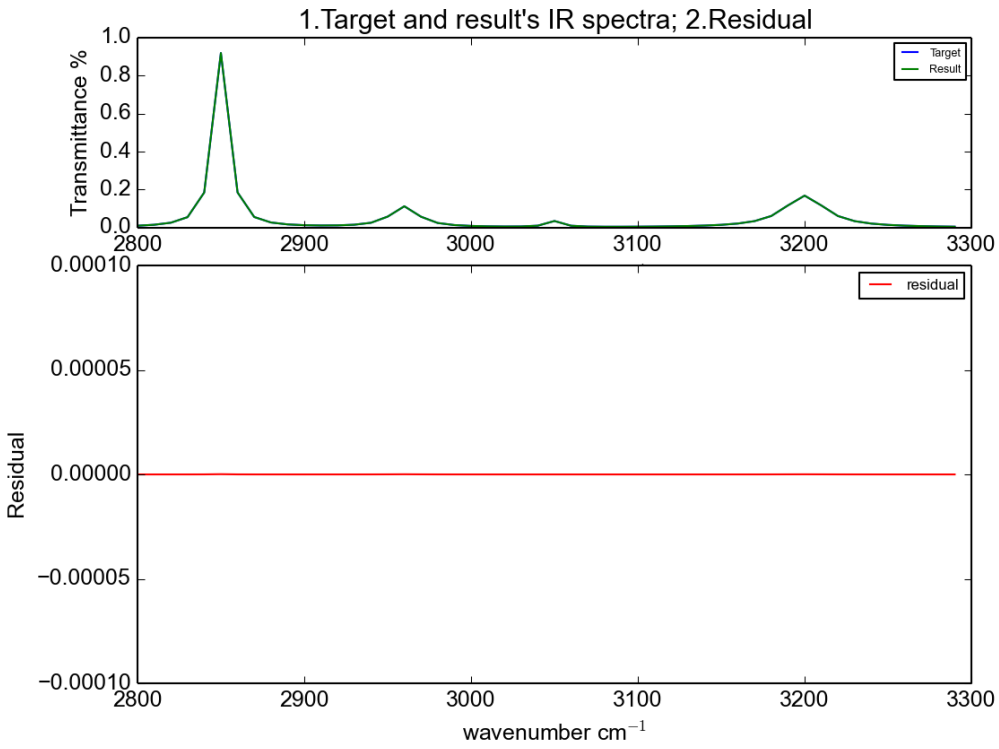

Figure 3.3: Toy Model Result Plotting for 10 Candidates on IR Cosine Projection

This demonstrates that there are multiple compositions can achieve to construct the spectra that are close to the target ones. However,the numerical limitation in the LP model helps the LP solver to converge to a unique one. Moreover, the reason for the LP model in Experiment 1 to return a composition that matches to the target one, is that the spectral information used to construct the LP model is competent. The constraints constructed in the LP model eventually converge to the target one, which results in better numerical outcome. The LP model is more complete compared to the one in Experiment 2.

In order to add the necessary information to construct the constraints in our LP model, IR's second polarization introduced for the toy model: the sine polarization. Figure **??** describes how the spectra are presented for 10 candidates same as Experiment 3. Experiment 4 and 5 will include both polarizations' spectral information when build the LP model. Table **??** displays the setting for Experiment 4 and 5. This setting is based on Experiment 2, with sine-polarization IR spectrum added. 100 data points are selected from this additional spectra, then converted to additional decision

Experiment 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(cos) + 100(sin)	100(cos) + 100(sin)
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: Experiment 4 and 5 Setting

variables and constraints in the LP model. Same with Experiment 5, it is based on Experiment 3, with sine-polarization IR spectral information added. For both Experiment 4 and 5, the return composition now matches to the target one. This further proves that as long as we have sufficing information to build the constraints, the LP solver will return a composition matches to the target one.

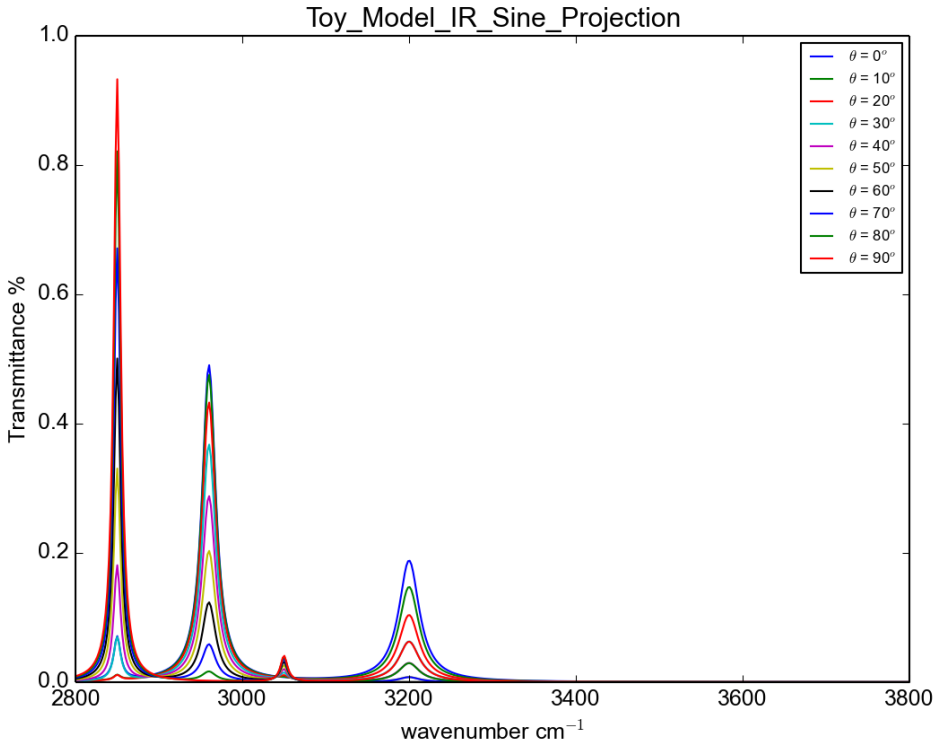

Figure 3.4: Toy Model Candidates IR Sine Projection

3.5 Constraint Study Based on Experiment 4

From Experiment 1 to 5, we know having sufficient information in our LP model is the key to obtain the target composition. Having sufficient information means having enough constraints to help LP model converge to a desired result. Moreover, the information is coming from the valuable data points selected along the spectra. This leads us to do a further study on the constraints in order to see how many data points are enough to get the desired composition.

Based on Experiment 4, experiments about formulating the LP model with different data information are conducted in Table **??**. The number of data points indicates how many data points are selected. Points Selection shows how data points are selected. [2800, 3300, 50] means along wavenumber from 2500 to 3300, every 25 wavenumber. For example, Experiment 6 contains 10 data points from cosine-polarized IR spectrum. Every 50 wavenumber, one data point is selected. Similarly, for Experiment 7, 8, 9, 10, 11, every 25, 20, 15, 10 and 5 wavenumber, one data point is select. From Experiment 12 to 16, data points are selected from both cosine-polarized and sine-polarized IR spectrum.

(TODO: rethink: What can we exactly get from the following two tables? Should we include this study?)

One interesting result from Table **??** is that: from Experiment 1 to 9, the result composition is the same. To the contrary, from Experiment 10, the return composition gets changed to the target one. Furthermore, if we plot the return composition of [0, 0.796962, 0.103038, 0.1] and the target one [0.1, 0.5, 0.4, 0] in Picture **??**. In this picture, we can see that the spectra generated by these two composition are identical.

3.6 Constraint Study Based on Experiment 5

When the same constraint study is applied to the data based on Experiment 5 in Table **??**, the observation is the same as the experiments in Table **??**. This further proves that: We can obtain different solutions by have different constraints. When the result composition [0, 0, 0.730541, 0, 0.212061,0, 0, 0.0573978, 0, 0] and target one [0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0] are plotted together, they are almost identical as well, as shown in Figure **??**.

Experiment Index	Number of Data Points	Points Selection	Result
6	10	[2800, 3300, 50]	[0, 0.796962, 0.103038, 0.1]
7	20	[2800, 3300, 25]	[0, 0.796962, 0.103038, 0.1]
8	25	[2800, 3300, 20]	[0, 0.796962, 0.103038, 0.1]
9	32	[2800, 3300, 15]	[0, 0.796962, 0.103038, 0.1]
10	50	[2800, 3300, 10]	[0, 0.796962, 0.103038, 0.1]
11	100	[2800, 3300, 5]	[0, 0.796962, 0.103038, 0.1]
12	$100 + 1$	[2800, 3300, 5], [2800, 3300, 500]	[0, 0.796962, 0.103038, 0.1]
13	$100 + 5$	[2800, 3300, 20], [2800, 3300, 100]	[0, 0.796962, 0.103038, 0.1]
14	$100 + 10$	[2800, 3300, 20], [2800, 3300, 50]	[0, 0.796962, 0.103038, 0.1]
15	$100 + 50$	[2800, 3300, 20], [2800, 3300, 10]	[0.1, 0.5, 0.4, 0]
16	$100 + 100$	[2800, 3300, 20], [2800, 3300, 5]	[0.1, 0.5, 0.4, 0]

Table 3.4: Constraint Study Based on Experiment 4

Experiment Index	Points	Point Selection	Result
17	10	[2800, 3300, 50]	[0.156758, 0, 0, 0.825977, 0, 0, 0, 0, 0, 0.017265]
18	25	[2800, 3300, 20]	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
19	50	[2800, 3300, 10]	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
20	100	[2800, 3300, 5]	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
21	500	[2800, 3300, 5]	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
22	$100 + 1$	[2800, 3300, 5], [2800, 3300, 500]	[0, 0, 0.730541, 0, 0.212061, 0, 0, 0.0573978, 0, 0, 0]
23	$100 + 10$	[2800, 3300, 5], [2800, 3300, 50]	[0.361587, 0, 0.312061, 0.326352, 0, 0, 0, 0, 0]
24	$100 + 20$	[2800, 3300, 5], [2800, 3300, 25]	[0.174023, 0, 0, 0.791447, 0, 0, 0.0345301, 0, 0, 0]
25	$100 + 25$	[2800, 3300, 20], [2800, 3300, 20]	[0.174023, 0, 0, 0.791447, 0, 0, 0.0345301, 0, 0, 0]
26	$100 + 50$	[2800, 3300, 5], [2800, 3300, 10]	[0, 0, 0.753209, 0, 0.146791, 0, 0.1, 0, 0, 0]
27	$100 + 84$	[2800, 3300, 5], [2800, 3300, 6]	[0.174023, 0, 0, 0.791447, 0, 0, 0.0345301, 0, 0, 0]
28	$100 + 100$	[2800, 3300, 5], [2800, 3300, 5]	[0.1, 0, 0.5, 0, 0.4, 0, 0, 0, 0, 0]

Table 3.5: Constraint Study Based on Experiment 5

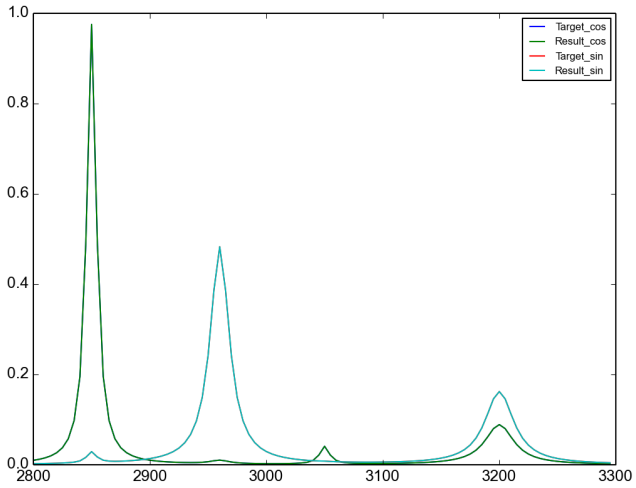

Figure 3.5: Toy Model Constraint Study 1

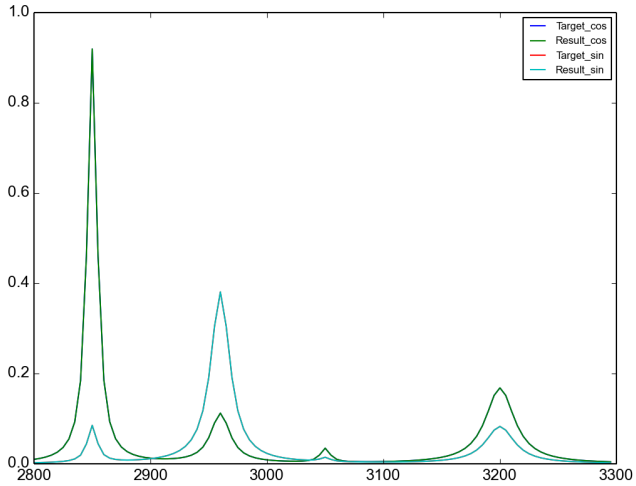

Figure 3.6: Toy Model Constraint Study 2

3.7 Discussion and Conclusion

With all the experiments conducted with the toy model, we have learnt that the reason, that our LP model does not return a composition that matches the target one, is that the model does not have sufficient information to build the constraints. However, with the limited information, the optimal solution returned by our LP model does build the perfect target spectrum. This means that the solution for the composition that achieves minimum residual of the objective function is not unique. However, in real experiment, because numerical restriction, an unique optimal solution is obtained

from the LP model.

Above analysis simulates the following question: how can we know there is enough information to achieve the target composition? In the next step, we will experiment with real molecules. The goal is to investigate with all the spectral information that we can obtain for real molecules, can our LP model return the target composition for the target spectrum? If yes, can we apply the LP model systematically? Furthermore, to maximally explore the capacity of our LP model, and study its limitation. Finally, come up with some general instructions for applying our LP model. These are the main focus for the following chapters.

Chapter 4

Realistic Molecular Model

4.1 Description

After experimenting with our toy problem, lacking sufficient information for the LP model is the key cause for the failure of obtaining the correct composition of the target spectra. First of all, in the toy model, there are only four vibrational modes, and the wavenumber range is limited. The number of data points selected is limited. Therefore, the constraints for the LP model is not sufficient. Secondly, the similarity among the candidates is high, as all the candidates are coming from one same molecule with the only difference in value θ. Third, the data points are only extracted from IR spectra.

In this chapter, real molecules are introduced. They contain more abundant spectroscopy information. In addition to IR, both Raman and SFG are introduced for real molecules, which makes the study one step closer to the overall goal and scope. Similar experiments applied to the toy model are now applied to a real molecule. The real molecule focused on this chapter is an amino acid - Methionine.

Same as the toy model, in order to limit the possible candidate space of Methionine, *twist* and *azimuthal* angular distributions are assumed to be isotropic, which are integrated. Only θ in Euler angles is considered in Methionine's surface orientation distribution function. In Chapter 2 section Generating model spectra, how a molecule's IR, Raman and SFG spectra are generated have been explained. Two unique IR spectra can be obtained from x, and z polarizations. Four unique Raman

Experiment index	1	2	3	4
Number of Candidates	4	4	4	4
Candidates	[0, 20, 40, 60]	[0, 20, 40, 60]	[0, 20, 40, 60]	[0, 20, 40, 60]
Target Composition	[0.1, 0.5, 0.4, 0]	[0.1, 0.5, 0.4, 0]	[0.1, 0.5, 0.4, 0]	[0.1, 0.5, 0.4, 0]
Number of Data Points	200(irx)	200(irz)	200(irx) + 200(irz)	200(irx) + 200(ramanxx)
Return Composition	[0.701654, 0, 0, 0.298346]	[0.701654, 0, 0, 0.298346]	[0.701654, 0, 0, 0.298346]	[0.1, 0.5, 0.4, 0]

Table 4.1: Experiment 1 to Experiment 4 Setting for Methionine Candidates

spectra can be obtained from xx, xy, xz and zz polarizations. Three unique SFG spectra can be obtained from yyz, yzy and zzz polarizations.

All the three spectroscopy techniques are applied in the experiments regarding real molecule. The goal is to see if those spectral information is enough to construct a LP model that returns the correct coordination distribution information about an amino acid at interfaces. If yes, we need to figure out what spectral information is needed to construct the LP model. If no, we need to check if the cause of the failure is the same as the toy model.

4.2 Experiments

Table **??**, four experiments are set up with four candidates and one same target composition. These four candidates each has θ of the following degree: 0°, 20°, 40° and 60°. The only difference among these four experiments is the spectroscopy information we select to construct the LP model, and it is indicated by the Number of Data Points. In Experiment 1, only IR x-polarization spectral information is used. This means that only data points from IR x-polarization are selected to build the LP model. Same for Experiment 2, data points are obtained from spectra of IR's z-polarization. In Experiment 3, the spectral information of IR's x and z-polarizations are combined. At last, for Experiment 4, spectral information of IR x-polarization and Raman xx-polarization are combined. The LP model we build for each experiment is different as the data points are selected differently. As the return composition indicates, Experiment 4 contains the most abundant information, as its return composition matches to the target one.

When merely using IR information, the return composition is the same for Experiment 1, 2 and 3. Figure **??** displays the result spectra generated by using the

return composition obtained from the first three experiments. The resulting spectra is almost identical to the target ones. It again proves that with the information only coming from IR spectra is not sufficient to get the target composition. However, the return composition could perfectly re-product the target spectra. This indicates that we would need further information for the constraint of LP model, in order to further refine the result. The more constraints are introduced, the more accurate the return composition will be.

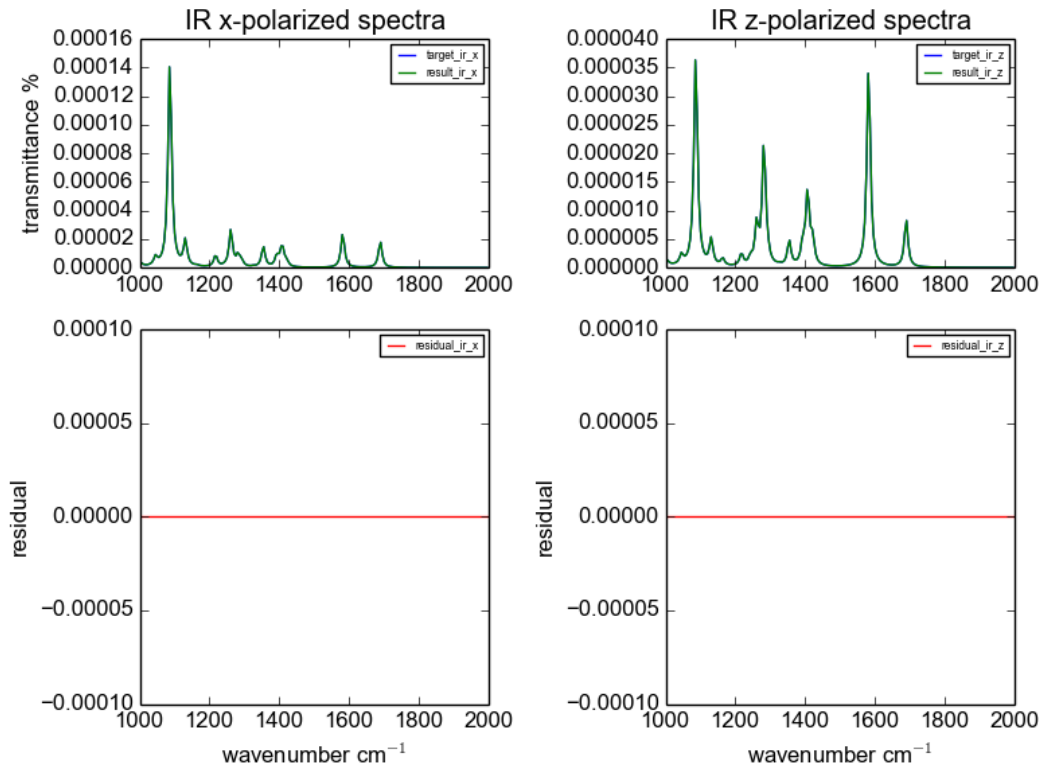

Figure 4.1: Compare target spectra with spectra generated by composition returned by LP model with only IR spectra of x and z projection

For the setting in Experiment 1 to 4, the LP model that constructed from combining IR and Raman spectral information is sufficient to obtain the target composition. When the difference in θ degree for candidates is smaller than 20°, which is 10°, we need to check if Raman and IR together still sufficient enough to derive the target composition. Therefore, the following experiments are conducted as shown in Table **??**.

Number of Candidates	4	
Candidates	[0, 10, 20, 30]	
Target Composition	[0.1, 0.5, 0.4, 0]	
Experiment index	Number of Data Points	Result Composition
5	200(irx) + 200(irz)	[0.752528, 0, 0, 0.247472]
6	200(irx)+200(irz)+ 200(ramanxx)	[0.1, 0.5, 0.4, 0]
7	200(ramanxx)+ 200(ramanxy)+ 200(ramanxz)	[0.1, 0.5, 0.4, 0]
8	200(ramanxx)+ 200(ramanxy)+ 200(ramanzz)	[0.1, 0.5, 0.4, 0]
9	200(ramanxx)+ 200(ramanxy)+ 200(ramanxz)+ 200(ramanzz)	[0.1, 0.5, 0.4, 0]

Table 4.2: Experiment 5 to Experiment 9 Setting for Methionine Candidates

Experiment 5 shows that the LP model constructed by merely using IR spectral information is not sufficient enough to derive the target composition for the current candidate setting. Experiment 6 indicates that combining IR and Raman spectral information helps to derive the target composition. What's more, Experiment 7 to 9, illustrates that Raman spectral information itself is sufficient to obtain the target composition as well.

For experiment setting in Table **??** and Table **??**, combining IR and Raman spectral information to construct a LP model is sufficient enough to obtain the target composition. In order to study the limitation of the LP model, the complexity of the experiment setting needed to be increased. Therefore, another group of experiments have been designed as shown in Table **??**. There are 5 candidates included in the experiments. Each candidate has θ with the following degree: 0°, 10°, 20°, 30° and 40°. The target composition is more complex than previous experiments, each candidate takes 20% in the mix.

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

As the result of Experiment 10 to 16 indicates that even combining all the spectral information of IR, Raman and SFG, it is still not sufficient to attain the target composition for the experiments set up in Table **??**. The LP model is showing its limitation in these experiments. In order to confirm if the reason causing the LP model to return a different composition is because of insufficient information. Further experiments are conducted as shown in Table **??**.

Number of Candidates	5	
Candidates	[0, 10, 20, 30, 40]	
Target Composition	[0.2, 0.2, 0.2, 0.2, 0.2]	
Experiment index	Constraints	Result
10	200(irx) + 200(irz)	[0.607766, 0, 0, 0, 0.392234]
11	200(ramanxx) + 200(ramanxy) + 200(ramanxz) + 200(ramanzz)	[0.247792, 0, 0.502139, 0, 0.250069]
12	200(sfgyyz) + 200(sfgyzy) + 200(sfgzzz)	[0.321014, 0, 0.31018, 0.163041, 0.205764]
13	200(irx) + 200(irz) + 200(ramanxx) + 200(ramanxy) + 200(ramanxz) + 200(ramanzz)	[0.247792, 0, 0.502139, 0, 0.250069]
14	200(ramanxx) + 200(ramanxy) + 200(ramanxz) + 200(ramanzz) + 200(sfgyyz) + 200(sfgyzy) + 200(sfgzzz)	[0.321014, 0, 0.31018, 0.163041, 0.205764]
15	200(irx) + 200(irz) + 200(sfgyyz) + 200(sfgyzy) + 200(sfgzzz)	[0.321014, 0, 0.31018, 0.163041, 0.205764]
16	200(irx) + 200(irz) + 200(ramanxx) + 200(ramanxy) + 200(ramanxz) + 200(ramanzz) + 200(sfgyyz) + 200(sfgyzy) + 200(sfgzzz)	[0.321014, 0, 0.31018, 0.163041, 0.205764]

Table 4.3: Experiment 5 to Experiment 9 Setting for Methionine Candidates

Number of 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]	
Experiment index	Number 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 4.4: Experiments to Explain the Limitation of LP Model for Methionine Molecule

4.3 Experiments to Explain the Limitation of LP Model for Methionine Molecule

In order to further explore the reason that LP model reaches its limitation for the real molecule, Experiment 17 and 18 are conducted. Methionine candidates are still used. To make the study case more general than Experiment 1 to 16, candidates' θ values are expanded from 0° to 80°. In total, there are 9 candidates. Because the SFG spectra for θ of 90° is a straight line, it is excluded from all the experiments. For target composition, five candidates are randomly selected to be presented. The difference between Experiment 17 and 18 is that different amount of data points are selected to build the LP model. From all three spectroscopy techniques' spectral information, every 5 wavenumber a data point is selected for Experiment 17. Every 500 wavenumber a data point is selected for Experiment 18. As a result, Experiment 17 and 18 each returns a different composition. Both compositions do not match to the target one.

However, for both Experiment 17 and 18, when the return composition is used to generate the IR, Raman and SFG spectra, and plotted together with the spectra created by target composition. All the spectra are almost identical for IR, Raman and SFG. Figure **??**, **??** and **??** display the spectra plotted by using return composition and target one for Experiment 17. Every spectrum is almost identical to each other in the figures. Same for Experiment 18 as shown in Figure **??**, **??** and **??**. These figures prove again that there are more than one composition that can perfectly construct the target spectra. The information to construct the LP model is not sufficient to converge to the one exactly matches to the target composition. This conclusion exactly fits the result obtained from the experiments have done with the toy molecule.

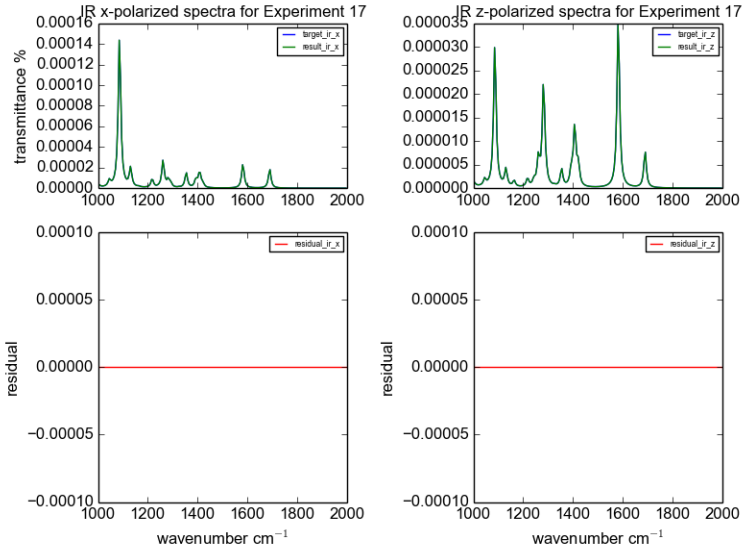

Figure 4.2: IR spectra plotted by using target composition and return composition of Experiment 17

4.4 Extra Experiments

TODO: this part of experiments are similar as what are done in Chapter 5 and 6. Think how to involve this part properly.

From Experiment 1 to 18, LP model helps to return the target composition for some cases, and not for others. We want to figure out if there a clean line indicating the information used to generate the LP model is not sufficient to obtain the target composition for one molecule. In order to answer this question, more systematic experiments needed to be organized. Therefore, the following experiments are conducted. The Methionine candidate space is the same as Experiment 17 and 18. spread from 0° to 80° on θ.

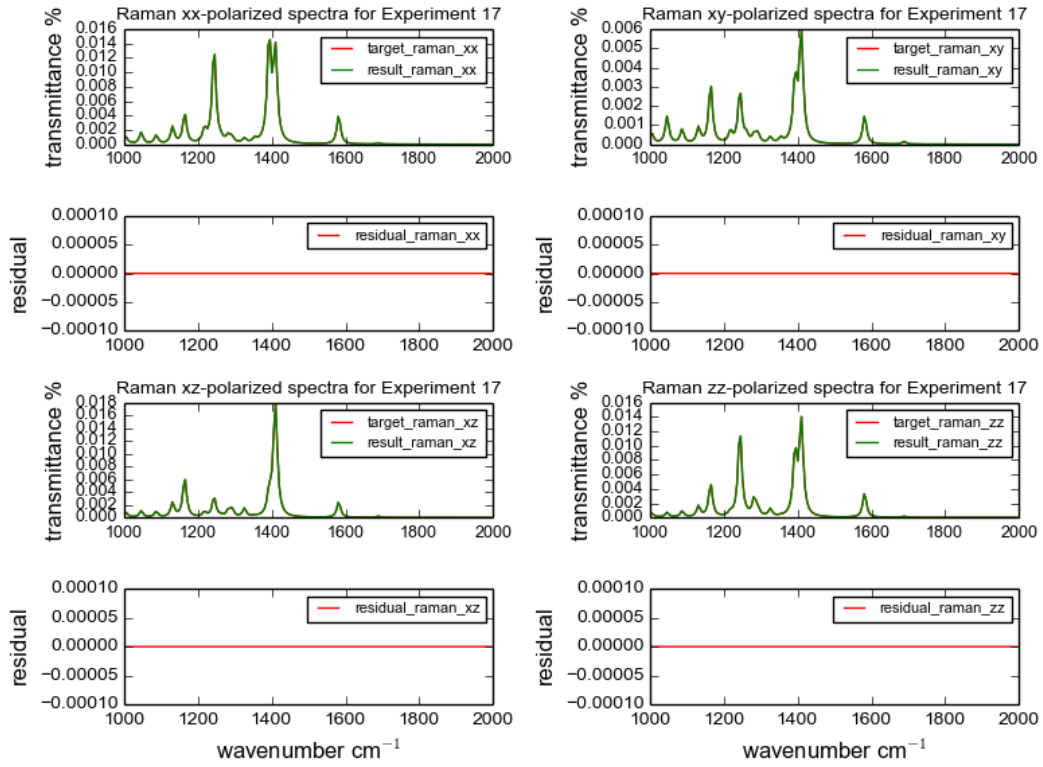

Figure 4.3: Raman spectra plotted by using target composition and return composition of Experiment 17

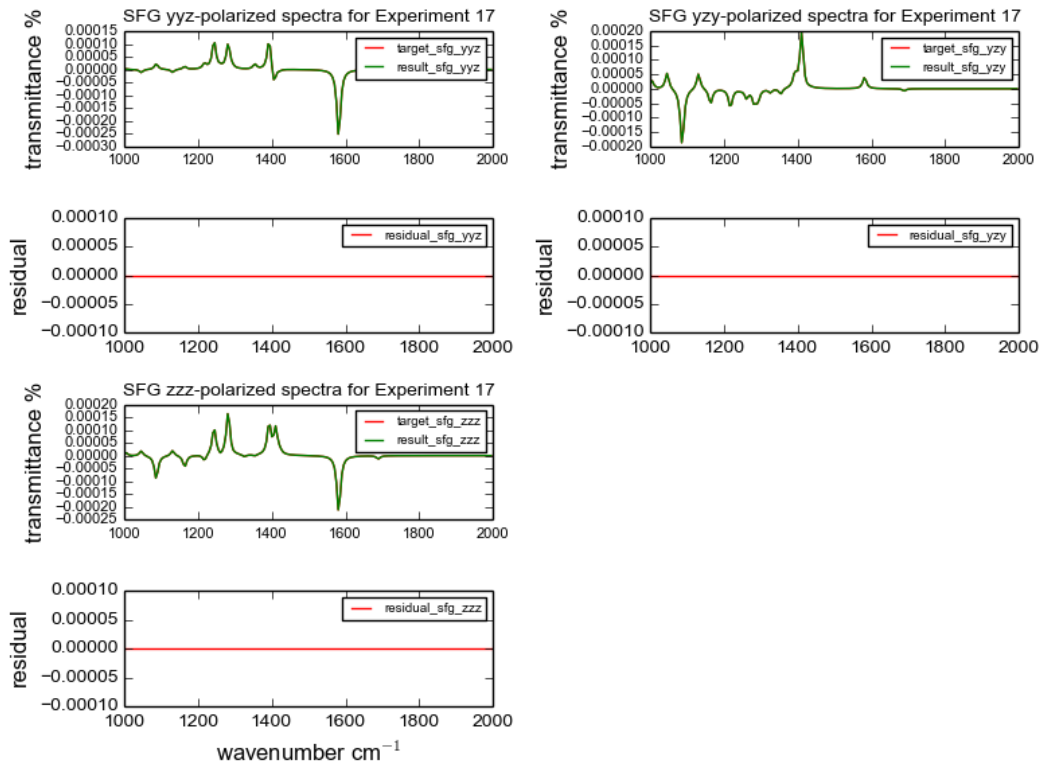

Figure 4.4: SFG spectra plotted by using target composition and return composition of Experiment 17

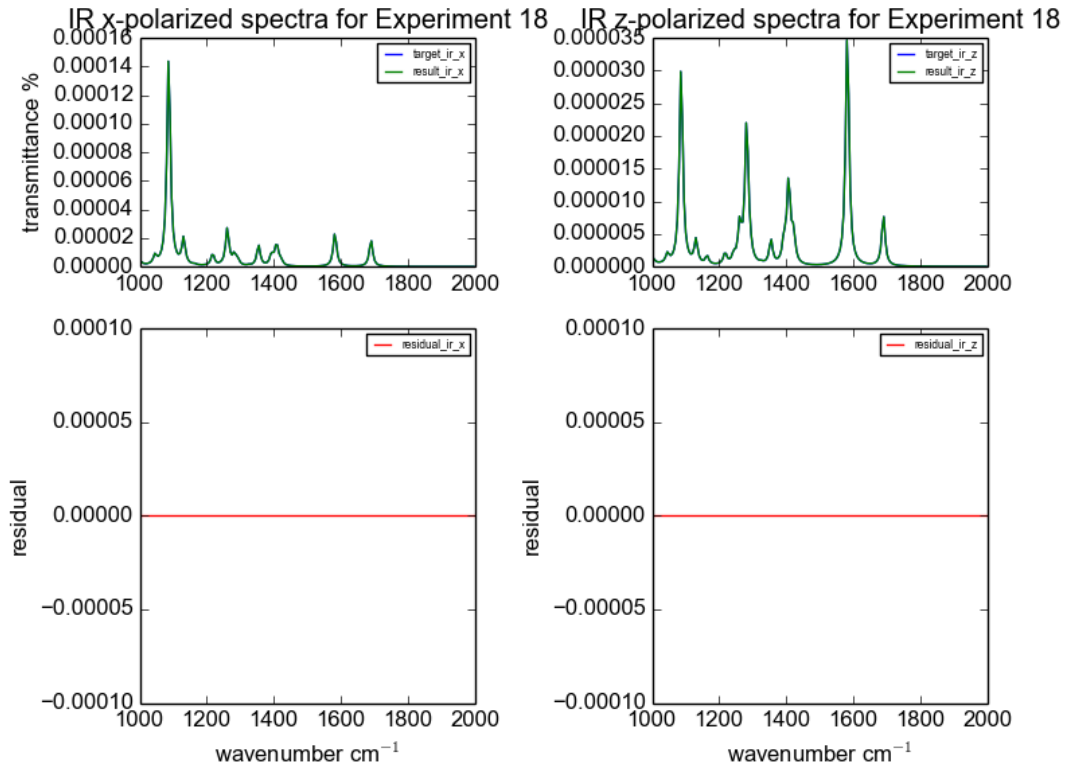

Figure 4.5: IR spectra plotted by using target composition and return composition of Experiment 18

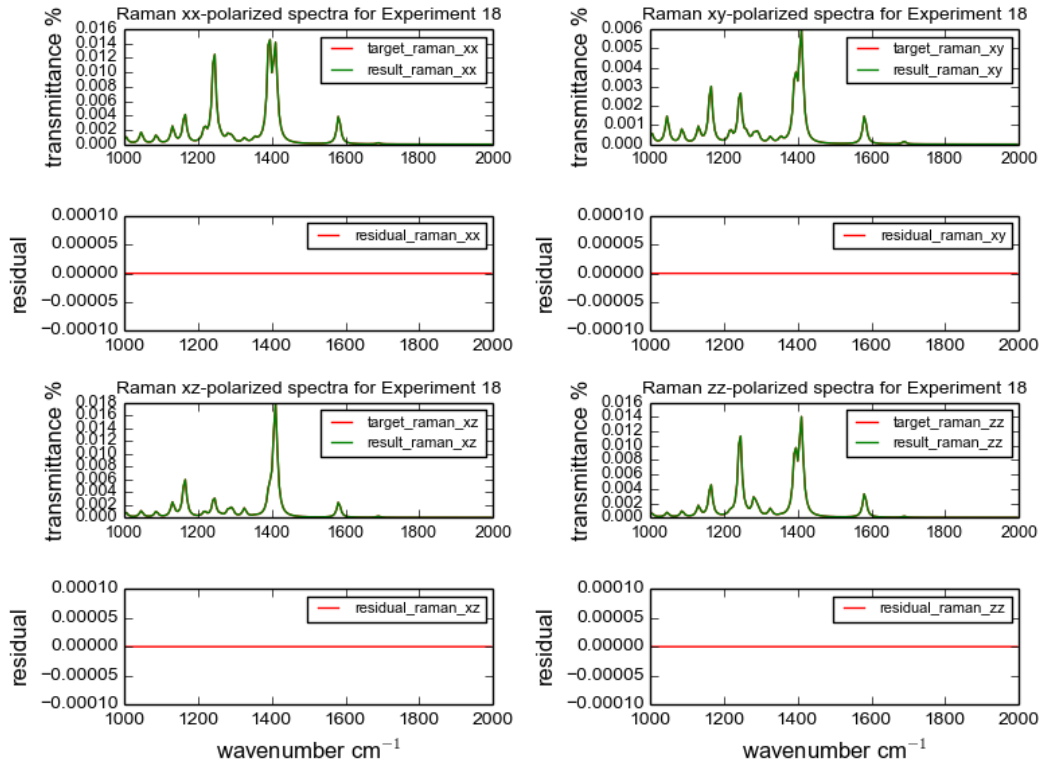

Figure 4.6: Raman spectra plotted by using target composition and return composition of Experiment 18

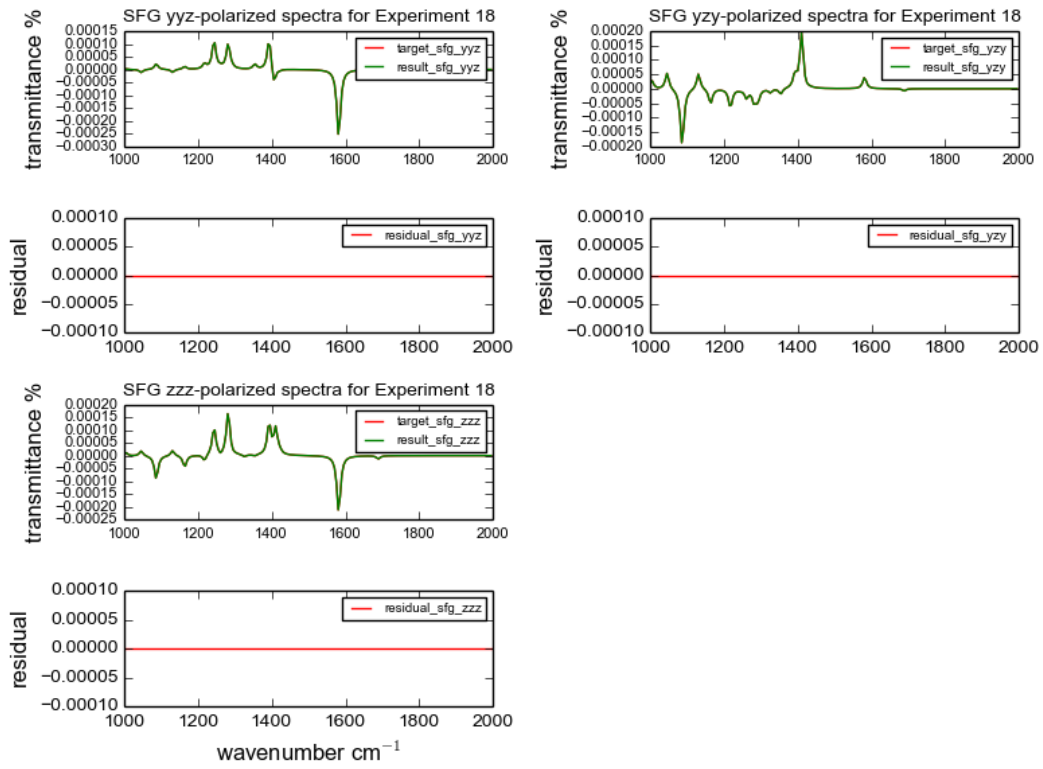

Figure 4.7: SFG spectra plotted by using target composition and return composition of Experiment 18

Chapter 5

Mixture

5.1 Description

In Chapter 4, experiments indicate that for one type of molecule at interfaces, even combining all the three spectral information, the constructed LP model cannot return the target composition in most cases. The existing spectral information is not adequate to obtain the target composition of molecule coordination distributions at interfaces. Multiple return compositions can build the spectra that are almost exactly the same as the target ones. These compositions are returned by different LP models that use different amounts of spectral information. In each LP model, because of numerical limitation, it returns an optimal composition solution.

For one type of molecule, even combining all three spectral information is limited in obtaining the target composition. It seems that our LP models have hit its limitation. However, there is another factor of we want to focus: the study of multiple different molecules at interfaces. For a mixture of different molecules at interface, we want to figure out whether our LP models can help to obtain the target composition. If the LP models success in obtaining the target composition, then its rate of accuracy is the key factor of the study as well.

5.2 Experiments

To achieve the study of mixture molecules at interfaces, further experiments are constructed. The experiments have the following common settings.

First, there are six different amino-acids in the mixture: methionine, leucine, isoleucine(ile), alanine, threonine and valine. For each amino acid, only θ difference is considered, the other two Euler angles are integrated. Each amino acid molecule has 9 candidates in the mixture, each with θ of the following degrees: 0°, 10°, 20°, 30°, 40°, 50°, 60°, 70° and 80°. Because when θ equals 90°, the SFG spectra is a straight line. The corresponding candidate is excluded from all the experiments. As a result, there are 54 candidates in the mixture.

Second, the target composition need to be generated. The operation includes two steps: randomly pick one candidate from each amino acid's 9 candidates, and randomly generate a percentage for the selected candidate. The target composition is made of six randomly selected candidates coming from six different amino acids. The rest 48 candidates have 0 percentage in the target composition. Namely, six selected candidate makes 100% component of the mixture.

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

Each experiment in the experiment set contains different spectroscopy information as shown in Table **??**. In Experiment 1, candidates' IR x and z polarization spectra are obtained. The target's IR x and z polarization spectra are generated by dot product of the target composition and all the candidates' spectral data. Then the corresponding LP model is conducted using Equation 3.4. Therefore, we claim that the LP model in Experiment 1 only contains IR information.

Similarly, Experiment 2 only contains Raman spectral information in the following four polarizations: xx, xy, xz and zz. Experiment 3 only contains SFG spectral information of yyz, yzy and zzz three polarizations.

Starting from Experiment 4, spectral information of different spectroscopy tech-

Experiment Index	Spectrum Information
Experiment 1	x and z polarized IR spectra
Experiment 2	xx, xy, xz and zz polarized Raman spectra
Experiment 3	yyz, yzy and zzz polarized SFG spectra
Experiment 4	x and z polarized IR spectra; xx, xy, xz and zz polarized Raman spectra
Experiment 5	x and z polarized IR spectra; yyz, yzy and zzz polarized SFG spectra
Experiment 6	xx, xy, xz and zz polarized Raman spectra; yyz, yzy and zzz polarized SFG spectra
Experiment 7	x and z polarized IR spectra; xx, xy, xz and zz polarized Raman spectra; yyz, yzy and zzz polarized SFG spectra

Table 5.1: Detailed Experiment Group Setting

niques are combined. In E4, IR spectral information is combined with Raman. In Experiment 5, IR spectral information is combined with SFG. In Experiment 6, Raman and SFG spectral information are incorporated. At the end, in E7, all three spectral information are put together: IR, Raman and SFG.

The LP models in each experiments are built using the same formula as shown in Equation 3.4. Because every experiment is built with different spectral information, each LP model is different.

Finally, this experiment set is run 100 times in order to see which experiment in Table **??** gives the target composition with the highest accuracy. This accuracy is measured by the time of each experiment returns the target composition. The scoring mechanism to measure whether a return composition matches to the target one is described in the next section.

5.3 Scoring methods

At the first glance, 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, in most experiments conducted earlier, the spectra generated by the return composition are almost identical to the ones created by the target one. Therefore, this sum of residuals is also negligible. It is not appropriate to use it as a scoring criteria.

Another way to measure the accuracy of the return composition, is to compare it directly with the target one. Therefore, calculating the sum of the residuals between a target composition and a return one directly is a faster approach to evaluate the accuracy of each experiment. The shortage of this approach is that it cannot be used to measure in real experiments where the target composition is unknown. However, in the current experiments, this approach can be a way to evaluate the different spectroscopy techniques sensitivity in molecular orientation study at interfaces.

The return composition of each experiment in the experiment 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 $1e-7$. Then the return composition is considered to be the same as the target one.

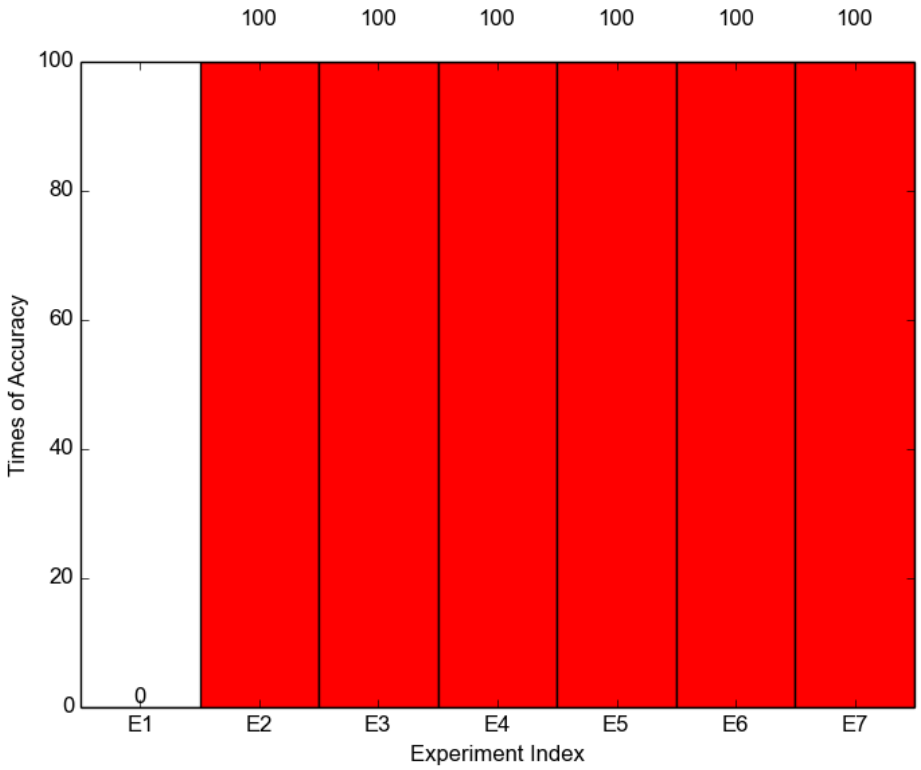

Figure 5.1: Accuracy analysis for experiments considering a mixture of amino acids with candidates from 0° to 80° on θ for each amino acid

The experiment set is ran for 100 times, the result is shown in Figure **??**. In Experiment 2, the return composition of the LP model constructed by using Raman spectral information meets the target composition 100 times. This means that within this set of experiments, Raman is sufficient to obtain the correct composition of the target spectra. Moreover, the accuracy is 100%.

Experiment 3 is the LP model constructed by using SFG spectral information. Its accuracy is 100% as well as shown in Figure **??**. This indicates that SFG spectral information is as abundant as Raman for this set of experiments. .

Experiment 4, Experiment 5, Experiment 6, and Experiment 7 each contains either spetral information of Raman, or SFG, or both. Therefore, the corresponding

LP model can help to get target composition with the same accuracy as Experiment 2 and Experiment 3.

The only exception is Experiment 1. The accuracy is not as high as the other experiments. The accuracy of LP model constructed using IR spectral information is 0. The low performance can be caused by the insufficient spectral information of IR.

When this experiment set is re-run 100 times, only Experiment 1's returned composition is analyzed and focused. In each run, IR x and z polarized spectra are plotted both by the returned composition and the target one. The result is these two polarizations' spectra conducted by the two different compositions are almost identical to each other in every run. For example, a random run is picked, then the two polarizations' spectra are plotted in Figure **??**. The spectra plotted by the return composition are identical to the ones plotted by the target composition. This indicates that the optimum composition returned by the LP model conducted with only IR spectral information has achieved its best in obtaining a composition that best fit the target spectra.

(TODO: rewrite or remove this paragraph) Comparatively, SFG has three unique polarizations, and Raman has four unique polarizations. From each projection's spectrum, we evenly select 200 data points. This means that one more projection will bring in 200 more constraints or 400 more (if we take the absolute sign off) constraints to the LP model. This would make a huge difference in LP model, in term of further refining the candidate selection in target composition. However, it is still too early for us to say that Raman has more coordination information because it has four unique polarizations. Because for Raman's any polarization, the spectrum of candidate with θ equals to one degree is identical to the one of candidate with this θ degree's complementary. For example, the Raman spectra for candidate with θ of 10°, is the same as candidate with θ of 170°. And for IR, it is the same case. Only SFG tells the differences between these two degrees, as the spectra for candidate with θ of one degree is symmetric to its complementary along wavenumber.

To further study the capacity of the LP models built for the mixture of molecules,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 experiments in Table **??** is used. The only difference is randomly select

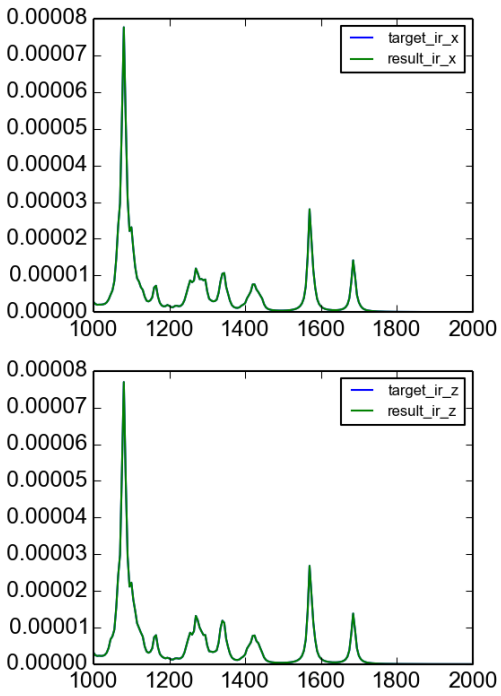

Figure 5.2: IR Spectra Plotted by Result Composition and Target Composition. TODO: add residual graph

one candidate from 18 candidates, instead of 9. 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 Experiment 1 is still low. This is not surprising as the complexity of the candidates has increased. Moreover, IR spectra for candidate with θ of one degree is identical to the one with θ of this degree's complementary, as shown in Figure 5.7. This also increases the difficulty for the LP model constructed by using IR spectral information to return the target composition.

However, it should be noticed that the accuracy for Experiment 2 has dramatically dropped. This is because the Raman spectra for one candidate with θ on one degree is identical to the one with this degree's complementary. Further explanation is provided in Chapter 6 .

Also in Figure 5.3, the accuracy for Experiment 3 is no longer high neither. After increasing the number of amino acid candidates from 9 to 18, the complexity of

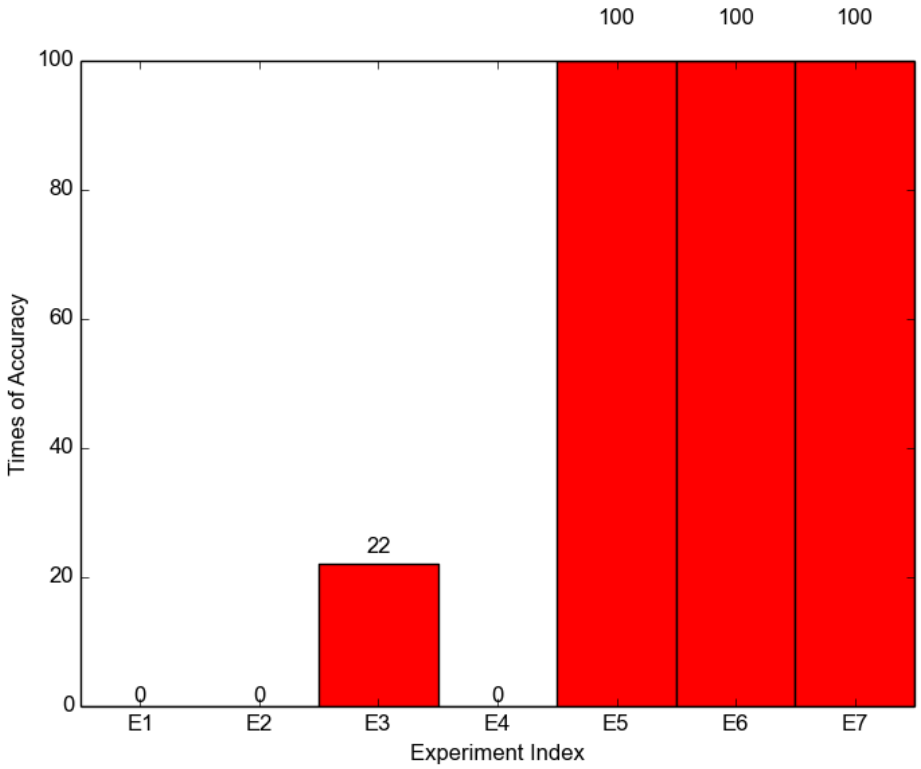

Figure 5.3: Accuracy analysis for experiments considering a mixture of amino acids with candidates from 0° to 180° on θ for each amino acid

the corresponding LP model has increased. From each projection, 200 data points are selected from both the target and the candidates' spectra. Therefore, while the number of constraints is the same as before, the number of candidates are twice bigger than before. Although the added candidates' SFG spectra are symmetric along wavenumber which may greatly increase the uniqueness of the candidates. The spectral information is still insufficient to converge the composition to the target one.

The good result starts to emerge when using the combinations of IR and SFG or Raman and SFG. Figure 5.3 shows that Experiment 5, Experiment 6, Experiment 7 all have 100% accuracies. This phenomenon can be explained as follow: SFG helps to distinguish a candidate from its complementary on θ value. The extra spectral information coming from IR or Raman helps to further refine the LP model, which can then converge the return composition to the target one.

Although the accuracy in Experiment 2 is low when each amino acid's candidates

spread from 0° to 180° on θ. There are still some noticable result in the return composition: for each amino acid, the percentage assigned to it is correct; however, the candidate presented may be the one with the correct degree, or the one with the correct degree's complementary. For example, a random run is selected. Figure **??** displays the target composition and Figure **??** displays the return composition of Experiment 2. Figure **??** is the return composition of Experiment 6. From the three figures, when extracting the non-zero values to generate a list, the three lists are the same. However, when overlapping Figure **??** with Figure **??**, the position of each non-zero value is not identical. For example, value of 0.299586 appears at θ of 150° for Valine in Figure **??**. In Figure **??**, it appears at θ of 30° for Valine. Same for value of 0.021196, 0.00662804, 0.000642609, and 0.00789. The LP model of E2 fails to tell which candidate is the exact one between the correct one and its complementary on θ's degree. This observation is a general case across all the experiment groups. From Experiment 6, as long as the spectra data from SFG is plugged into the LP model, the return composition is the same as the target one.

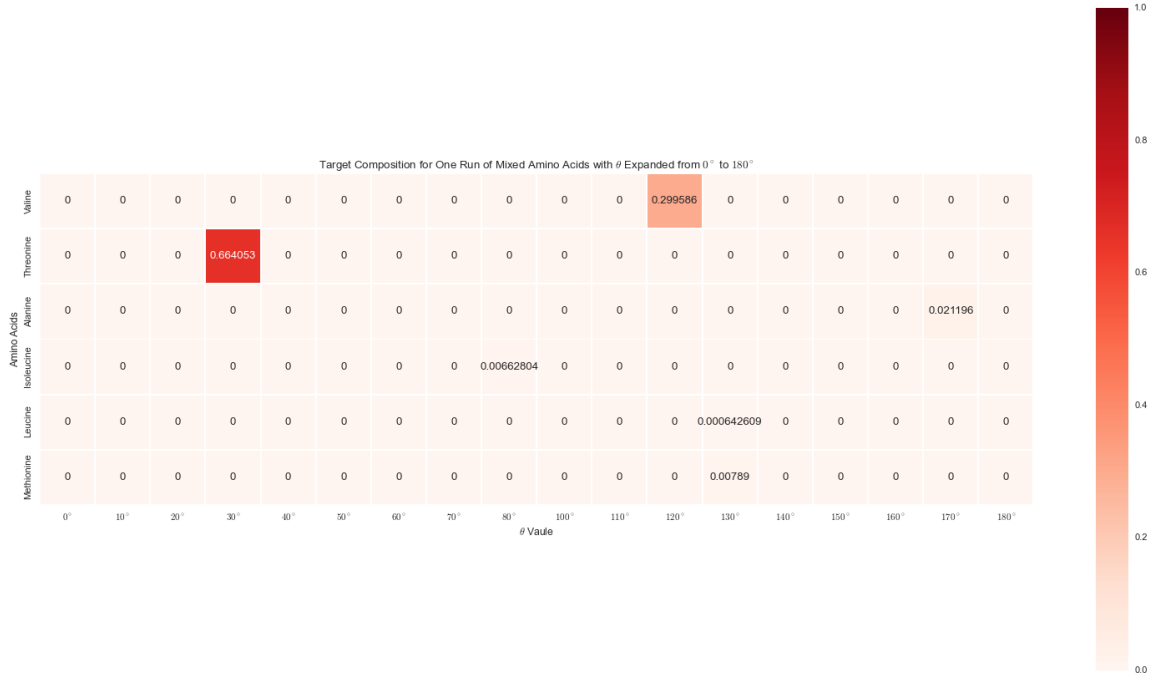

Figure 5.4: Target Composition for One Run of Mixed Amino Acids with θ Expanded from 0° to 180°

From the above analysis, Experiment 2 appears the ability of limiting the number

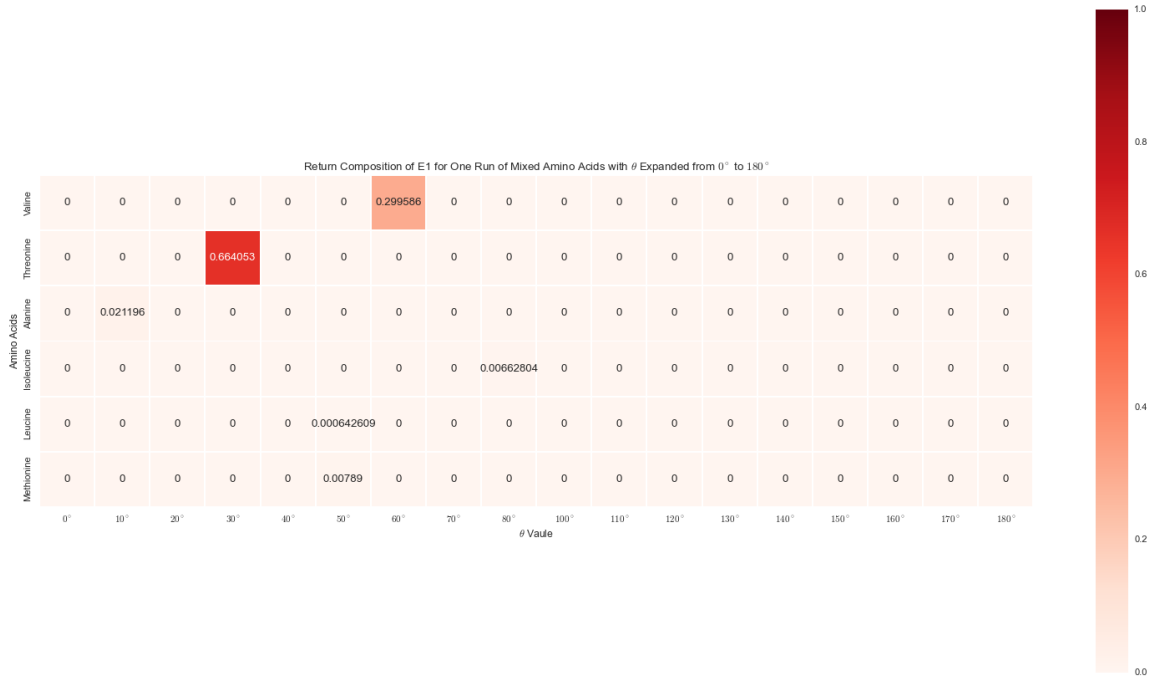

Figure 5.5: Return Composition of Experiment 2 for One Run of Mixed Amino Acids with θ Expanded from 0° to 180° on θ

of candidates to 2 for each amino acid. These two candidates are complementary on θ degree, with one of them to be the correct one for the target composition. The return composition of Experiment 4 is the same as the one of Experiment 2, which means IR spectra information is not helping in this case. Spectral information from SFG is needed in order to study the cases that having θ expanded from 0° to 180°.

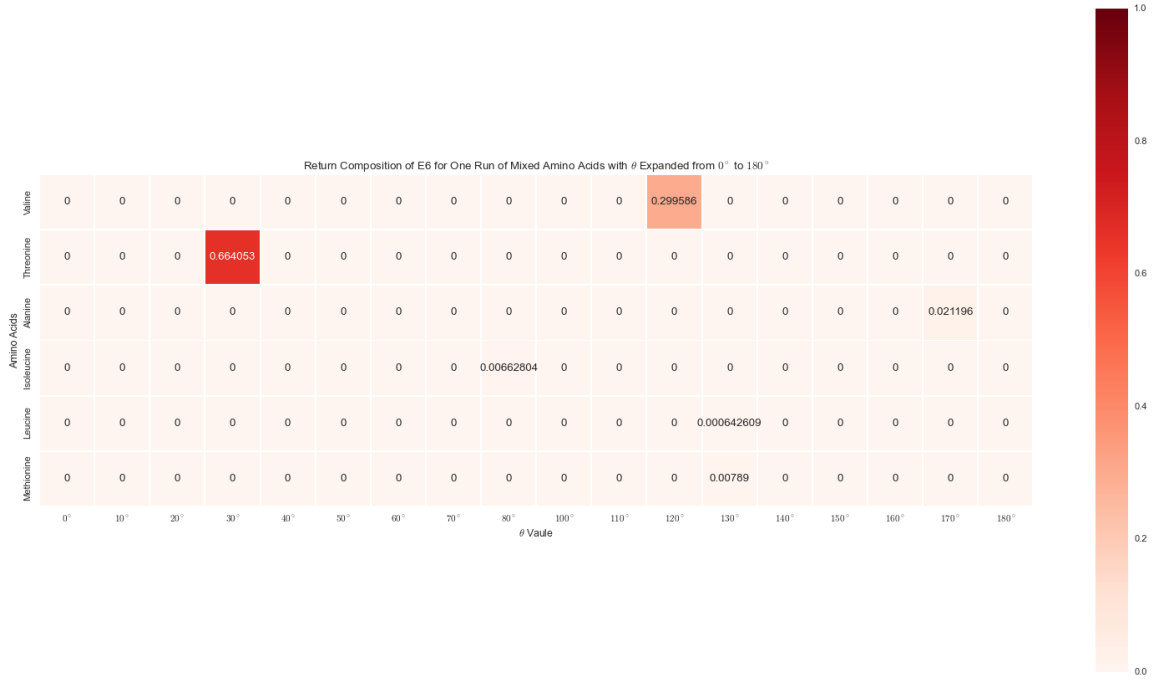

Figure 5.6: Return Composition of Experiment 6 for One Run of Mixed Amino Acids with θ Expanded from 0° to 180° on θ

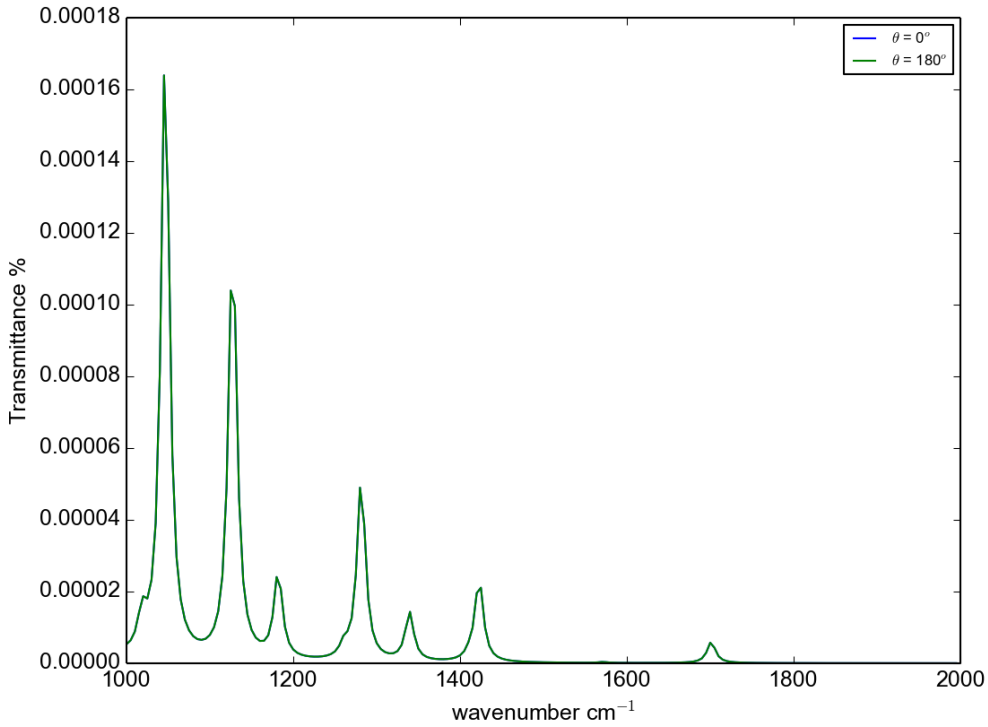

Figure 5.7: IR z projection spectrum for Alanine Candidate with θ of 0° is identical to Alanine Candidate with θ of 180°

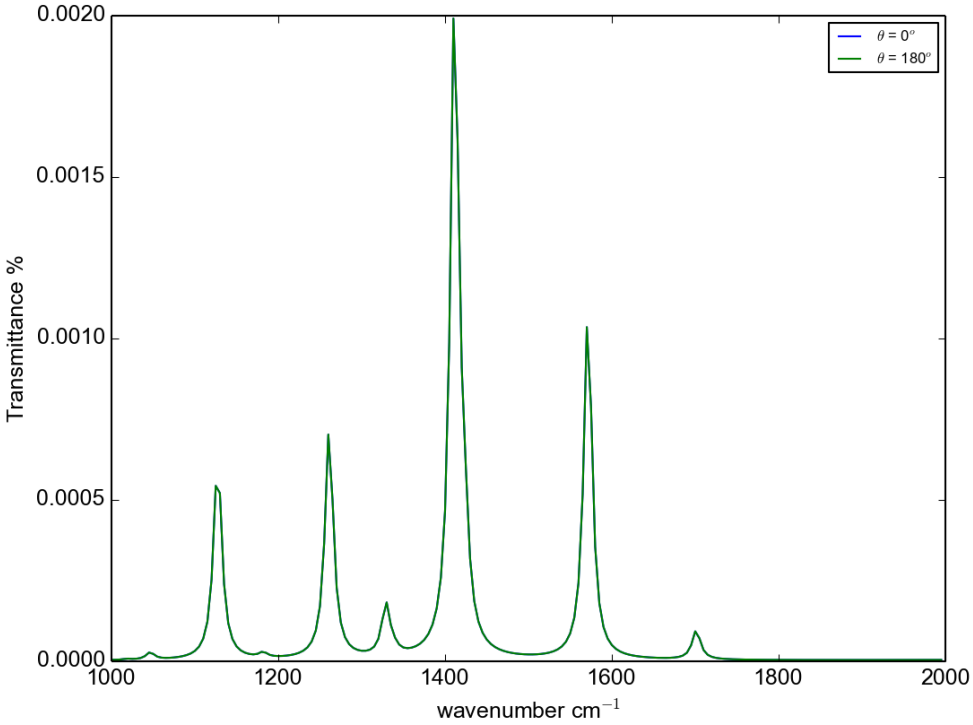

Figure 5.8: Raman zz projection spectrum for Alanine Candidate with θ of 0° is identical to Alanine Candidate with θ of 180°

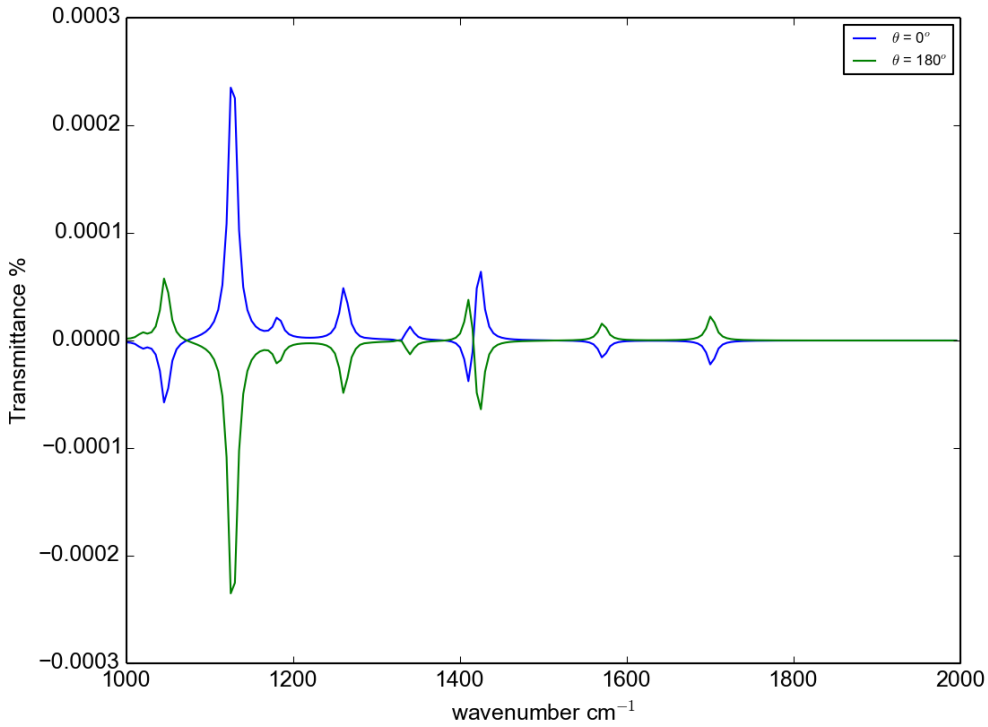

Figure 5.9: SFG zzz projection spectrum for Alanine Candidate with θ of 0° is not identical to Alanine Candidate with θ of 180°, but symmetric along wavelength

Chapter 6

Possibilities for treating experimental data

6.1 Description

The experimental spectral data that we obtain from IR, Raman or SFG techniques have a amplitude scaling factor when comparing to the candidate spectra that we generate mathematically. This means that between candidates' theoretical spectra and the target one, there is an unknown scaling factor. However, this scaling factor is the same for any spectra obtained within a particular spectroscopy technique. For example, for IR, the scaling factor for spectrum of x projection is the same as the one for spectrum of z projection. Therefore, we need to introduce this scaling factor to our LP models. Since the LP models constructed by the 7 experiments(E2, E3, E4, E5, E6, and E7 for θ ranged from 0° to 80°) in Chapter 5 are doing well in retrieving target composition for the mixed amino acids, we would like to know if the same LP models can be applied directly to real experimental data for the same θ range.

The experiment set-ups are the same as what is listed in Table 5.1 in Chapter 5. A group of 7 experiments with different spectroscopy information. The goal is also the same as what we have in Chapter 5, we want to figure out which spectroscopy technique's data will help us to retrieve the correct composition when we combine spectroscopy information with LP models. The only difference is that, for each experiment group, we will generate an arbitrary scaling factor for IR, Raman and SFG respectively. Therefore, the target spectra is not only composed by the percentage

composition of all candidates, but also need to multiple it by a randomly generated scaling factor.

In addition, to start with, we limit the scaling factor to be smaller than 1.

After a few runs of the experiment group, we observe that the returned compositions always contains one extra variable for each experiment.For E2, E4, E6 and E7, the models that are doing well in Chapter 5, the returned composition contains the right candidates, however, the percentages of the candidates are different from the target one. If we take a further look, the ratio between returned percentage and the generated percentage equals one when it adds to the extra variable. For example, taking E2's data from one experiment group as an example, the percentages of candidates we use to generate the target spectra is shown as Array 6.1. And the return percentages is what's in Array 6.2.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.03218 & 0 \\ 0 & 0.73929 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.19745 & 0 & 0 & 0 & 0 & 0 \\ 0.00173 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01819 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01116 & 0 & 0 \end{bmatrix} \tag{6.1}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.019308 & 0 \\ 0 & 0.443574 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.11847 & 0 & 0 & 0 & 0 & 0 \\ 0.001038 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.010914 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.006696 & 0 & 0 \\ 0.4 & & & & & & & & \end{bmatrix} \tag{6.2}$$

Comparing Array 6.2 with Array 6.1, there is one extra variable in Array 6.2, which is 0.4. As Array 6.3 shows, the ratio between every non-zero return percentage and every non-zero target percentage is the same which is 0.6. Moreover, when we

add 0.6 up with last variable 0.4, we get a total of 100%. Consistent with LP terminology, we call this last variable slack variable (Theory backup here).

From the above observation, we know that the slack variable always equals 1 minus the scaling factor. And the ratio between the return percentage of existing candidate and target percentage equals to scaling factor. This meets the theory of LP when scaling factor is smaller than 1 (Theory backup here).

$$\frac{0.019308}{0.03218} = \frac{0.443574}{0.73929} = \frac{0.11847}{0.19745} = \frac{0.001038}{0.00173} = \frac{0.010914}{0.01819} = \frac{0.006696}{0.01116} = 0.6 \quad (6.3)$$

Based on this observation, we want to know if the conclusion can be applied generally. If it can, which experiment in the group will help us to achieve this conclusion. In addition, with how much accuracy.

After running the experiment group 100 times, we have obtained the following picture Picture **??**.

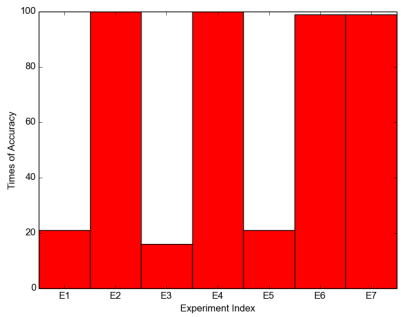

Figure 6.1: Experiment Accuracy Analysis for Experiments using experimental spectra data that contains scaling factor that is smaller than 1 and candidates with θ from 0° to 80°

As the picture indicates the LP model that built by using only Raman data is sufficient to help us to meet the above conclusion. The scaling factor equals the ratio between the return percentage of existing candidate and target percentage. And the scaling factor plus the slack variable equals 1. Therefore, by using experimental Raman spectra alone, we will be able to know the correct composition of the target

spectrum. Moreover, knowing the scaling factor.
Although the accuracy for E2 is high, it is not the case for E3 which LP model only contains SFG experimental spectrum. As we can see that the percent that it hits our previous observation is low, even lower than the LP model only contains IR spectra. From Chapter 5, we know that for candidates with θ from 0° to 80°, Raman or SFG alone are both sufficient to obtain the composition of target spectrum. However, when we introduce the scaling factor, the result for SFG is not sufficient any more. (Not sure how to explain this one theoretically)

Even for E5, which combines IR and SFG spectra data, the result is not much better than E1 or E3. However, any experiment that contains Raman spectra data,result in good accuracy.

For E4, E6 and E7, the scaling factor for each of these experiments is the same as the experiment only contains Raman spectra data. As well as the slack variable, its value for these experiments is the same as the one from LP model with only Raman spectra data. (Looks like Raman is dominating the result here.)
What happens when scaling factor is greater than 1?
When the scaling factor is greater than 1, all LP models built by the 7 experiments will fail to obtain the correct composition of target spectrum.
All the experiments will fail.(Need to discuss how we can resolve this problem)

Nevertheless, when we expand the candidates from 0° to 180° on θ, with the scaling factor still smaller than 1. The result we observe is interesting as well.

The LP model with only Raman spectra information E2, is able to tell us for each amino acid, candidate with which θ or this θ's complementary would exist in our target spectra. Because Raman spectra for candidate with θ on one degree is the same as its supplementary. This LP model can not distinguish between candidate with θ and the one with its complementary.However, with the help of SFG, we may be able to know which one between the above two dominates one amino acid's the total fraction, like what we have learnt from Chapter 5 about E6. Therefore we exam E6 here, and it displays which one of the two takes the major composition in the target spectra. With this information, we can decide between the θ and its complementary. With the information coming from E2 and E6, we therefore, obtain the right composition

of the target spectrum.

Here goes the example, in one run of the experiment group. The composition for the target spectrum is Array 6.4. In this target spectrum, we have 0.14799 of $\theta = 40°$ Methionine, 0.74202 of $\theta = 50°$ Leucine, 0.08989 of $\theta = 150°$ Ile, 0.01135 of $\theta = 40°$ Ala, 0.00715 of $\theta = 0°$ Thr, 0.0016 of $\theta = 20°$ Val.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0.14799 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.74202 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.08989 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.01135 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.00715 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0016 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (6.4)$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0.102936 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.516118 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0625238 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0078945 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.00497324 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.00111289 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.304441 & & & & & & & & & & & & & & & & \end{bmatrix} \quad (6.5)$$

The result returned by E2 is shown in Array 6.5. You may notice same as Array 6.2 to Array 6.1, Array 6.5 contains one more value than Array 6.4, 0.304441, which is the slack variable. We already know that the scaling factor is 0.695560510845(generated randomly, but recorded). When we add the slack variable and the scaling factor, the total comes up to 1.0.

In Array 6.5, we get 0.102936 of $\theta = 40°$ Methionine, 0.516118 of $\theta = 50°$ Leucine, 0.0625238 of $\theta = 30°$ Ile, 0.0078945 of $\theta = 40°$ Ala, 0.00497324 of $\theta = 0°$ Thr, 0.00111289 of $\theta = 20°$ Val. From Array 6.6, we can also deduce the value for the scaling factor.

$$\frac{0.102936}{0.14799} = \frac{0.516118}{0.74202} = \frac{0.0625238}{0.08989} = \frac{0.0078945}{0.01135}$$
$$= \frac{0.00497324}{0.00715} = \frac{0.00111289}{0.0016} = 0.695560 \tag{6.6}$$

At first glance, we may guess that this LP model actually return the correct composition. However, not all amino acids' composition is correct. For Ile, it should be 0.0625238 of $\theta = 150°$, but the result returned 0.0625238 of $\theta = 30°$, which is the complimentary of 150°. It is because these two degrees' Raman spectra are identical, there is no way for current LP model to distinguish these two.

With this information, we know the only thing we need to make sure is: is the θ returned by LP model the exact one in target spectrum or its complementary? (The above conclusion can also be applied to the experiments in Chapter 5 without the scaling factor. The LP model with only Raman spectra information can help us to see which θ of candidate and its complimentary should be for each amino acid. I did not observe this before.)

To answer this question, we need the help of SFG data. Because only SFG can tell us the difference between one angle and its complementary, as their spectra are symmetry, not identical around the axis of wevenumber.

In this experiment group, the result returned E6 is Array 6.7 (Second example???)

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0.0776716 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0252641 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.3894440 & 0 & 0 & 0 & 0 & 0 & 0.1266740 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.0153456 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0471782 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.00595697 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00193762 & 0 & 0 & 0 & 0 \\
0.0037526 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00122061 \\
0 & 0 & 0.000839749 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.000273144 & 0 & 0 \\
0.304441 (any relation for the fraction???) & & & & & & & & & & & & & & & &
\end{bmatrix} \tag{6.7}$$

The value of the slack variable is the same as what returned by E2. However, the returned composition is totally different than what returned by E2. The interesting thing is that for each amino acid, the return existing candidates are complementary on θ. The total percentage of these two candidates, take Methionine as an instance, $0.0776716 + 0.0252641$ makes 0.1029357 which is the same as what is returned by E2. This is the same for every amino acid. What's more, the composition returned by the

E6 indicates which θ dominates the composition for one amino acid. For Methionine, $\theta = 40°$ takes major part; for Leucine, $\theta = 50°$ does; for Ile, $\theta = 30°$ does; for Ala, $\theta = 40°$ does; for Thr, $\theta = 0°$ does; for Val, $\theta = 20°$ does; And those candidates are the correct components for target spectra.

IR+SFG, can you do anything with it???

6.2 Results

6.3 Discussion

6.4 Conclusions

Chapter 7

Conclusion and Future Work

7.1 Conclusion

7.1.1 Contributions

7.2 Future Work

Appendix A

Additional Information

This is a good place to put tables, lots of results, perhaps all the data compiled in the experiments. By avoiding putting all the results inside the chapters themselves, the whole thing may become much more readable and the various tables can be linked to appropriately.

The main purpose of an Appendix however should be to take care of the future readers and researchers. This implies listing all the housekeeping facts needed to continue the research. For example: where is the raw data stored? where is the software used? which version of which operating system or library or experimental equipment was used and where can it be accessed again?

Ask yourself: if you were given this thesis to read with the goal that you will be expanding the research presented here, what would you like to have as housekeeping information and what do you need? Be kind to the future graduate students and to your supervisor who will be the one stuck in the middle trying to find where all the stuff was left!

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

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

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