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

A surface is an interface of two phases. The phases can be in any forms (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, and so on. Therefore, surface characterization identification remains an active area of research in the physics, chemistry, biotechnology communities and modern electronic technology, and 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 the following aspects: adhesion, lubrication, catalysis, bio-membrane functions and so on. [**?**]

Many experimental techniques have been applied in the study of molecular orientation at interfaces. Moreover, among them the optical methods are more preferable, for example, infrared absoption (IR), Raman scattering (Raman) and sum-frequency generation (SFG). All these vibrational spectra carry quantitative structural information of molecules at interfaces. Although each of them may have its own shortcomings, they all share the following advantages comparing with the other non-optical methods. First of all, they 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[**?**][**?**]. Moreover, the other 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 analysis are required. Combing different spectroscopy techniques is an very effective way to achieve the goal of molecular study. However, finding the most effective ways to combine them 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, projecting the vibrational mode properties from molecular frame to laboratory frame, and so on. In addition to the study, these three spectroscopy techniques are employed with a special mathematical programming: Linear Programming (LP). In order to explore how LP can further facilitate in extracting the quantitative structural information of molecules at interfaces.

The approach is to start with small molecule model, to study the properties of the LP models built by using the spectroscopy information. Then real molecules will be introduced, which are six well-known amino acids: methionine, leucine, isoleucine, alanine, threonine and valine.

Before the details of the LP models and the molecules are introduced. The basic theory of the IR, Raman and SFG spectra, and Linear Programming will be introduced.

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, which are the derivatives of these two factors. The dipole moment and polarizability are changing as the molecule's conformation is changing. Current study will focus on amino acid molecules, which have less complex conformation. It helps to reduce the possible number of the dipole moment and polarizability of a molecule.

For IR, its absorption is the absorption-transmission-reflexion mode (resonant).

The physical principle is the variation of the static dipole moment μ (the first rank tensor) along the normal coordinates Q: $\frac{\partial \mu}{\partial Q}$.

$$\langle 1|\overline{\mu}|0\rangle \approx \frac{1}{\sqrt{2m_Q w_Q}} \frac{\partial \mu}{\partial Q} \tag{1.1}$$

where $|0\rangle$ is the vibrational ground state, $\langle 1|$ is the vibrational excited state, m is the reduced mass of the normal mode, $\overline{\mu}$ is the transient dipole moment, and w is the resonance frequency. The dipole moment μ is a vector with x, y and z. Therefore, the dipole moment derivatives can be expressed as Equation 1.2. The IR spectra can be obtained from 3 projections: x, y, z.

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

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

$$\langle 1|\overline{\mu}|0\rangle \approx \frac{1}{\sqrt{2m_Q w_Q}} \frac{\partial \alpha^{(1)}}{\partial Q} \tag{1.3}$$

where $|0\rangle$ is the vibrational ground state, $\langle 1|$ is the vibrational excited state, m is the reduced mass of the normal mode, and w is the resonance frequency. The polarizability is coupled by (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 projections of Raman spectra: $xx, yy, zz, xy, xz, yx, zx, yz$ and zx.

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

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 dot product of dipole moment and polarizability, $\chi^{(2)}$ (the third rank tensor): $\frac{\partial \mu}{\partial Q} * \frac{\partial \alpha}{\partial Q}$. Therefore, there are 27 elements for SFG spectra, which result in 27 projections of SFG spectra.

$$\langle 1|\overline{\alpha}|0\rangle\langle 1|\overline{\mu}|0\rangle \approx \frac{1}{2m_Q w_Q}\left(\frac{\partial \alpha^{(1)}}{\partial Q} \otimes \frac{\partial \mu}{\partial Q}\right) \tag{1.5}$$

1.3 Linear programming

Linear Programming (LP) is an optimization problem. The standard form of LP is a minimization problem that has an objective function and constraints as shown in Equation 1.6 [2]:

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

The goal is to minimize the objective function, with $x_{1,...,n}$ as the decision variables. The constraints are the conditions that the decision variables need to subject to.

The diet problem is a popular example to illustrate the concept of LP. For example, a restaurant would like to achieve the minimal nutrition requirement with the lowest price of the food selection as shown in Picture 1.3. For each meal, the minimal requirements for vitamin A, vitamin C and dietary fiber are $0.5mg$, $15mg$ and $4g$. The restaurant has three options: raw carrot, raw white cabbage and pickled cucumber. The table also displays the nutrition contains and the price of each ingredient. With all this information, we need to know how much carrot, cabbage and cucumber to put into each meal, so that the minimal nutrition requirements can be met with the lowest price. In result, the goal is to minimize the price, and the constraints are the

nutrition requirements. Therefore, we come up with a model as shown in Equations from Equation 1.7 to 1.13.

Food	Carrot, Raw	White Cabbage, Raw	Cucumber, Pickled	Required per dish
Vitamin A [mg/kg]	35	0.5	0.5	0.5 mg
Vitamin C [mg/kg]	60	300	10	15 mg
Dietary Fiber [g/kg]	30	20	10	4 g
price [€/kg]	0.75	0.5	0.15*	—

Figure 1.1: Diet Problem

$$\text{minimize} \quad 0.75x_1 + 0.5x_2 + 0.15x_3 \tag{1.7}$$

$$\text{subject to} \quad 35x_1 + 0.5x_2 + 0.5x_3 \geq 0.5 \tag{1.8}$$

$$60x_1 + 300x_2 + 10x_3 \geq 15 \tag{1.9}$$

$$30x_1 + 20x_2 + 10x_3 \geq 4 \tag{1.10}$$

$$x_1 \geq 0 \tag{1.11}$$

$$x_2 \geq 0 \tag{1.12}$$

$$x_3 \geq 0 \tag{1.13}$$

In this LP model, x_1, x_2 and x_3 are decision variables, each presents the amount of the ingredients. Equation 1.2 is the objective function to minimize. Equation 1.8 to Equation 1.10 are the nutrition requirements. Equation 1.11 to Equation 1.13 restrict 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 are only three kinds solutions: feasible and bounded, feasible and unbounded, and infeasible. 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 infinite optimal solutions[1].

For a general LP problem, it can be a minimization or a maximization problem. Its constraints can be equalities or inequalities. However, there are ways to convert the LP problem into its standard form. Furthermore, for a LP problem that con-

tains n decision variables, its solution would be in n-dimension 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 is(are) feasible solution(s). This makes LP a convex problem. The benefit of a convex problem is that the local optimal solution is the global optimum. The optimal solution that the LP solvers return is the global one. If a LP problem has a unique optimal solution, then this solution is definitely a vertex of the convex polyhedron.

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

Another advantage of LP is that it can deal with thousands of variables, which makes it suitable for the study on 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.

Among these algorithms, the most common one is Simplex method. Most popular LP solvers have implemented Simplex and Interior point algorithms. Comparing to Interior point, Simplex method is comparatively easier to understand and implement. 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. As the previous research conducted by Kuo Kai Hung (Kai) was mainly using Simplex method, Simplex method will be used for the further experiments as well.

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 that it starts at a vertex, then pivot from vertex to vertex, until it reaches the optimum. Although it has proved that Simplex method is not a polynomial algorithm, however, 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 the study. The approach is to build LP model, and check if the optimal solutions returned by these models match to the target composition that used to generate the target spectra or not. The LP models are built using spectra of different techniques. Therefore, there are different LP models. From these models, then analyze which model helps to reachthe goal with the highest accuracy. With this result,knowing which spectroscopy technique(s) is/are more 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 get the accurate composition of the target spectra.

Chapter 2

Methods

2.1 Current approaches to structure elucidation

Previous work.

2.2 Structure of molecules adsorbed to surfaces

(TODO: check with Dennis, how to expand this part.)

2.3 Generating model spectra

Creating candidate spectra is an essential step before building LP models. This part of research has been done thoroughly by Kai [**?**]. In this chapter, the content in his thesis that is related to the current study will be explained, as well as what this thesis has extended an.d is different from Kai's thesis.

As mentioned in Chapter **??**, before analyzing the vibrational spectra of amino acids, there are a few factors to address first.

First of all, in order to generate these amino acids' vibrational spectra, a molecule's vibration modes needed to be modelled in the molecular frame, then transferred to the laboratory frame to work with the systems where interfaces exist. Chapter 2 in Kai [**?**] describes how to use GAMESS /cite??? to obtain every atom's dipole moment derivative and polarizability derivative of one molecule. The paper then introduces

how to use Direction Cosine Matrix (DCM) to transfer these two derivatives from the molecular coordinate system to laboratory one. After that, Euler angles are extracted from DCM. Euler angles are labelled by θ, ϕ and ψ as shown in Figure 2.1, and referred as *tilt*, *azimuthal* and *twist* angles respectively. *Tilt* angle θ is the one between z and c; *azimuthal* angle ϕ is the rotation about z; and *twist* angle ψ is a twist about c /citehore0033-rotations.pdf. Euler angles are used to describe a molecule's coordination at interfaces. After three steps of successive rotations of Euler angles, molecule properties can be transferred from the molecular frame to the lab frame.

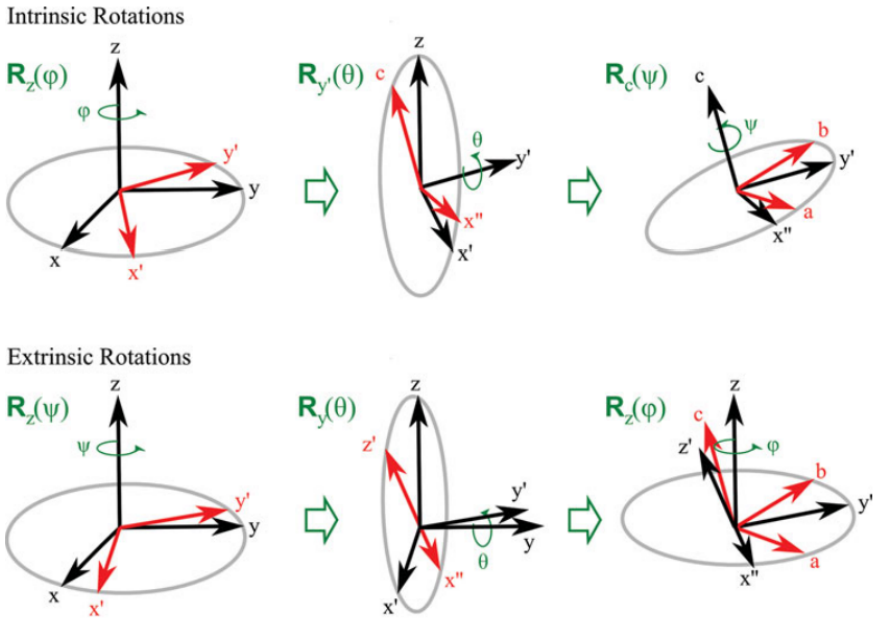

Figure 2.1: The Euler angles represented as the spherical polar angles θ, ϕ and ψ, and the illustration of the three successive rotations that transform the lab (x, y, z) coordinate system into the molecular (a, b, c) frame.

When the molecules lay on an interface, the orientation for each molecule varies. This requires either do a simulation to study the distribution of molecule orientations at the interface, or come up with a analytic orientation distribution function. In this study, the second method is preferred, i.e., using Delta distribution function to represent the molecule orientation distribution that models the spectrum signals.

Furthermore, the experiments are limited to only consider *tilt* angle distribution

of Euler angles, and assume isotropy in *twist* and *azimuthal* angular distributions. Therefore, these two Euler angles are integrated to create a uniform distribution for them. For angle ψ, it requires the surfaces to be not striped. However, it can be no anisotropy in the plane of the surface. Because of this, assumption is simplified by integrating ψ angle. 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 titled, 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 is still enormous when only considering the θ angle. The possible combinations of all these amino acids candidates that will be consider are still excessive.

The above is the reason to seek a mathematical model with appropriate computational speed to help the study of the composition of molecule(s) at interfaces with given experimental target spectra and candidates spectra.

In first, GAMESS was used to do an Hessian calculation, next 7 snapshots of a molecule vibrating in different modes were taken, then the derivatives of dipole moment and polarizability based on these 7 modes' interpolation were obtained. Because the two obtained directives are in GAMESS's coordinate system (molecular frame), he then used DCM to convert them into lab frame, and abstracted Euler angles from DCM. After this transformation, he restored the derivatives information in Python pickle files for any further usage.

For my study, I will directly use the pickle files to generate the following amino acids' spectroscopy information: methionine, leucine, isoleucine, alanine, threonine and valine. Within each pickle file, it contains derivatives of dipole moment and polarizabilities of vibrational mode from 0 to 36. Furthermore, We use the following formulations to generate each amino acid's IR, Raman and SFG spectrum.

For infrared absorption spectroscopy (IR), it is a harmonic approximation, the intensity of the x-polarized absorption spectrum is given by

$$I_x(w_{IR}) = \sum_q \frac{1}{2fm_q w_q} \left\langle \left[\frac{\partial u_x}{\partial Q} \right]_q^2 \right\rangle \frac{\Gamma_q^2}{(w_{IR} - w_q)^2 + \Gamma_q^2} \tag{2.1}$$

where I_x represents x-polarized intensity. The same equation applies to I_y and I_z. w_{IR} is the frequency of the probe radiation. μ is the dipole moment. m_q is the reduced mass. w_q is resonance frequency. f is ???, which is 0.936 for all the experiments we will conduct in the following chapters. Γ_q is the homogeneous line width, which is 6 for all the following experiments as well. Q_q is the normal mode coordinate of the qth vibrational mode. And all the value of w_{IR}, μ, m_q, Q are obtained from the pickle files. Furthermore, because we integrate ϕ and ψ angles, the y-polarized spectrum is identical with the z-polarized one. Therefore, in our future experiments, we can only obtain two different polarized spectrum for one molecule. (TODO: need to double check the accuracy with Dennis)

For Raman spectroscopy, a spectrum with an x-polarized excitation source, will collect the x-polarized component of the scattered radiation, and can be approximated from

$$I_{xx}(\Delta w) = \sum_q \frac{1}{2fm_q w_q} \left\langle \left[\frac{\partial \alpha_{xx}^{(1)}}{\partial Q} \right]_q^2 \right\rangle \frac{\Gamma_q^2}{(\Delta w - w_q)^2 + \Gamma_q^2} \tag{2.2}$$

where Δw is the Stokes Raman shift. $\alpha_{xx}^{(1)}$ is one component of the 9-element polarizability tensor. m_q is the reduced mass. w_q is resonance frequency. f is (TODO: confirm the meaning with Dennis) the same as what's for IR spectra, which is 0.936 for all the future experiments. Γ_q is homogeneous line width, which is 6 for all the future experiments. Q_q is the normal mode coordinate of the qth vibrational mode. All the value of w_{IR}, μ, m_q, Q are obtained from the pickle files. Similar to IR spectroscopy, because of the integration on ϕ and ψ angles, we obtain only 4 unique spectrum from the following polarization: xx, xy, xz and zz. (TODO: double check the accuracy of the content with Dennis).

For SFG spectroscopy, the response intensity is governed by

$$I^{(2)}_{xxx}(w_{IR}) = \sum_q \frac{1}{2f m_q w_q} \left\langle \left[\frac{\partial \alpha^{(1)}_{xx}}{\partial Q}\right]_q \left[\frac{\partial u_x}{\partial Q}\right]_q \right\rangle \frac{1}{w_q - w_{IR} - i\Gamma_q} \tag{2.3}$$

where I^2_{xxx} is the second-order susceptibility tensor. It is probed by a x-polarized visible incoming beam at frequency w_{vis} and a x-polarized infrared beam incoming with frequency w_{IR} are incident to the sample, and then the x-component of SFG at frequency $w_{SFG} = w_{vis} + w_{IR}$ is selected for detection. As $i = \sqrt{-1}$ is in the denominator, $X^{(2)}$ is a complex value/citeKai thesis. And the SFG response is the imaginary component of the second-order susceptibility. Same as IR and Raman spectroscopy, all the value of w_{IR}, μ, m_q, Q are obtained from the pickle files. Because of the integration on ϕ and ψ angles, we only get 3 unique spectrum from the following polarization: yyz, yzy and zzz. (TODO: double check the accuracy of the content with Dennis).

With these equations and the pickle files, we can then generate IR, Raman and SFG spectroscopy spectrum for a specific molecule with certain Euler angles, and here we only consider θ angle. Therefore, we only need to specify the amino acid and its θ angle value, and we will have the spectral information for a targeted candidate. A candidate here is a specific amino acid (or molecule) with specific Euler angles (orientation at interfaces). (TODO: check accuracy with Dennis).

In Picture 2.2, we have generated x-polarized IR spectrum for Methionine with θ of the following values: 0°, 20°, 40° and 60°. Their spectrum are prefixed with *candidate_* in the label. *ir_x_* indicates the spectroscopy technique, "number" indicates the θ angle's value. For the spectra labelled as *target_ir_x*, is generated by combining 10% of *candidate_ir_x_*0, 50% *candidate_ir_x_*20 and 40% of *candidate_ir_x_*40. (Putting the target spectra here is for comparing and visual)

Similarly, Picture 2.3, 2.4 and 2.5 are the spectrum for the same candidates and target for z-polarized IR, xx-projection Raman and yyz-projection SFG spectrum respectively. Taking Picture 2.2 as an example, the difference among the four candidates does not appear significantly. The biggest differences occur at each vibrational mode. The valid wavenumber is from 1000 to 2000. Therefore for each projection,

no matter IR, Raman or SFG, we can extra 200 data points by every 5 wavenumber (This has also proved to be most effective experimentally as well). With these data points, we will construct the corresponding LP model as we will describe in **??**.

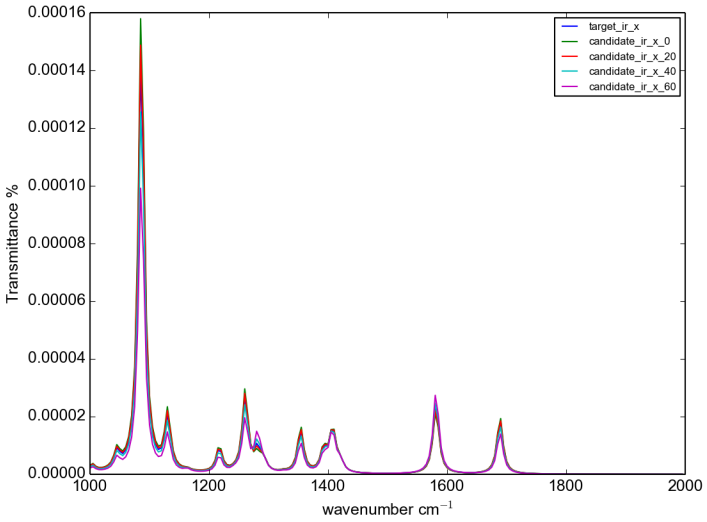

Figure 2.2: IR x Projection Spectra for Methionine Four Candidates and Taget

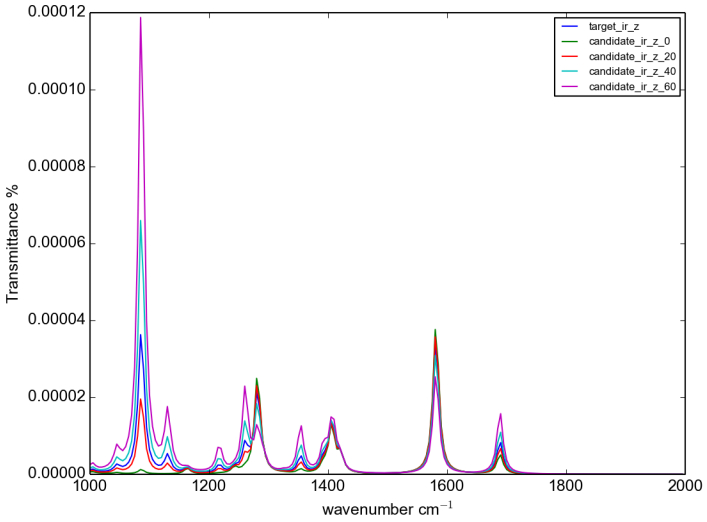

Figure 2.3: IR z Projection Spectra for Methionine Four Candidates and Taget

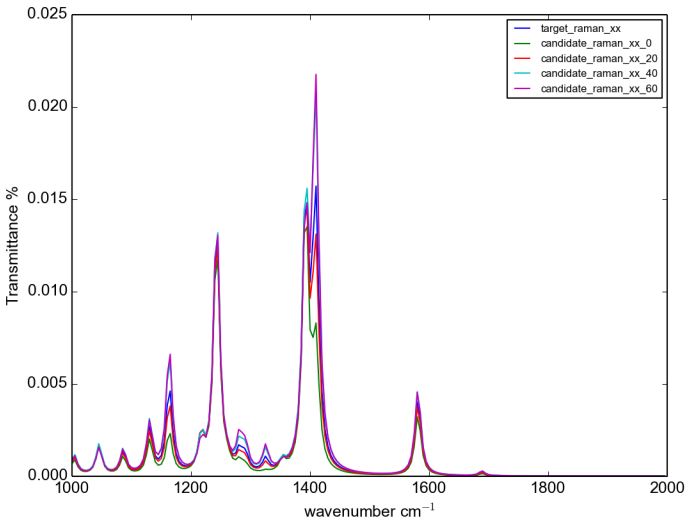

Figure 2.4: Raman xx Projection Spectra for Methionine Four Candidates and Taget

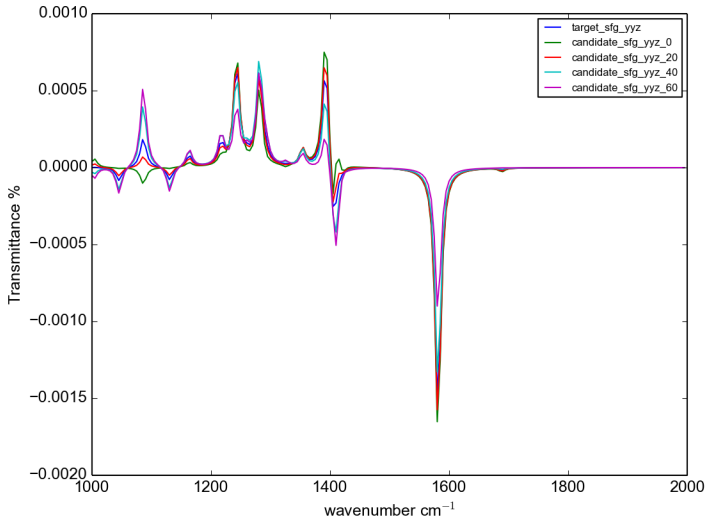

Figure 2.5: SFG yyz Projection Spectra for Methionine Four Candidates and Taget

2.4 Surface orientation distribution functions

(TODO: check with Dennis, how to expand this part.)

2.5 The properties of the LP models

In the following chapters, we will study the properties of the LP models that we are going to build. What's important, we want to study the maximum capacity of our LP models, and figure out what are the reasons that cause the limitation of our LP model. Meanwhile, we also want to compare the sensitivity of different spectroscopy techniques in the term of studying molecular orientation distribution at interfaces.

In order to achieve the above goals, we first study the properties of our LP models. In **??**, we will expand this part of study by using a toy model to gain an insight of what is the barrier of our LP models. With the further information we will gain in **??**, we will build more experience accordingly.

Chapter 3

Simplified Molecular Model

3.1 Description

The goal for Chapter **??** is to further expand our research focus and introduce the formulas we use to generate the LP models. Therefore, before jugmpin to the real molecules, we first study a toy molecule with limited vibration modes. By doing this, we would like to analyse the nature of the LP formulas we use in order to gain some further insights of the models we build. With the data information we have, could the LP models give us some valuable information? If yes, what are those applicable cases. If not, what could be the reason that causing the LP model to fail?

In order to answer the above question, we construct a toy molecule. For this toy molecule, there are 4 vibration modes, with vibrational peaks at frequency of $2850, 2960, 3050$ and 3200. Furthermore, the width for each peak is $5, 10, 5$ and 15 respectively. and the amplitude for each peak is $1, 0.7, -0.2$ and 0.5 respectively. At last, the comparing angle for each peak is $15, 90, 0$ and 60. (TODO: check with Dennis, how to explain those comparing angles?)

For this toy model, we will only consider IR spectroscopy. We use equation **??** to generate the cosine projection IR spectroscopy. Moreover, we integrate ϕ and ψ Euler angles for the toy model as well. Therefore, we only need to consider θ angle.

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

where A is the amplitude, $theta_q$ is the $tilt$ angle of the candidate, γ is the width, w_q is the frequency. (TODO: Double check the correct meaning of each symbol) Then 10 candidates' spectra with 10 different θ values: $\theta = 0°, 10°$, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 90° are generated in Picture **??**. From the picture, the 10 candidates have peaks at the same frequencies. However, one may wonder that the spectrum among all the candidates here may look quite similar to each other. This does not matter much in appearance, however, may make it difficult to obtain the exact composition for the targeted spectrum, as various combinations of candidates are possible to achieve the targeted spectrum. We will need to run certain experiments to further demonstrate this point.

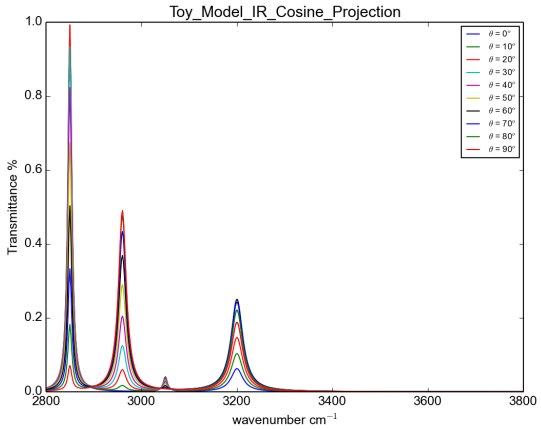

Figure 3.1: Toy Model IR Candidates Cosine Projection

As an example, we compose a target spectrum by combining 15 percent θ of 20 degree's candidate and 85 percent of θ of 70 degree's candidate: $0.15 * f_{20}(x) + 0.85 * f_{70}(x)$ for the following experiment.

3.2 Linear Programming Model for Spectra Study

The linear programming model that we construct to check if the optimal solution returned by LP solver actually matches the known composition is shown as equation **??**. Th model has also been used in the following studies: [**?**] and [**?**].

$$minimize \sum_{p=1}^{points} \left| Target - \sum_{c=1}^{candidates} p_c f_\theta(x) \right|$$
$$p_c : percentage\ of\ each\ candidate$$
$$p : number\ of\ points\ selected\ both\ from\ candidates\ and\ Target \tag{3.2}$$

In this model, the unknown percentage of each candidate is the decision variable, it is denoted by p_c. We select data points along the wavelength frequency, the data point is denoted by p here. Based on each data point, we calculate the absolute residual between target spectrum and the one composed by the decision variables. Then the objective function is to minimize of the sum of the absolute residual over all the data points.

However, in order to actually apply Linear Programming, we need to get rid of the absolute sign in the objective function. We achieve this goal by introducing one more variable X and two more constraints for each data point. Accordingly, variable X and the constraints are shown in 3.3. We then convert the previous model in equation **??** into one that can actually be solved by Linear Programming solver 3.4. And at last, we introduced the last constraint to restrict the sum of the percentage to 1, as shown in the last equality in 3.4.

For each point:

$$X = \left| Target - \sum_{c=1}^{candidates} p_c f_\theta(x) \right|$$

$$X \geq Target - \sum_{c=1}^{candidates} p_c f_\theta(x)$$

$$X \geq -Target + \sum_{c=1}^{candidates} p_c f_\theta(x) \tag{3.3}$$

$$minimize \sum_{p=1}^{points} X_p$$

$$X_1 - Target_1 + \sum_{c=1}^{candidates} p_c f_\theta(x_1) \geq 0$$

$$X_1 + Target_1 - \sum_{c=1}^{candidates} p_c f_\theta(x_1) \geq 0$$

$$\cdots$$

$$X_n - Target_n + \sum_{c=1}^{candidates} p_c f_\theta(x_n) \geq 0$$

$$X_n + Target_n - \sum_{c=1}^{candidates} p_c f_\theta(x_n) \geq 0$$

$$\sum_{c=1}^{candidates} p_c = 1 \tag{3.4}$$

3.3 Linear programming model implementation

In order to run the experiments, there are few steps we need to implement first. And all the code is implemented in Python. To start, we need to generate one file that contains all the candidates' spectral information we want to focus for one experiment. For this step, we make use of the pickle files. For a specific candidate, we give the amino acid pickle file and the θ degree to obtain the candidate's spectral information.

In first step, I create a class that called Candidate in *candidate_class.py*. This class defines one candidate's x- and z- polarized IR spectrum; xx-, xy-, xz-, and

zz- projection Raman spectrum; *yyz*-, *yzy*-, *zzz*- projection SFG spectrum. Given one candidate's pickle file and θ value, *candidate_class.py* will create a instance of this specfic candidate. For the toy model, it only contains IR spectral information, therefore, one candidate only contains *cosine*- and *sine*- polarized IR spectrum.

In the second step, when giving a list of needed candidates, the composition to generate target spectra, and the probe arrange (which is the range of the wavenumber, for the toy model, the probe arrange is from 2800 to 3300 wavenumber. For the further experiments in Chapter 4, **??** and **??**, the probe arrange is from 2000 to 3000 wavenumber), file *create_candidates.py* will generate all the candidates and target spectra information into one file called *candidate.txt*. If you want to run experiments that contain different spectroscopy information, you specify in *create_candidates.py* as well, and different candidate.txt will be created, which contains 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, we need to construct the LP model by using *candidate.txt*. For this step, the script *realProblemSolver.py* is written by Kai. In this script, it reads all the candidates and the target spectral information, then builds the LP model as shown in 3.4, then creates CPLEX LP input file called *lp_input.txt*.

In the fourth step, we run the following command in terminal: "$glpsol - o < outputfile > - - lp < CPLEXLPinputfile >$" to obtain the output of the LP solver.

3.4 Experiments

In order to further simplify the problem, in the first experiment set, it is necessary to limit the candidate numbers to 4. Table **??** illustrates the detailed setting for experiment 1 and 2. In the first experiment, there are 4 candidates with θ of $0°, 10°, 20°$, and $30°$. For the second experiment, the θ values have been changed to $0°, 5°, 10°$, and $15°$. Instead of having 10 degree variance in the theta, we have 5 degree difference on theta. This means that the candidates are becoming a bit more similar to each other, their spectra are getting more similar as well. The increase in similarity may

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

make the problem more difficult to resolve. For both experiments, 100 data points are selected evenly along the wavenumber from the spectra of *cosine*-polarized IR. The target composition for the candidates are the same for both experiments. For 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.

Why the composition returned from LP solver for experiment 1 matches to the target, but not the same case for experiment 2? In order to answer this question, for experiment 2, the spectra that generated by the return composition and the target one is plotted together in Picture **??**. Interesting thing is that these two spectrum are almost identical to each other. In order to see that if it is the general case, another experiment is set up in Table **??**. This experiment is a bit more complex than Experiment 1 and 2. 10 candidates are introduced, including the *theta* value range from 0° to 90°.

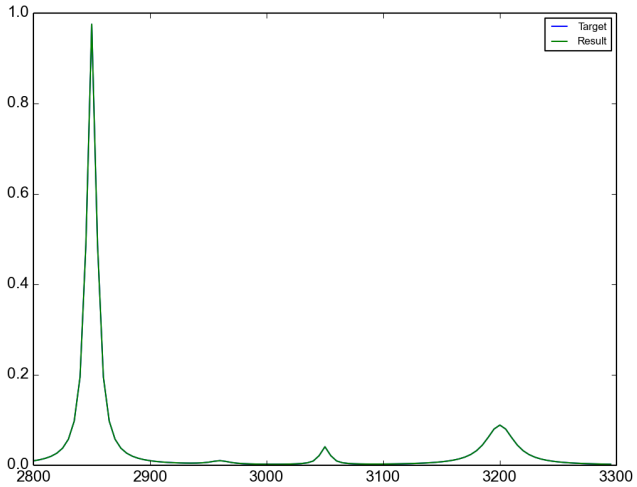

Figure 3.2: Toy Model Result Plotting for 4 Candidates on IR Cosine Projection

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

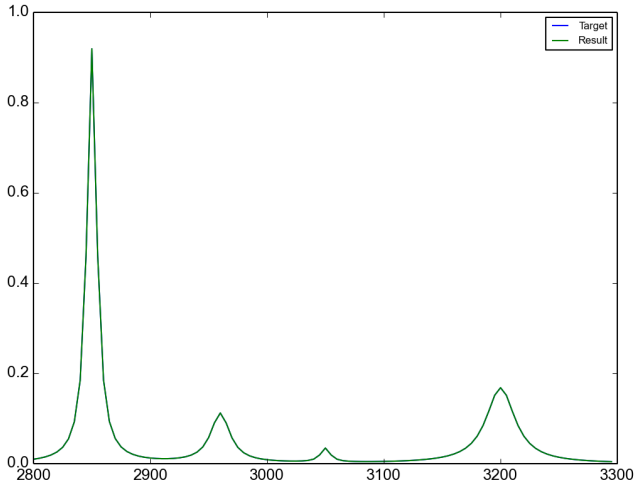

Figure 3.3: Toy Model Result Plotting for 10 Candidates on IR Cosine Projection

Table **??** indicates that: for Experiment 3, the return composition is different from the target one. Picture **??** shows that the spectra produced by the return composition is almost identical as the one generated by the target composition.This is the same case as Experiment 2. In addition, for Experiment 2 and 3, the sum of residual, between the spectra that composed by return composition and the one composed by target composition,is very small that can almost be negligible.

Among Experiment 1, 2 and 3, only Experiment 1's return composition matches to its target one. For Experiment 2 and 3, the return composition is totally different from the target one. However, the complexity of Experiment 2 and 3 is higher than Experiment 1. 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 difficulty for LP model to return the target composition. Meanwhile, from Picture **??** and **??**, for both Experiment 2 and 3, the spectra constructed by return composition matches to the one built by target composition. This illustrates that the corresponding LP model, built with the presenting spectral information, may lack of competent information to guarantee the return composition actually is the target one. Because of the high complexity in the setting of the candidates pool, the solution for the LP model may not be unique. However, in real LP problem solving, the numerical limitation helps LP solver to converge to an unique solution. This unique solution is the optimum one for the LP model with

the information currently available. Further demonstration will also be expanded in Chapter 4 when real molecules are introduced.

In order to bring in more information to the LP model, The second projection of IR is introduced: the sine projection. Picture **??** describes how the spectra is presented for 10 candidates same as Experiment 3. Experiment 4 and 5 will include both projections' spectral information when building the LP model. Table **??** shows the setting for Experiment 4 and 5. The setting is based on Experiment 2, in addition includes sine-projection IR spectra. 100 data points are selected from this additional spectra, then converted to decision variable and constraints in the LP model. Same with Experiment 5, it is based on Experiment 3, with sine-projection 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.

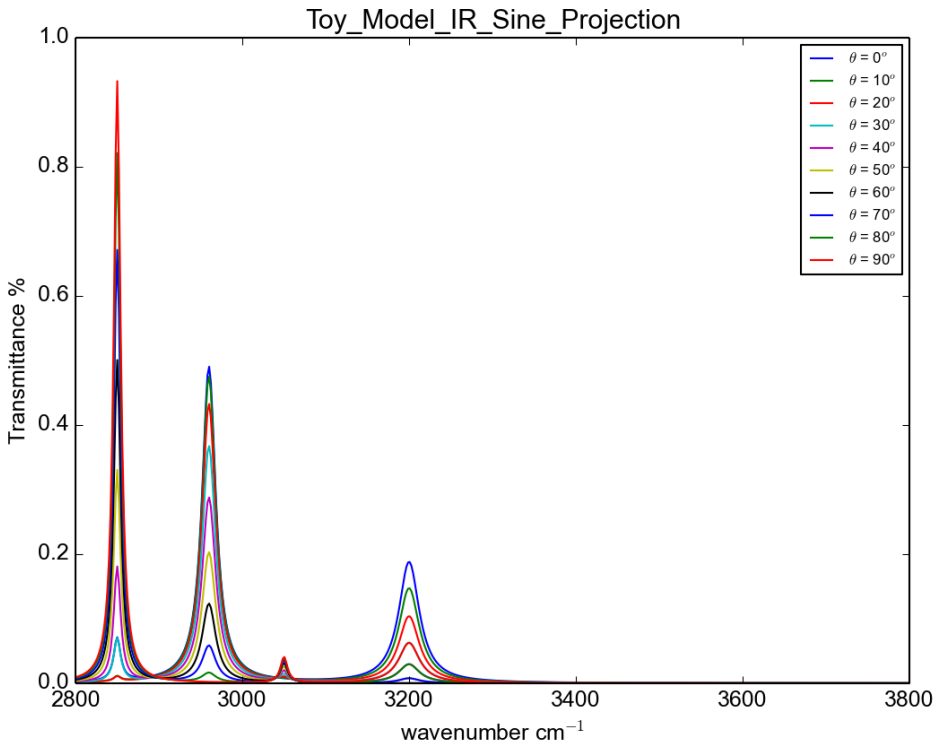

Figure 3.4: Toy Model Candidates IR Sine Projection

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

3.5 Constraint Study Based on Experiment 4

From Experiment 1 to 5, we know having sufficing information in our LP model is the key to obtain the target composition. Having sufficing 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.

Base on Experiment 4, experiments about constructing LP model with different data information are conducted in Table 3.5. 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 3.5 is that: from Experiment 1 to 9, the result composition is the same. Au 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 3.5. In this picture, we can see that the spectra generated by these two composition are identical.

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

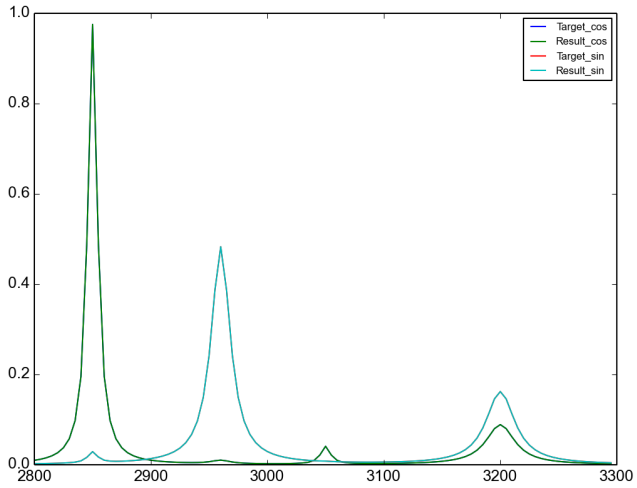

Figure 3.5: Toy Model Constraint Study 1

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

3.6 Constraint Study Based on Experiment 5

And when the same constraint study is applied to the experiments based on Experiment 5 in Table 3.6, the observation is the same as the experiments in Table 3.5. 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 Picture 3.6.

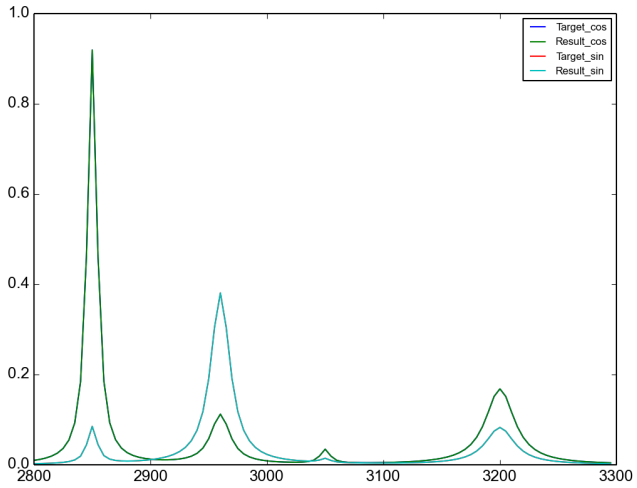

Figure 3.6: Toy Model Constraint Study 2

3.7 Discussion and Conclusion

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

The above analysis throws a question: how can we know there is enough information to achieve the target composition? In the next step, we will experiment with the real molecules. The goal is to check with all the spectral information 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 the toy problem, lacking sufficient information for the LP model is the key cause for the failure of obtaining the correct composition for the target spectra. However, in the toy model, there are only four vibrational modes, and the range of wave-length is limited. Only few data points can be selected to conduct the constraints for the LP model. Secondly, the similarity among the candidates is high, as all the candidates are coming from one same molecule, and the only difference is in θ degree. Third, we only obtain the data points from IR spectra. What if we introduce more spectroscopy techniques, such as: Raman and SFG that allow us to have more information to build the constraints?

In this chapter, real molecules are introduced. They contain more abundant spectroscopy information, meanwhile, making the study one step closer to the overall goal and scope.

The plan here is to apply similar experiments that have applied to the toy model to a real molecule. The real molecule is an amino acid - Methionine.

In order to limit the possible candidate space of Methionine, *twist* and *azimuthal* angular distributions are assumed to be isotropic, therefore are integrated. Only θ in Euler angles is considered in Methionine's surface orientation distribution functions. In Chapter **??**, how the spectra for IR, Raman and SFG have been generated for the

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

toy model has been explained. However, for an amino acide molecule, this process is a bit more complicated comparing to the toy model. The process has been explained in Chapter **??**.

Chapter 2 has introduced the following concepts: for IR, two unique spectra can be obtained from x, and z projections. For Raman, four unique spectra can be obtained from xx, xy, xz and zz projections. For SFG, three unique spectra can be obtained from yyz, yzy and zzz projections.

In current chapter, all the three spectroscopy techniques will be studied in the experiments. In order 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, is spectroscopy information from all the three techniques needed, or only some of them are needed to achieve the goal. If no, is the same reason as the toy model that causes the failure of the LP model.

4.2 Experiments

In Table 4.2, there are four experiments have been set up. Each experiment is set up with the above four candidates and the same target composition. They are all Methionine candidates, with differences in *tilt* angle, which is the θ degree. These four candidates each has θ with the following degree: 0°, 20°,40° and 60°. The difference in θ degree is 20°. The only difference among these four experiments is the spectroscopy information we select to construct the LP model that indicated in Number of Data Points. In Experiment 1, only IR x-projection spectra information is used, both for candidates and target. This means that only data points from IR x-projection are selected in order to build the LP model. Same for Experiment 2, data points are obtained from spectra of IR's z projection for both candidates and target. For Exper-

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

iment 3, the spectra information of IR's x and z projections are combined. At last, for Experiment 4, spectra information of IR x-projection and Raman xx-projection are combined. Because we select different spectroscopy information for each experiment, the LP model we build for each experiment is different consequently. However, it is possible that the last one contains the most information here, as the return composition actually matches to the target one.

One of the interesting phenomenon is that when we merely using IR information, the return composition is the same for Experiment 1, 2 and 3. This indicates that LP model constructed by using only IR information have achieved its best capacity, however, it is still not sufficient enough to study a case with the current set-up.

Furthermore, in Figure 4.2, the result spectra is 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. The optimal composition return by this LP model 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 our result. The more constraints are introduced, the more accurate the return composition will be.

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. What if the difference in θ degree for candidates is smaller than $20°$, which is $10°$, will Raman and IR together still be sufficient enough to derive the target composition? Therefore, the following experiments are conducted as shown in Table 4.2.

Experiment 5 shows that the LP model constructed by merely using IR spectral

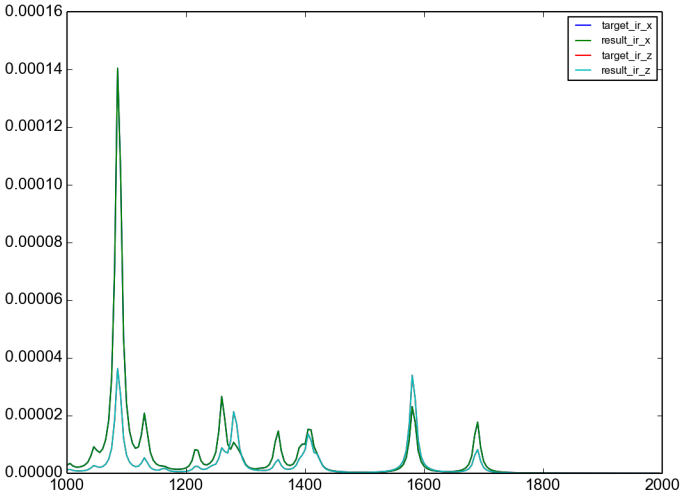

Figure 4.1: Compare target spectra with spectra generated by composition returned by LP model with only IR spectra of x and z projection

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 here.

For experiment setting in Table 4.2 and Table 4.2, 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 4.2. 5 candidates are 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 as well, 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 to 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. Therefore,

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

from Experiment, different kinds of spectral information are mixed. 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, at last Experiment 16 cooperates all the three spectral information, however, none of them returns a composition that matches to 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. The LP model has reached its limitation. 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 in Table 4.3.

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, Experiment 17 and 18 are conducted. Methionine is still used as an example. To make the study case more general than Experiment 1 to 16, candidates are expanded from θ

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

degree of 0° to 80°. In total, there are 9 candidates. Because the SFG spectra for θ of 90° is a straight line, we exclude θ of 90° 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 the spectral information of the three techniques, 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 of them 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 4.3 to Figure 4.3 displays 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 4.3 to Figure 4.3. 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.

4.4 Extra Experiments

From Experiment 1 to 18, LP model helps to return the target composition for some cases,and not for others.Is there a clean line indicating when the information used to generate the LP model is not sufficient to obtain the target composition. 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

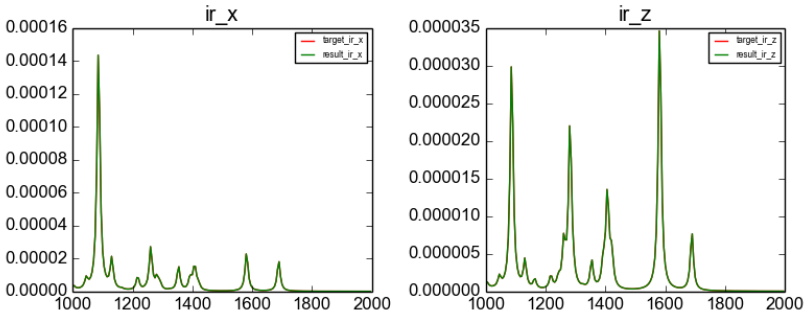

Figure 4.2: IR spectra plotted by using target composition and return composition of Experiment 17

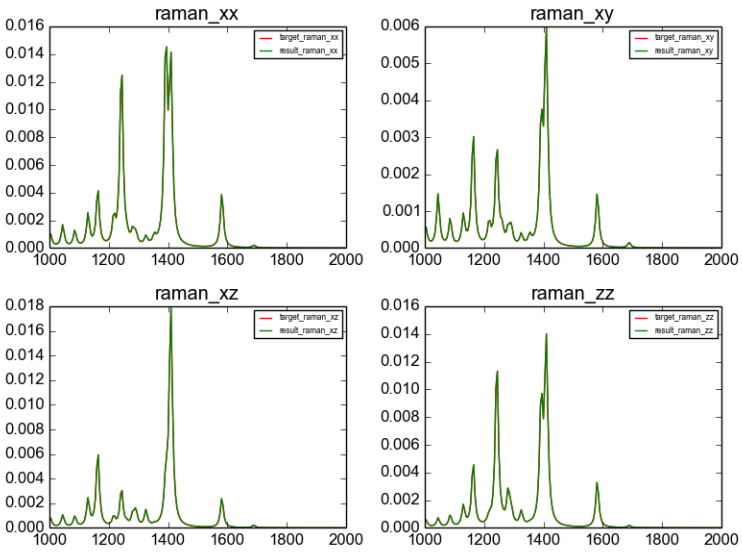

Figure 4.3: Raman spectra plotted by using target composition and return composition of Experiment 17

as Experiment 17 and 18. spread from 0° to 80° on θ.

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

The change is applied to the target composition. For the first group of experiments, only two candidates are selected in the target composition.

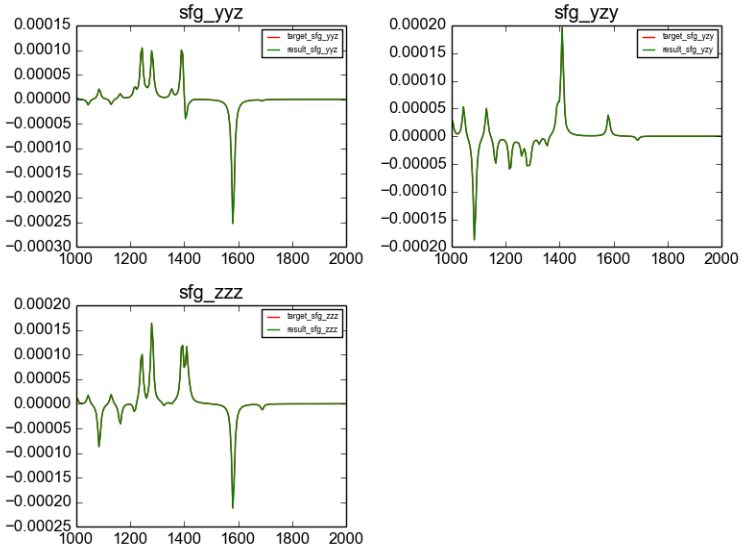

Figure 4.4: SFG spectra plotted by using target composition and return composition of Experiment 17

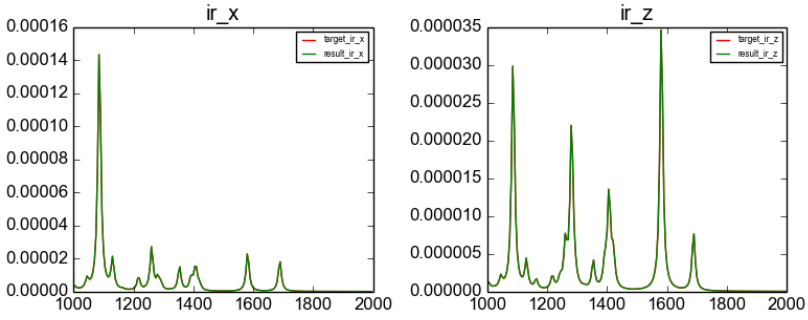

Figure 4.5: IR spectra plotted by using target composition and return composition of Experiment 18

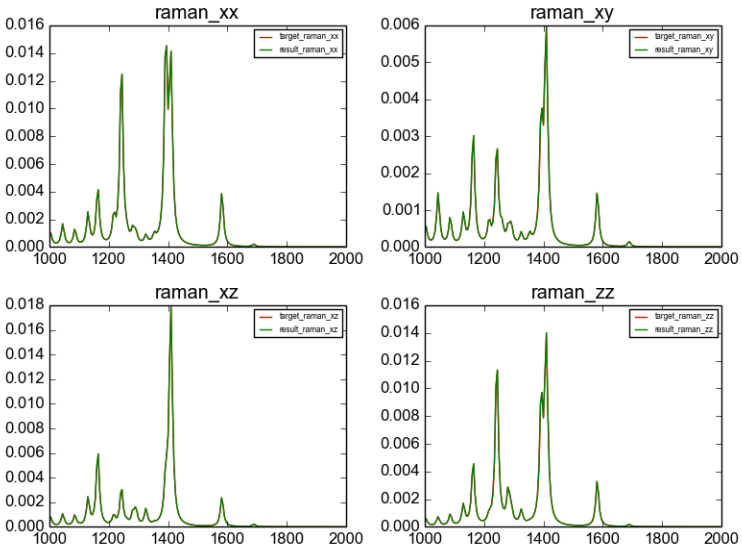

Figure 4.6: Raman spectra plotted by using target composition and return composition of Experiment 18

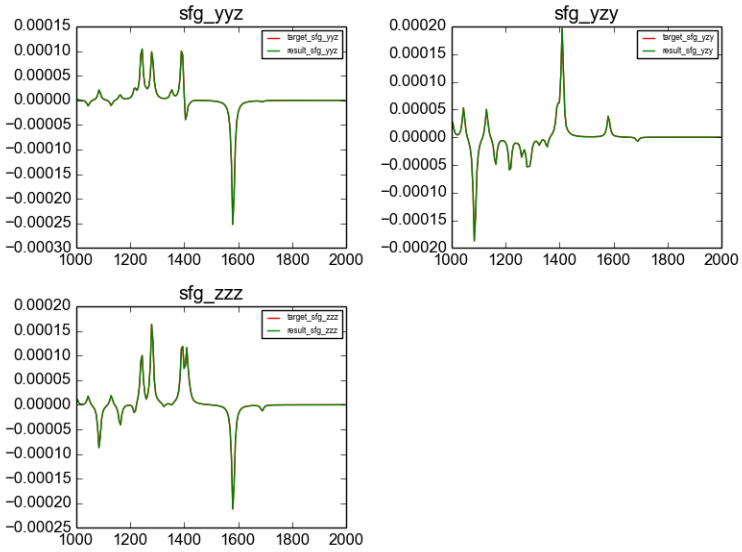

Figure 4.7: SFG spectra plotted by using target composition and return composition of Experiment 18

Chapter 5

Mixture

5.1 Description

In Chapter 4, experiments indicate that for one kind of molecule at interfaces, even combine all the three kinds of spectral information, the LP model constructed can not return the target composition for most cases. The existing spectral information is not adequate to obtain an unique solution for the composition of molecule coordination distribution at interfaces. Multiple compositions can build spectra that are exactly the same as the target ones. These compositions are return by different LP models that using different amount of spectra information. For each LP model, because of the numerical limitation, each LP model returns an optimal solution in the composition.

All the spectral information from three techniques has been extracted in building the LP model for one single kind of molecule at interface. The spectral information is limited in the goal of obtaining the target composition for one molecule. However, that's one factor of our focus. Next step, the focus is moved to the study of multiple different molecules at interfaces. For this mixture of different molecules, will LP models built with different spectral information help to return the given target composition. If any of the LP models success in obtaining the target composition, its accuracy is the key factor for the study as well.

Experiment Index	Spectrum Information
E1	x and z polarized IR spectra
E2	xx, xy, xz and zz polarized Raman spectra
E3	yyz, yzy and zzz polarized SFG spectra
E4	x and z polarized IR spectra; xx, xy, xz and zz polarized Raman spectra
E5	x and z polarized IR spectra; yyz, yzy and zzz polarized SFG spectra
E6	xx, xy, xz and zz polarized Raman spectra; yyz, yzy and zzz polarized SFG spectra
E7	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

5.2 Experiments

In order to achieve the above goals, further experiments are needed. These experiments have the following common settings.

First of all, 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 mix, which are θ of the following degrees: 0°, 10°, 20°, 30°, 40°, 50°, 60°, 70° and 80°. Because the SFG spectra for θ equals 90° is a straight line, this candidate is excluded in the experiments. As a result, there are 54 candidates in the mix.

Secondly, target composition needed to be generated. Two steps are operated: randomly pick one candidate from each amino acid's 9 candidates, then randomly generate a percentage for the selected candidate. The target composition is made up of 6 randomly selected candidates, with each one coming from six different amino acids. The rest 48 candidates contain 0 percentage in the target composition. The total percentage of 6 selected candidate makes 100% component of the mix.

Third, we need to generate the IR, Raman and SFG spectra for all the 54 candidates' and the target.

For each experiment in the experiment set, it contains different spectroscopy information as shown in Table 5.2. For E1 in the table, candidates' IR x and z projections spectra are obtained, as well as target's IR x and z projections spectra generated by dot product of the target composition and all the candidates' spectral data. Then the corresponding LP model is conducted. Therefore, the LP model for E1 only contains IR information.

Same idea for E2, it contains Raman's xx, xy, xz and zz four projections spectral information. E3 only contains SFG's yyz, yzy and zzz three projections spectral information.

Start from E4, spectral information from different spectroscopy techniques are combined. For E4, IR's x and z projections are combined with Raman's xx, xy, xz and zz four projections spectral information in the LP model. For E5, IR's x and z projections are combined with SFG's yyz, yzy and zzz three projections spectra information. For E6, Raman and SFG spectral information are incorporated. At the end, for E7, all three spectral information are put together: IR, Raman and SFG.

The LP models for each experiments are 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 5.2 gives us the target composition with the highest accuracy. This accuracy is measured by the time of each experiment hits the target composition. The scoring mechanism to measure a return composition matches to the target one is described in the section of Scoring Methods.

5.3 Scoring methods

At first glance, the sum of residuals between the spectra composed by the return composition and the target one, is considered to measure the accuracy of the return composition. However, for most experiments conducted earlier, the spectra generated by the return composition are almost identical to the one created by the target one. This sum of residuals is also negligible. It is impossible to use it as a scoring criteria.

The only way to measure the accuracy of the return composition, is to compare it with the target one. Therefore, calculating the sum of the residual between the target composition and the return one directly is the fastest way to evaluate the accuracy of each experiment. The shortage of this approach is that it cannot be used to measure in real life where the target composition is unknown. However, for current experi-

ments, this approach can be a way to evaluate the different spectroscopy techniques. In addition, a glance of which spectroscopy technique have the highest sensitivity can be studied.

The return composition of each experiment in the experiment set will be obtained for each run. Each return composition is compared with the target one to calculate the sum of the residual. If the residual is smaller than a certain threshold, which is $1e-7$, the return composition is considered to be the same as the target one.

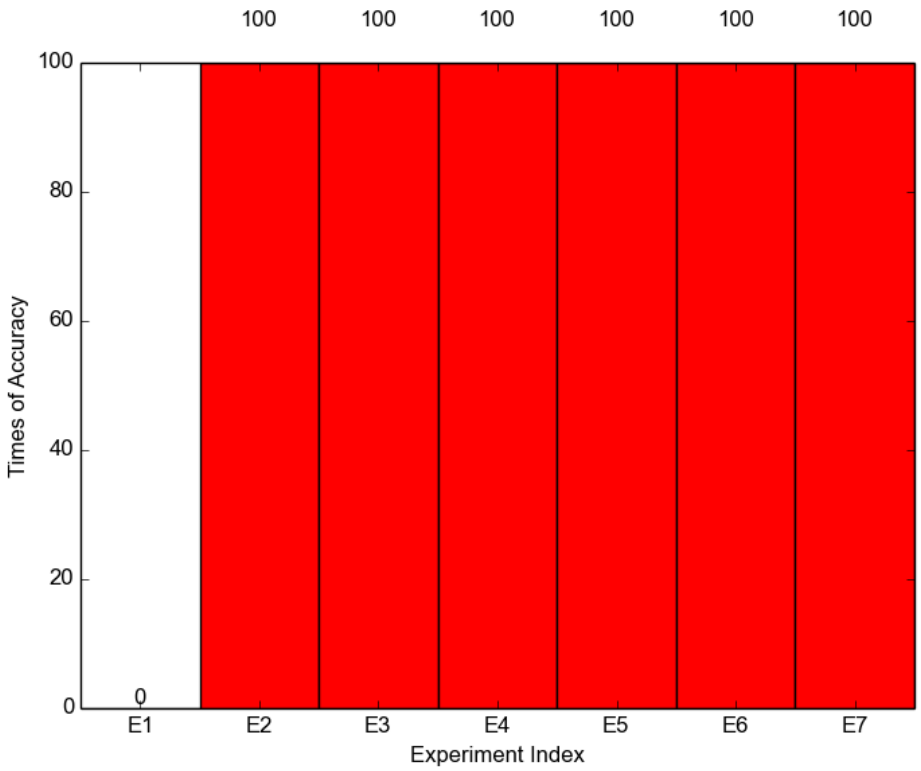

Figure 5.1: Accuracy analysis for experiments considering a mixture of amino acids with candidates from 0° to 80° on θ for each amino acid

After running the experiment set for 100 times, the interesting result is shown in Picture 5.3. E2, the return composition of the LP model constructed using Raman's four projections' spectral information meets the target composition 100 times. This means that within this set of experiments, Raman is sufficient enough to figure out the correct composition for the target spectra. Moreover, the accuracy is 100%.

E3 is the LP model constructed using SFG's three projections' spectral information. Same as E2, its spectral information is as abundant as Raman's spectra for this set of experiments. The accuracy is 100% as well.

E4, E5, E6, and E7, as they either contain information of Raman, or SFG, or both. Therefore, the corresponding LP model can help to get the desire target composition with same accuracy as E2 and E3.

The only exception is E1. The accuracy is not as high as the others. The accuray of LP model constructed using IR's two projections' spectral information is 0. The reason, for the LP model built using only IR spectral information having low performance, can be that the spectral information is not sufficient to obtain the target composition.

When this experiment set is re-run 100 times, only E1's returned composition is analyzed and focused. For each run, x and z projected IR spectra are constructed together both by the returned composition and the target one. The result is that: these two projections' spectra conducted by the two different compositions are almost identical to each other for every run. As an example, a random run is picked, then the two spectra are plotted in Figure A.1. The spectrum plotted by the return composition is identical to the one 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 projections, and Raman has four unique projections. 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 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 projections. Because for Raman's any projection, 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 wavelength.

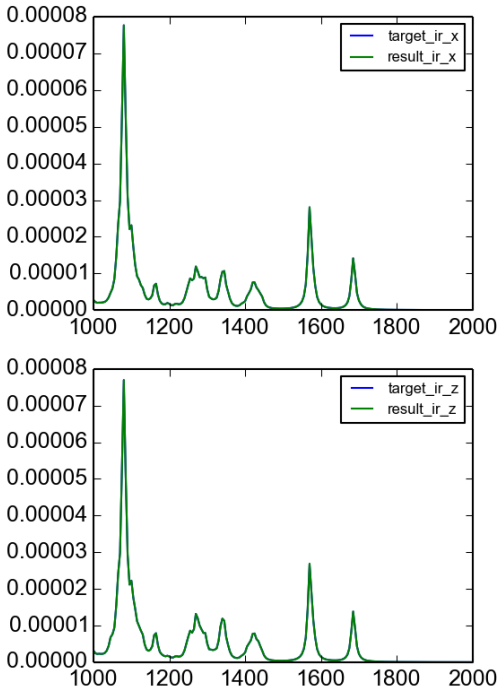

Figure 5.2: IR Spectra Plotted by Result Composition and Target Composition

To further study the capacity of the LP models for the mixture of molecules,the candidate pool is expanded from 0° to 180° on the value of θ degree. Therefore, each amino acid has 18 candidates. In total, there are 108 candidates in the mix. The same set of experiments in Table 5.2 is used. The only different is to to randomly select one candidate from 18 candidates, instead of 9. All 108 candidates' IR, Raman and SFG spectra are needed to be generated. After 100 runs, Figure 5.3 indicates the results. The accuracy for E1 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 5.3. This also increases the difficulty for the LP model constructed by using IR spectral information to return the correct composition.

However, it is very interesting to observe that the accuracy for E2 has dramati-

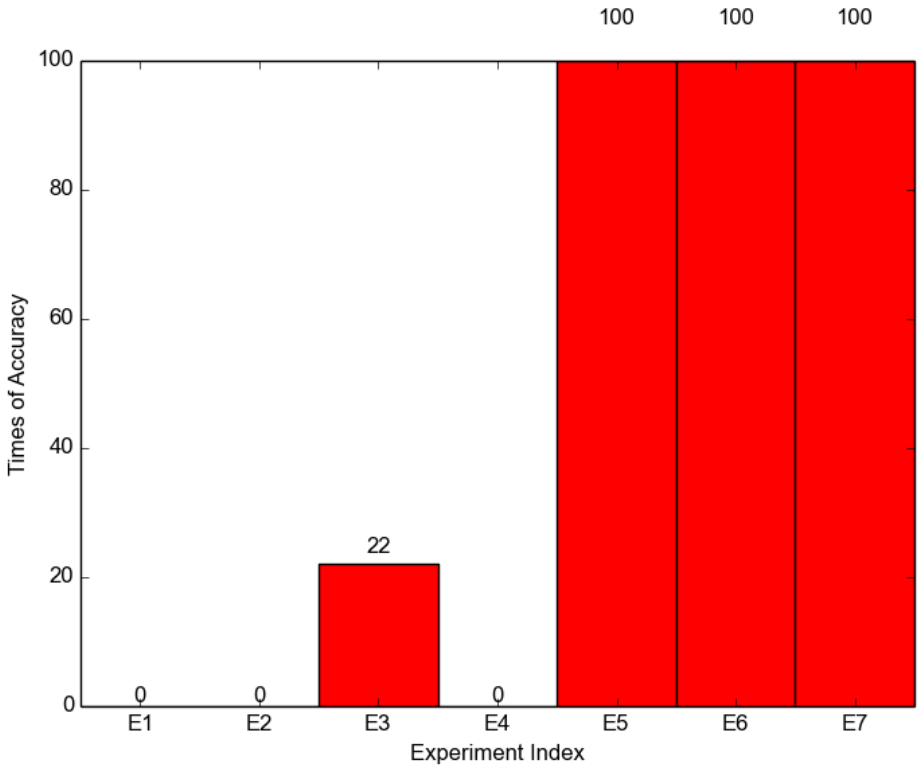

Figure 5.3: Accuracy analysis for experiments considering a mixture of amino acids with candidates from 0° to 180° on θ for each amino acid

cally 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 will be provided on this reason.

Also from Figure 5.3, the accuracy for E3 is no longer high neither. After increasing the candidate number from 9 to 18 for each amino acid, the complexity of the corresponding LP model has increased. From each projection, 200 data points are selected from both target and candidates' spectra. Therefore, there are same number of constraints, in contrary, the number of candidates are twice bigger as before. Although the extra candidates' SFG spectra are symmetric along wavelength which may greatly increase the uniqueness of the candidates. However, the spectral information is still insufficient to converge the composition to the target one.

The good result starts to emerge when we combine IR with SFG or Raman with SFG. Figure 5.3 shows that E5, E6, E7 all have 100% accuracy. This phenomenon

may not be hard to explain at the first glance. SFG helps to distinguish a candidate from its complementary on θ value. Extra spectral information coming from IR or Raman helps to further refine the LP model which can converge the return composition to the target one.
Although the accuracy of E1 is low when each amino acid's candidates spread from 0° to 180° on θ. There are still some interesting result buried in the return composition: for each amino acid, the percentage assigned to it is correct, however, the candidate presented may be the one with correct degree, or the one with correct degree's complementary. Among 100 runs, randomly select one run of the experiment set as an example.

Figure 5.3is the target composition and Figure 5.3 is the return composition for E2. Figure 5.3 is the return composition of E6. From all these three figures, when extracting the non-zero values to generate a list, these three lists are the same. However, when overlapping Figure 5.3 with Figure 5.3, 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 5.3. In Figure 5.3, 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 sometimes. This observation is a general case across all the experiment groups. From E6, as long as the spectra data from SFG is plugged into the LP model, the return composition is the same as the target one.

From the above analysis, E2 appears the ability of limiting the number 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 E4 is the same as the one of E2, 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°.

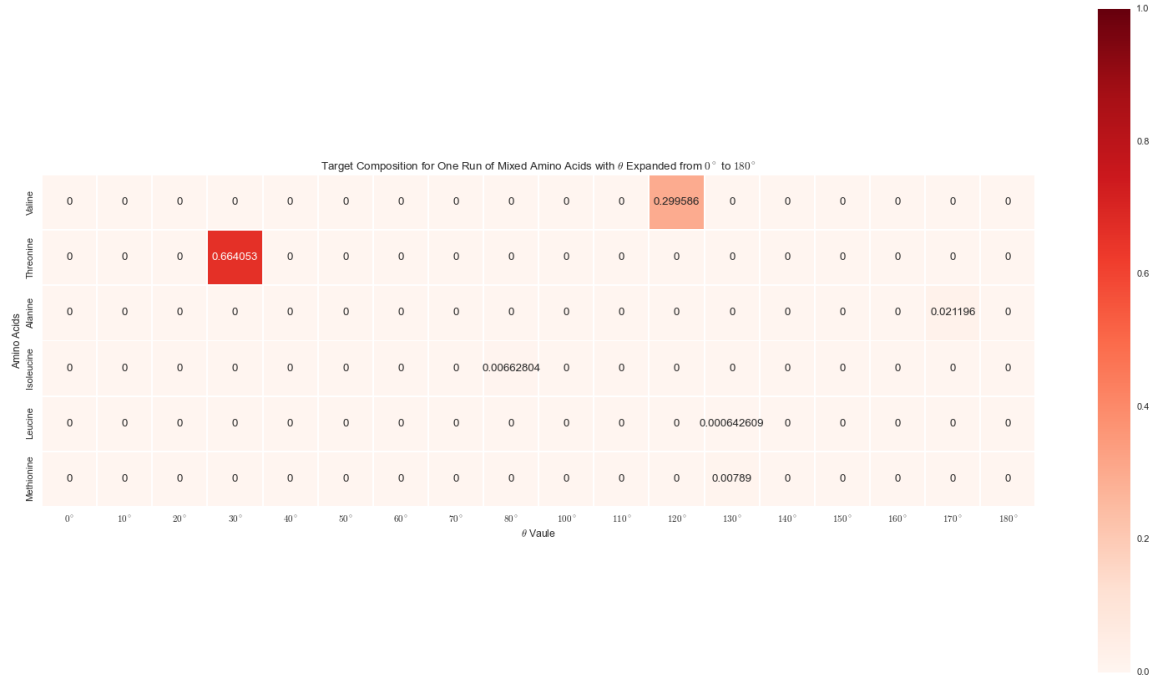

Figure 5.4: Target Composition for One Run of Mixed Amino Acids with θ Expanded from 0° to 180°

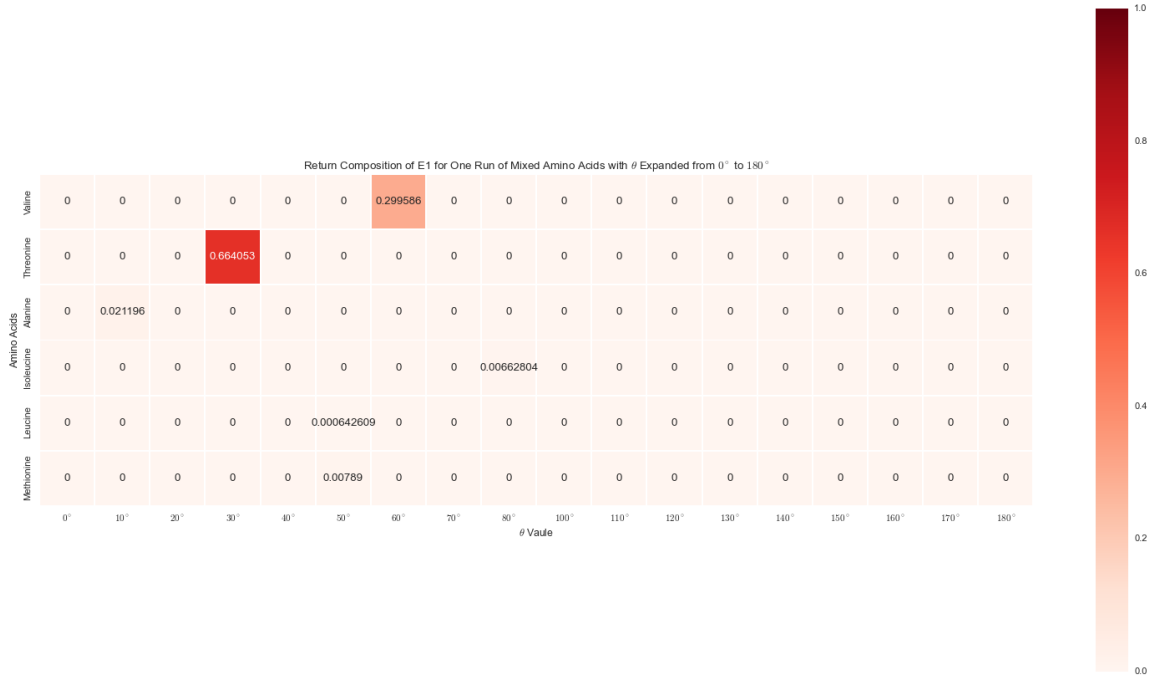

Figure 5.5: Return Composition of E1 for One Run of Mixed Amino Acids with θ Expanded from 0° to 180° on θ

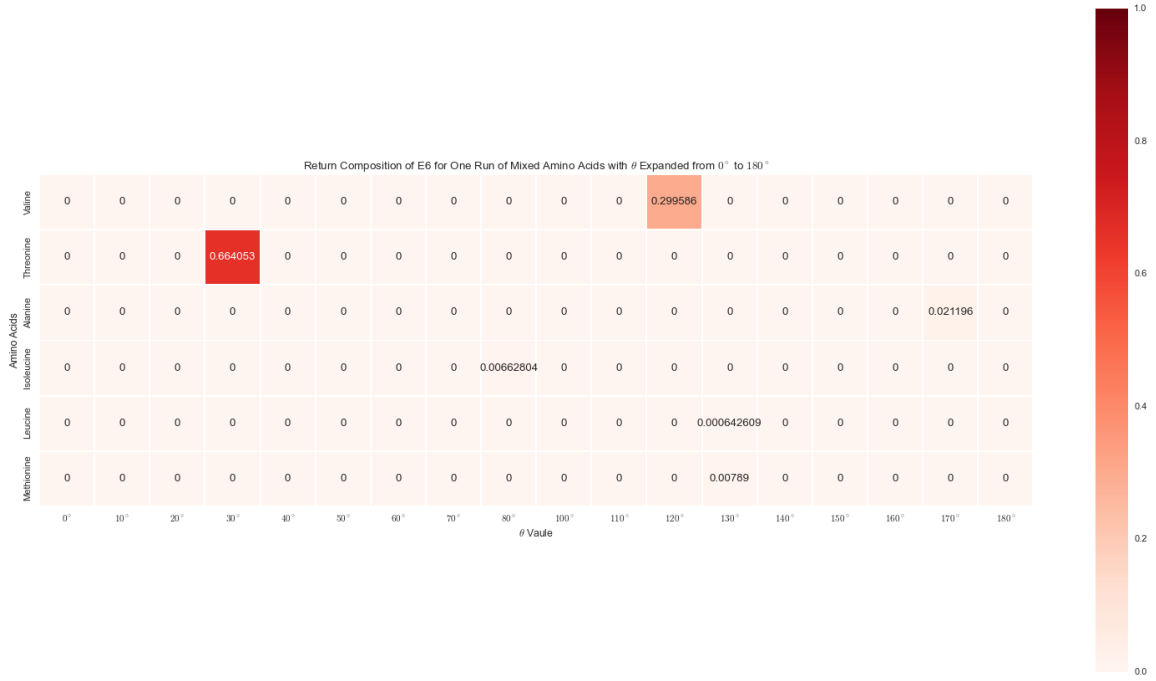

Figure 5.6: Return Composition of E6 for One Run of Mixed Amino Acids with θ Expanded from 0° to 180° on θ

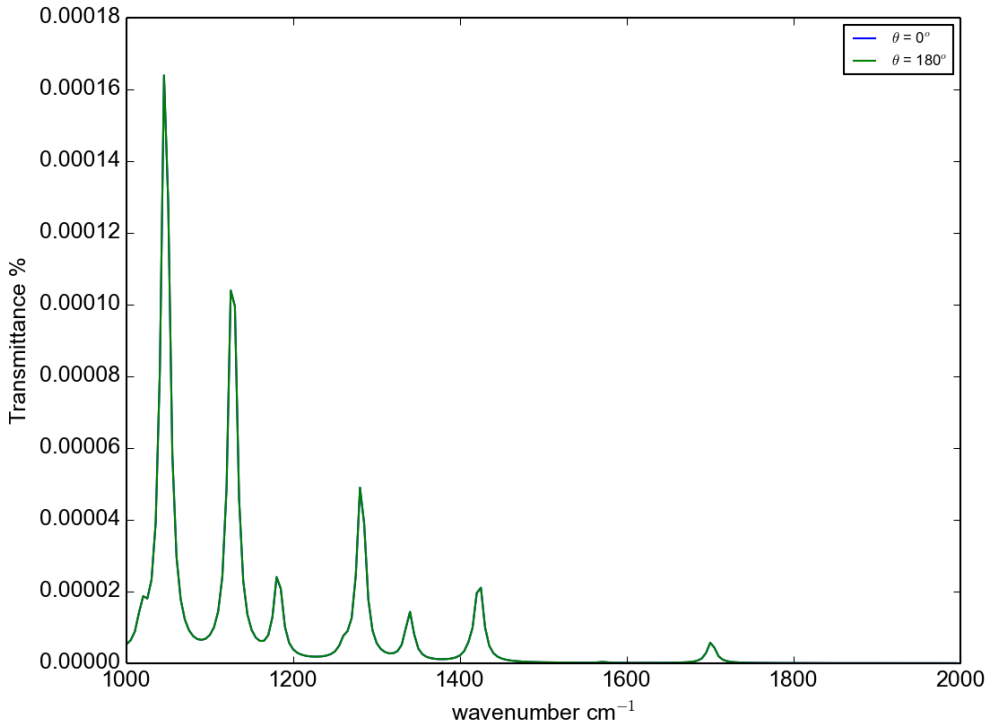

Figure 5.7: IR z projection spectrum for Alanine Candidate with θ of 0° is identical to Alanine Candidate with θ of 180°

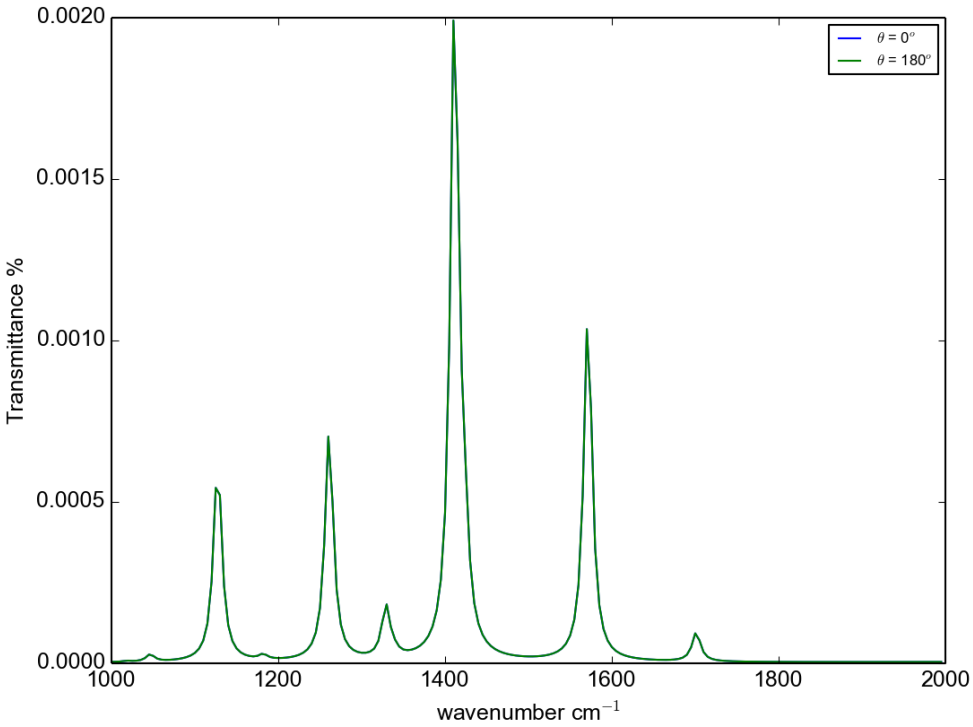

Figure 5.8: Raman zz projection spectrum for Alanine Candidate with θ of 0° is identical to Alanine Candidate with θ of 180°

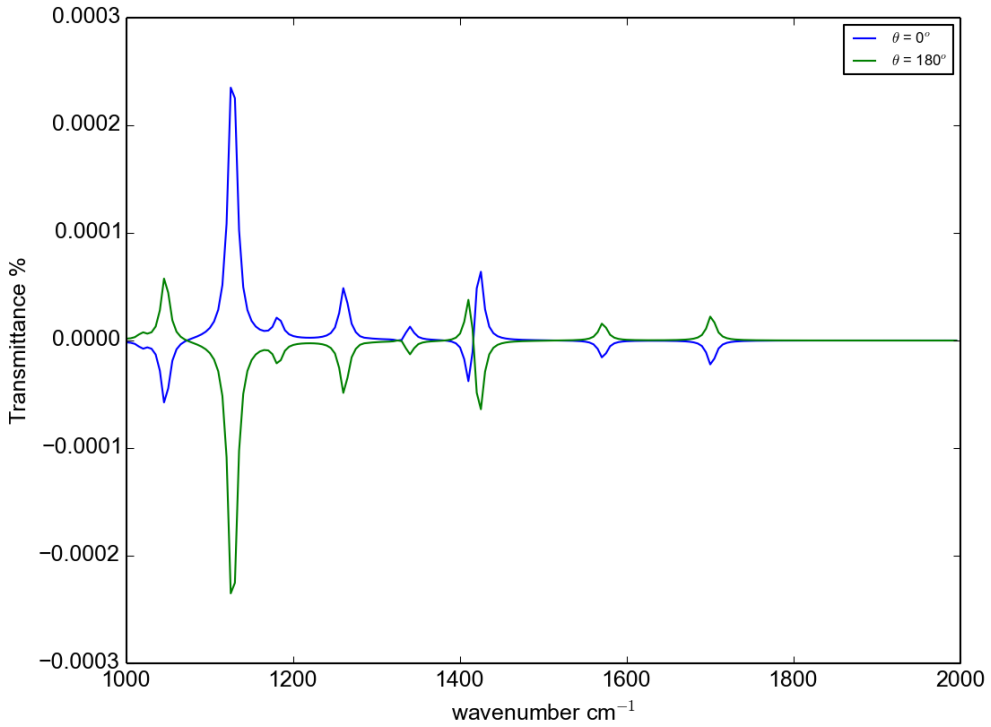

Figure 5.9: SFG zzz projection spectrum for Alanine Candidate with θ of 0° is not identical to Alanine Candidate with θ of 180°, but symmetric along wavelength

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 **??** 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 **??**. A group of 7 experiments with different spectroscopy information. The goal is also the same as what we have in Chapter **??**, 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 **??**, 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 6.1.

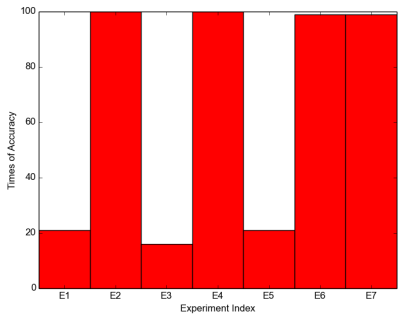

Figure 6.1: Experiment Accuracy Analysis for Experiments using experimental spectra data that contains scaling factor that is smaller than 1 and candidates with θ from 0° to 80°

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 **??**, 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 **??** 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.

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

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

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 **??** 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!

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

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

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

Bibliography

[1] Vasek Chvatal. *Linear Programming.* W. H. Freeman and Company, 1983.

[2] Bernd Cartner Jiri Maousek. *Understanding and Using Linear Programming.*