Emerald Challenge Questions

Quantitative & Qualitative Reasoning

PLEASE CHOOSE 1 OF THE 3 QUESTIONS

Question 1: Mass Spec Data [50 points]

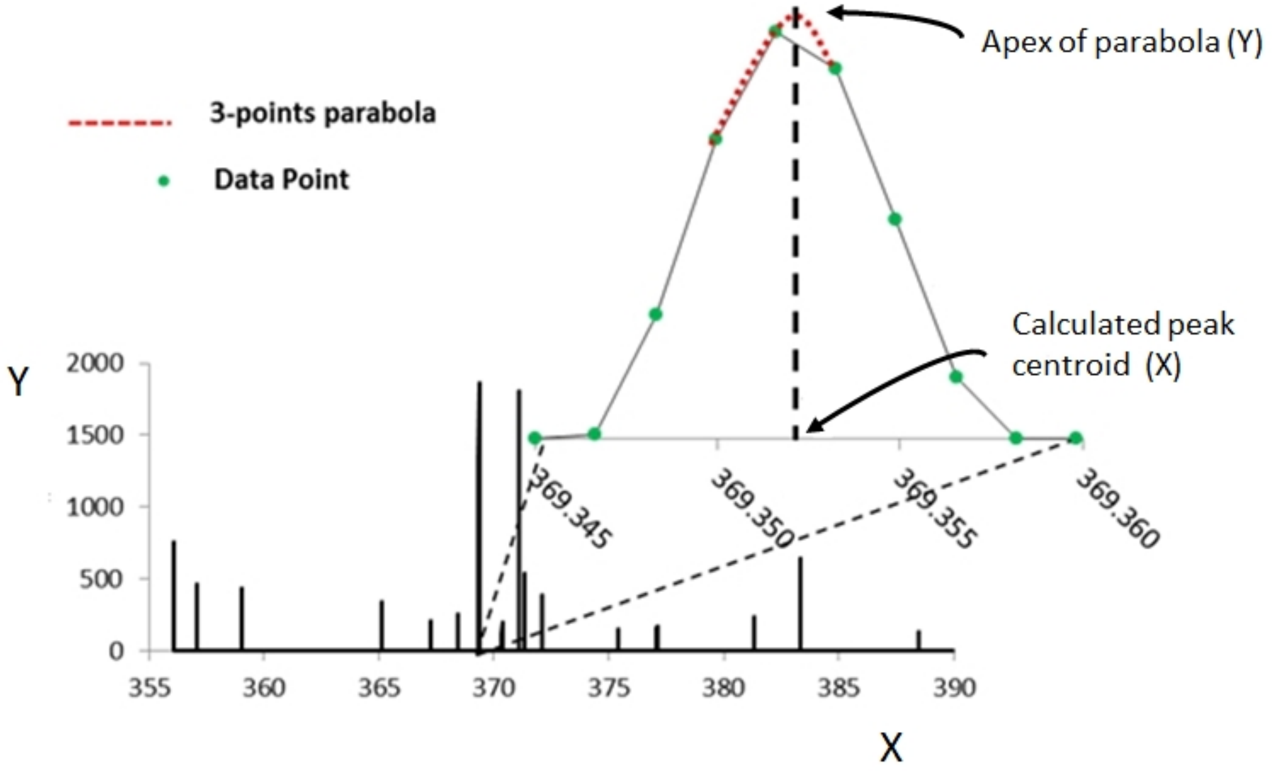
You collected data using an old mass spectrometer for your research project. Unfortunately the software package provided by the manufacturer does not allow you to perform the type of analysis that you really need to do. Determined to perform your own data analysis, you managed to export the raw data for your experiment in a .txt file (see attachment end of this problem). However, just to add to your misery, the exported data is formatted in such a way that x and y value for each data point is embedded in XML-like markup constructs as shown below:

```
...
...
<begin>
    <datapoint>10<datapoint>
    <xunit>mz<xunit>
    <xvalue>150.025461<xvalue>
    <yunit>ab<yunit>
    <yvalue>81.815033<yvalue>
    <time>1412448635<time>
<end>
<begin>
    <datapoint>11<datapoint>
    <xunit>mz<xunit>
    <xvalue>150.025597<xvalue>
    <yunit>ab<yunit>
    <yvalue>110.508636<yvalue>
    <time>1412448635<time>
<end>
...
...
```

Extract the x and y data from the RawSpectrumData.txt file and plot the corresponding mass spectrum (y=f(x)). (15 Points)

Peak picking (35 Points)

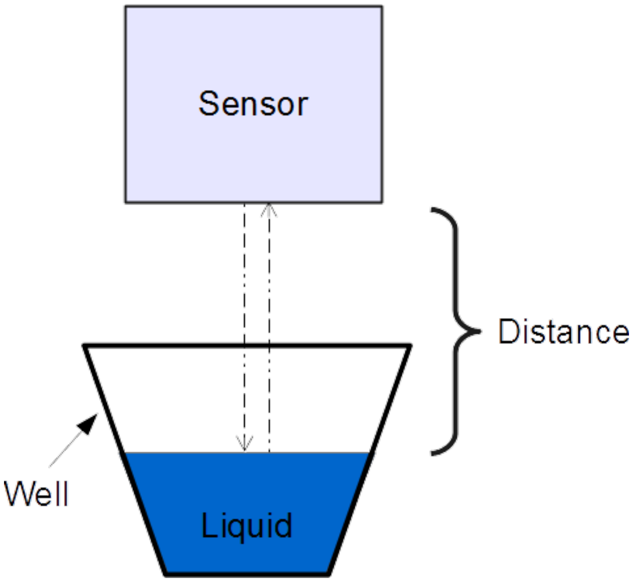
There are different techniques that can be used to calculate the centroid location of peaks in a spectrum. A simple approximation is to fit a parabola on each local maximum and its 2 adjacent points as shown in the figure below. Provide the calculated peak centroids (x) and corresponding calculated y values (apex of parabola) for all peaks that are greater than y=2000.



3-points parabola
Data Point
Apex of parabola (Y)
Calculated peak centroid (X)
Y
2000
1500
1000
500
0
355
360
365
370
375
380
385
390
X
369.345
369.350
369.355
369.360

Question 2: Calibration of Volume Measurements [60 Points]

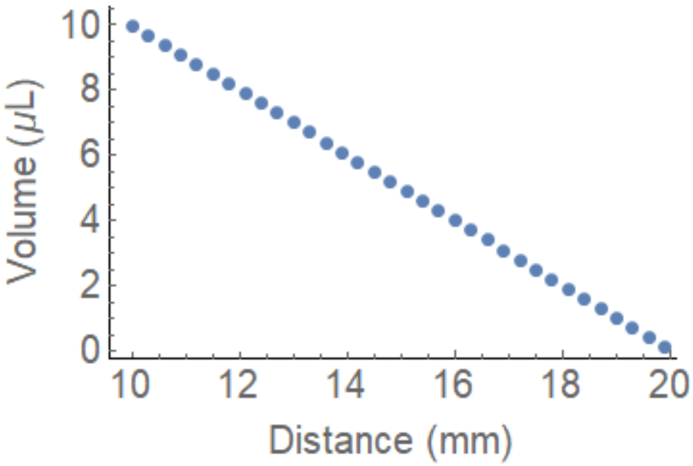
There is an instrument in your lab that helps calculate the volume of liquid in different types of wells. The instrument contains a sensor that sends a signal down to the surface of the liquid in the well, computes the distance travelled by the signal based on elapsed time. Using this measured distance, it's possible to determine the amount of liquid in the well, provided we have an accurate calibration curve for the well.



Since the relationship between distance measured by the instrument and volume in the well depends on the shape of the well, every different well shape needs its own calibration curve that can be used to transform distances measured by the instrument into actual volumes.

Your lab recently bought a new type of well and it is your job to calibrate it. The first step of calibration is to measure out volumes ranging from 0μL to 10μL, and then measure the distance for each of those volumes in the new well.

For example, if the well was cylindrical, the data might look like



The next step is to find a function that accurately describes the relationship between volume and height for that well. That is, find ‘f’ such that

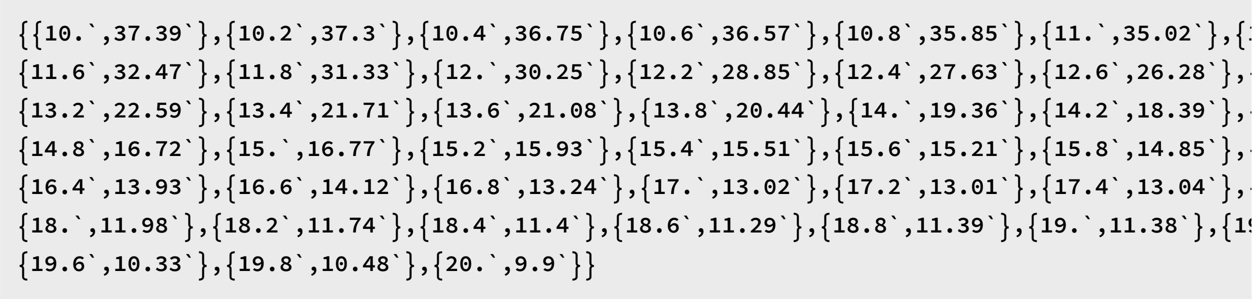
Volume = f (Distance)

accurately describes your data. Once you have that expression you can convert back and forth between volumes and heights for that particular well.

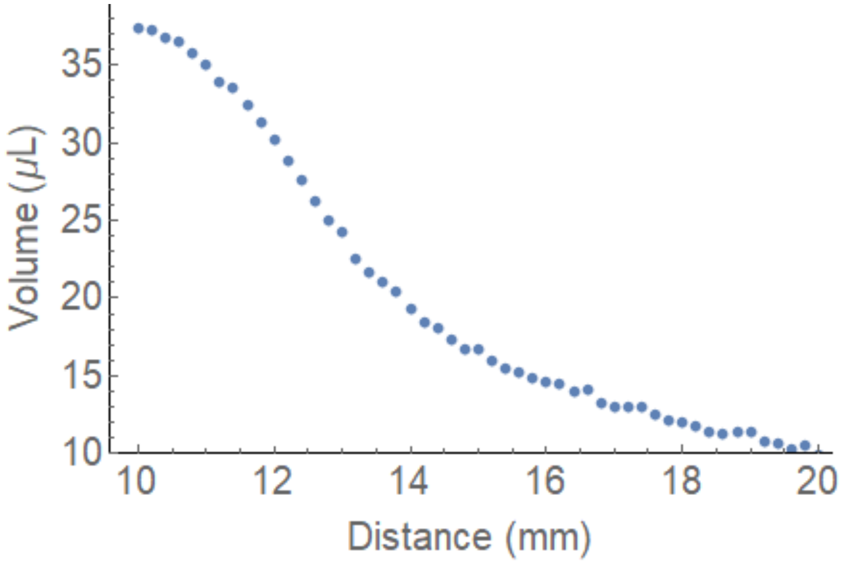
The last step is to perform an error analysis of your calibration, taking into account as many sources of error as possible

1) If the new well happened to be proportioned exactly like the one shown in the first image above (the well containing the blue Liquid), what would the Volume vs Distance curve look like? Be as precise as possible. [10 pts]

2) The actual data measured is given below. Find a function ‘f’ that best fits the data. Describe how you chose it and what criteria you used to determine quality of the fit. [15 pts]



```
{{10.`,37.39`},{10.2`,37.3`},{10.4`,36.75`},{10.6`,36.57`},{10.8`,35.85`},{11.`,35.02`},{1
{11.6`,32.47`},{11.8`,31.33`},{12.`,30.25`},{12.2`,28.85`},{12.4`,27.63`},{12.6`,26.28`},{
{13.2`,22.59`},{13.4`,21.71`},{13.6`,21.08`},{13.8`,20.44`},{14.`,19.36`},{14.2`,18.39`},{
{14.8`,16.72`},{15.`,16.77`},{15.2`,15.93`},{15.4`,15.51`},{15.6`,15.21`},{15.8`,14.85`},{
{16.4`,13.93`},{16.6`,14.12`},{16.8`,13.24`},{17.`,13.02`},{17.2`,13.01`},{17.4`,13.04`},{
{18.`,11.98`},{18.2`,11.74`},{18.4`,11.4`},{18.6`,11.29`},{18.8`,11.39`},{19.`,11.38`},{19
{19.6`,10.33`},{19.8`,10.48`},{20.`,9.9`}}
```



3) What does the cross section of the new well look like? Draw or plot the shape. Be as precise as possible. [10 pts]

4) Your colleague tells you the calibration predicted a volume of 3μL for their measurement. What is the height of the liquid in the well? [10 pts]

5) In order to quantify the uncertainty of the volume measurements, you perform repeated measurements of the same volumes (below). Additionally, your pipetter claims to have an error of 2%. Using this new information, what is the uncertainty in your predicted volume if the instrument measures 15 mm? [15 pts]

Note: these measurements are from the same instrument, but from a DIFFERENT well, so the distance-volume values don't match your other data points.

Each point is {volume,measuredDistance}

```
{{5.`,15.2`},{5.`,14.93`},{5.`,15.13`},{5.`,14.91`},{5.`,15.06`},{5.`,14.96`},{5.`,14.85`
{10.`,20.06`},{10.`,19.81`},{10.`,19.92`},{10.`,20.3`},{10.`,20.22`},{10.`,19.99`},{10.`,
{15.`,24.96`},{15.`,24.85`},{15.`,24.46`},{15.`,25.26`},{15.`,24.71`},{15.`,25.35`}}
```

Question 3: International Encoding of Amino Acids [35 points]

You have been asked to head a committee designed to encode a database of human protein sequences, which are all built from the following list of twenty amino acids, and the code for completion of a message (stop codon):

```
   Alanine      Arginine   Asparagine  Aspartic Acid    Cystine
Glutamic Acid  Glutamine    Glycine      Histidine     Isoleucine
   Leucine      Leucine      Lysine      Methionine   Phenylalanine
   Proline       Serine     Threonine    Tryptophan     Tyrosine
    Valine
```

The committee has approved the use of two international symbols to be used in strings to represent each amino acid in the database, a red circle and a blue triangle:

Number	1	2
Symbol	●	▲

Initially the committee suggests using a 5 character string encoding system that uses a combination of five symbols to represent each amino acid as follows:

Amino Acid	Encoding
Alanine	●●●●●
Arginine	▲●●●●
Asparagine	●▲●●●
Aspartic Acid	▲▲●●●
Cystine	●●▲●●
Glutamic Acid	▲●▲●●
Glutamine	●▲▲●●
Glycine	▲▲▲●●
Histidine	●●●▲●
Isoleucine	▲●●▲●
Leucine	●▲●▲●
Lysine	▲▲●▲●
Methionine	●●▲▲●
Phenylalanine	▲●▲▲●
Proline	●▲▲▲●
Serine	▲▲▲▲●
Threonine	●●●●▲
Tryptophan	▲●●●▲
Tyrosine	●▲●●▲
Valine	▲▲●●▲
Stop Codons	●●▲●▲

However, through your superior knowledge of biochemistry, you know that each of these amino acids do not appear in your dataset with equal frequency. After doing a quick survey of the database you find that the distribution of amino acids is actually the following:

Amino Acid	Frequency (%)
Alanine	7.4
Arginine	4.2
Asparagine	4.4
Aspartic Acid	5.9
Cystine	3.3
Glutamic Acid	5.8
Glutamine	3.7
Glycine	7.4
Histidine	2.9
Isoleucine	3.8
Leucine	7.5
Lysine	7.2
Methionine	1.8
Phenylalanine	4.
Proline	5.
Serine	8.
Threonine	6.2
Tryptophan	1.3
Tyrosine	3.3
Valine	6.8
Stop Codons	0.1

A) Given your knowledge of the frequencies of each amino acid, what encoding system could you use to most efficiently store the entire dataset with the least number of symbols?(15 points)

B) How much space would you save (as a function of % fewer symbols) by encoding the dataset this way? (5 points)

C) Suppose the committee introduces another 2 symbols and a new default encoding (see below). Again, given your knowledge of the frequencies of each Amino acid, what encoding system could you use to most efficiently store the entire dataset with the least number of symbols? (10 points)

Number	1	2	3	4
Symbol	🔴	▲	🟩	⬟

Default Encoding:

Amino Acid	Encoding
Alanine	🔴🔴🔴
Arginine	▲🔴🔴
Asparagine	🟩🔴🔴
Aspartic Acid	⬟🔴🔴
Cystine	🔴▲🔴
Glutamic Acid	▲▲🔴
Glutamine	🟩▲🔴
Glycine	⬟▲🔴
Histidine	🔴🟩🔴
Isoleucine	▲🟩🔴
Leucine	🟩🟩🔴
Lysine	⬟🟩🔴
Methionine	🔴⬟🔴
Phenylalanine	▲⬟🔴
Proline	🟩⬟🔴
Serine	⬟⬟🔴
Threonine	🔴🔴▲
Tryptophan	▲🔴▲
Tyrosine	🟩🔴▲
Valine	⬟🔴▲
Stop Codons	🔴▲▲

D) How much space would you save (as a function of % fewer symbols) by encoding the dataset as described the question above? (5 points)

Powers of Estimation

PLEASE ANSWER ALL OF THE 5 QUESTIONS

Please attempt to estimate solutions to the following question, listing any assumptions, data sources, formulas, and comparables you used to reach the conclusion.

Question 1: On a typical day, how many times in the world is the English word "Hello" spoken? [10 points]

Question 2: How large of an area of solar panels would you need to set up be able to fully change a Tesla Model 3 in one day of charging in the Sahara desert? [10 points]

Question 3: How large of a swimming pool would you need to hold all of the coffee sold by Starbucks each day? [10 points]

Question 4: In designing a futuristic Jet, how fast would it need to travel to be able to visit the worlds the 10 most populous cities in a single day? [10 points]

Question 5: How much would it cost in materials to tag every functioning laptop in the world with a unique QR code? [10 points]

Life Sciences

PLEASE CHOOSE 1 OF THE 3 QUESTIONS

Question 1: Microbiology [50 points]

A) You identified a novel bacterium from the sample collected from your latest expedition to Amazonia. This strain has some intriguing properties–for example, it grows really fast that double time is about 15 minutes in common growth media. You figured that it has potential to be an alternative to classic lab bacterial strain such as *Escherichia coli* and could facilitate scientific researches, but some works need to be done first.

A1) Safety first. You would like this bacterium to be a safe biological agent in general and satisfy the requirement for Biosafety Level 1 (BSL-1). Describe what characteristics of this bacterium and how do you examine them to make sure it's safe. How would you you further improve the safety feature of this bacterium? (10 Points)

A2) What other features does this bacterium need in order to serve as a useful research tool for fields like biochemistry, molecular biology and genomics to study nucleic acids and proteins? List at least three and explain why they are important. (10 Points)

A3) If some of these features are missing, how would you implement them into this bacterium? Choose two features and describe the strategy you would use. (10 points)

B) You've successfully got the strain you want, and it's named *Emeraldia Chlora* LB1. The fact that *E. Chlora* grows so fast essentially means it's efficient in biosynthesis and gene expression, and hence you want to use its cell extract to develop a *in vitro* gene expression system.

B1) Describe how would you develop *E. Chlora in vitro* gene expression system that can take DNA or RNA molecules as input and produce the proteins encoded. (5 Points)

B2) You've got some preliminary results with a working *E. Chlora in vitro* gene expression system, using green fluorescent protein (GFP) as output so you can use fluorescence signal to evaluate the protein synthesis efficiency. Unfortunately, the output signal of GFP is not as high as you expected–far lower than the amount of amino acids provided for building proteins. Describe what's your hypothesis about the low protein synthesis efficiency, how to test it, and how to improve it. (5 Points)

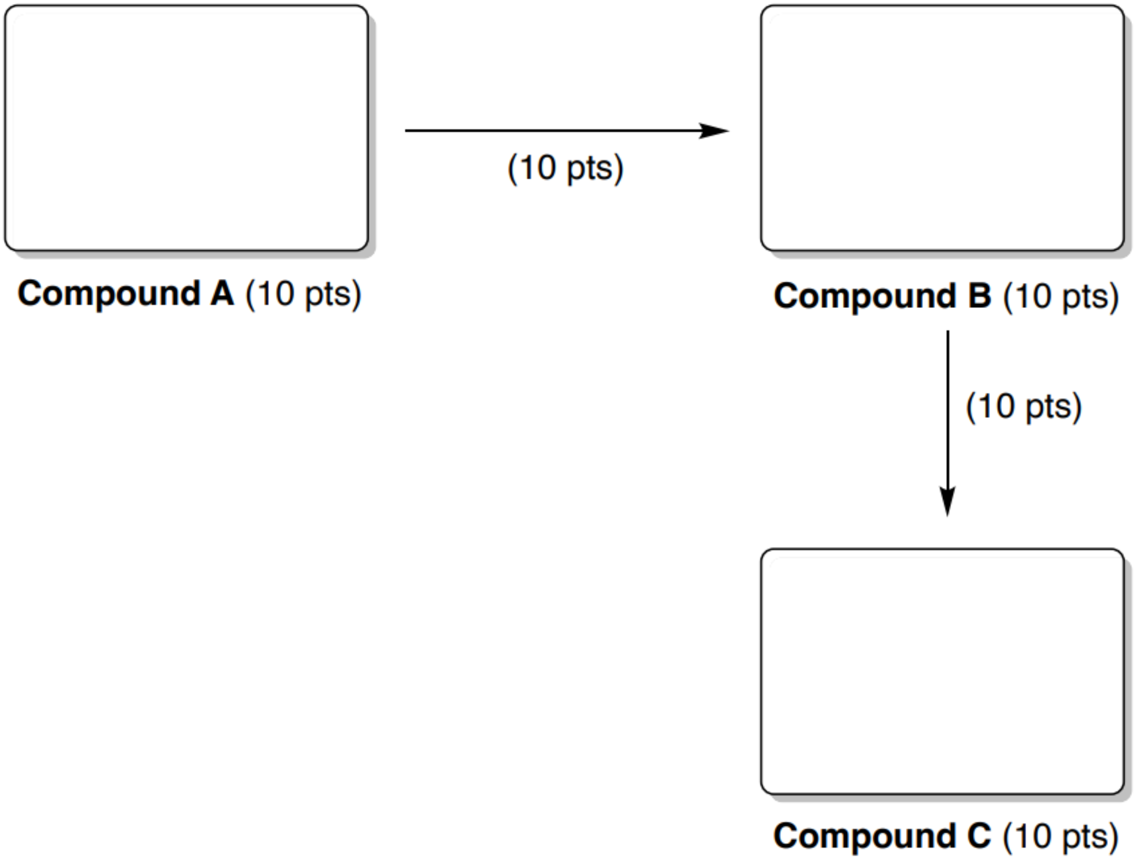
B3) Design one logic gate (AND, OR, NOR, etc) genetic circuit with living E. Chlora cells or E. Chlora in vitro system that can perform Boolean operation that will produce fluorescent signal if output is True and no signal if False. Describe which logic gate, which system, what are your inputs, and the molecular mechanism of your circuit. (5 Points)

B4) Design an genetic oscillator with living E. Chlora cells or E. Chlora in vitro system. Describe the molecular mechanism of your circuit and based on your design, how could you tune the key parameters of oscillation, such as amplitude or period? (5 Points)

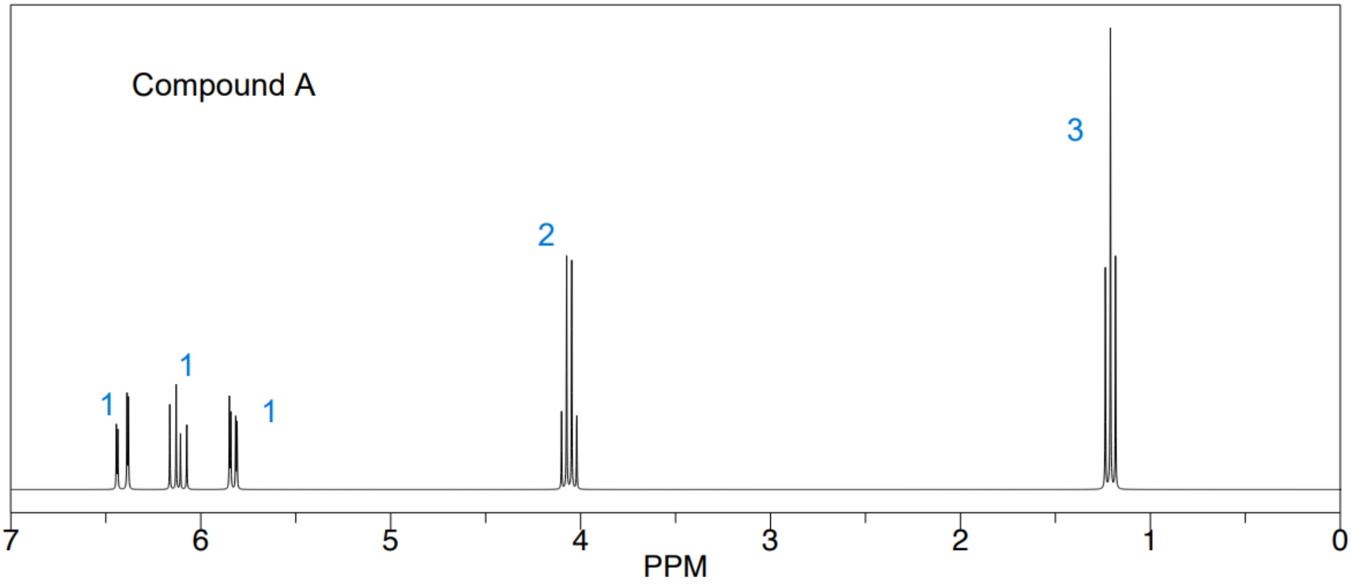
C) You would like to further expand the capability of E. Chlora that it can utilize a novel tryptophan analog synthesized in the lab and put it into a specific site on a protein it's producing. Describe your strategy you would use to achieve this goal. (10 Points)

Question 2: NMR [40 points]

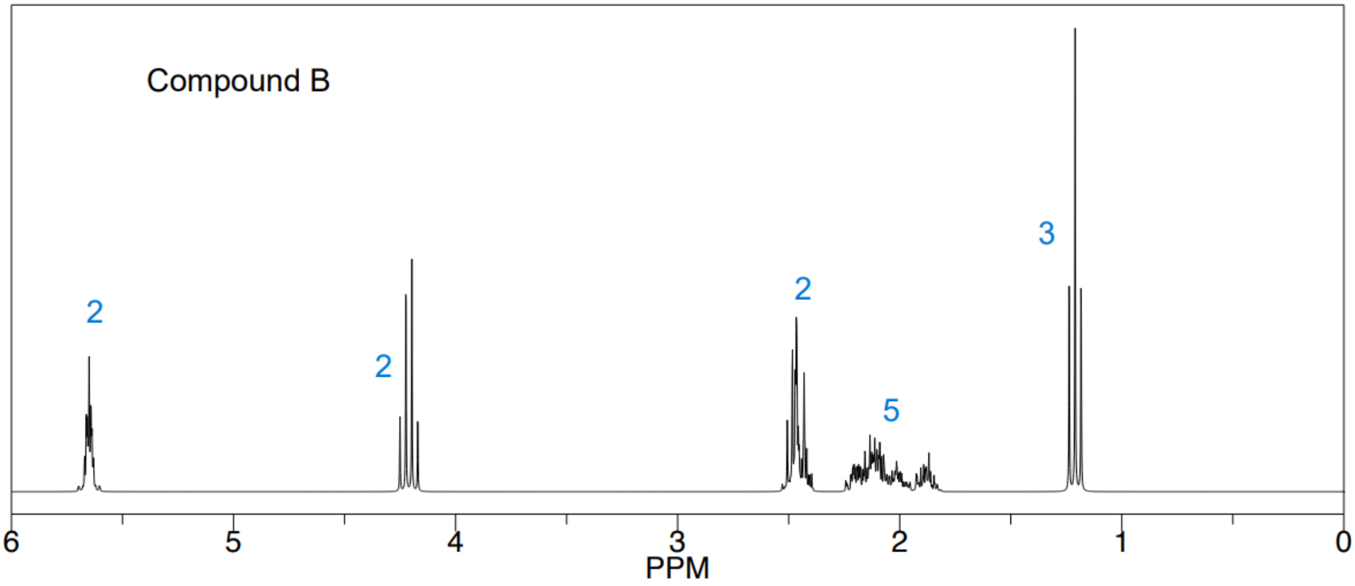
Given the included NMR spectra (integrations of peaks indicated with blue numbers), draw the structures for **Compound A**, **Compound B**, and **Compound C**, as well as the conditions/reagents/etc necessary for converting **A** to **B** and **B** to **C** (10 points per structure and 10 points per conditions)



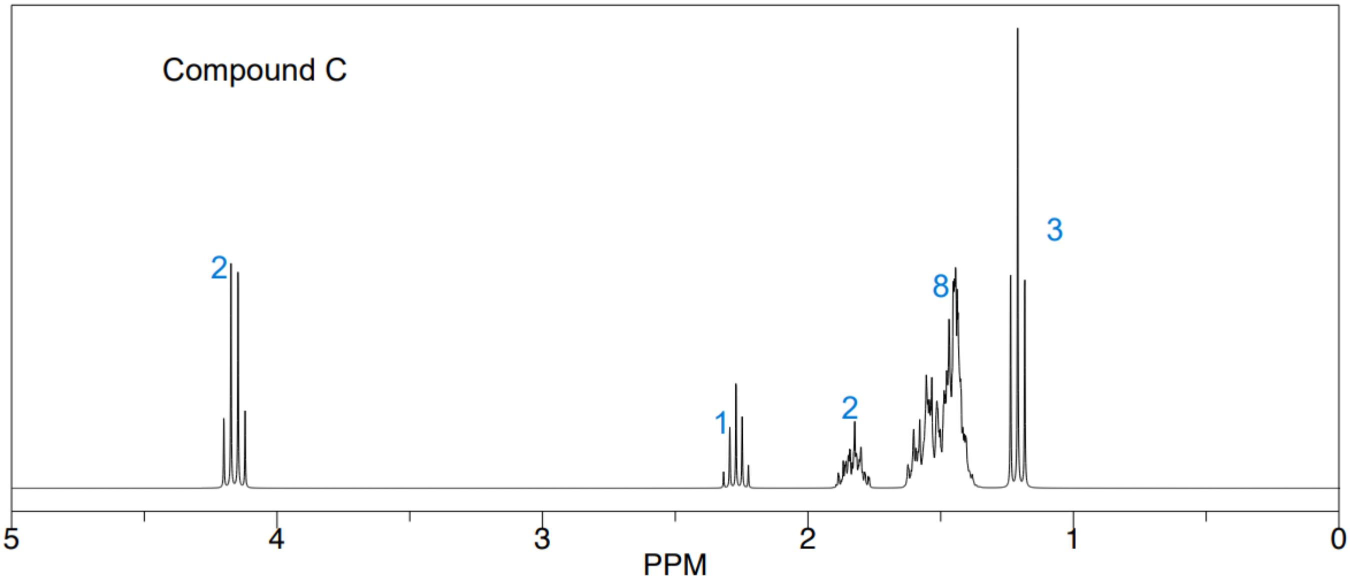
Compound A



Compound B



Compound C



Question 3: Kinetics [45 Points]

To successfully answer any of the following parts, please provide a clear presentation of the steps required to solve the problem. Also, please state any assumptions used at the beginning of your answer. No points will be awarded for simply providing numeric answers without any justification or explanations.

Since Emerald is a computationally driven company, computational answers are encouraged. In this case, please provide any code/scripts used to arrive at your conclusions.

Part A: Bimolecular reaction

Given the following simple reaction mechanism:

$A + B \rightarrow C$

Where the forward rate constant for this reaction at 25° Celsius is known to be 1.6×10^4 Molar^{-1} Second^{-1} and unless otherwise noted initial concentrations are as follows:

$[A]_0 = 8.8\ \mu$M
$[B]_0 = 6.4\ \mu$M
$[C]_0 = 0$

If the reaction is run at 25° Celsius for 10 seconds, what would you expect the concentration of [C] to be? (5 points)

Part B: Autocatalysis

Now, let's consider a bimolecular reaction with the following mechanism:

$A + B \rightarrow 2B$

The forward rate constant for this reaction at 25° Celsius is known to be 1.1×10^5 Molar^{-1} Second^{-1}, and the initial conditions are as follows:

$[A]_0 = 7.8\ \mu$M
$[B]_0 = 0.1\ \mu$M

Part 1: Plot the concentration of B from 0 to 10 seconds. (5 Points)
Part 2: Describe how the behavior of this reaction is fundamentally different from the reaction in Question A (5 Points)

Question C: Chemical clocks

Let's consider a system of auto-catalytic reactions with the following mechanisms:

(Reaction 1):	$A + B \rightarrow 2B$
(Reaction 2):	$B + C \rightarrow 2C$
(Reaction 3):	$C + A \rightarrow 2A$

The rate constants are as follows:

$k_1 = 1.1 \times 10^5$ Molar^{-1} Second^{-1}

$k_2 = 1.0 \times 10^5$ Molar^{-1} Second^{-1}

$k_3 = 0.9 \times 10^5$ Molar^{-1} Second^{-1}

The initial conditions are:

$[A]_0 = 7.8\ \mu\text{M}$

$[B]_0 = 0.1\ \mu\text{M}$

$[C]_0 = 2.3\ \mu\text{M}$

The B molecule has a visible green color, the others are transparent.

Part 1: Plot the concentration of B from 0 to 20 seconds (10 Points)

Part 2: At what frequency does the chemical system oscillate (in 1/seconds)? (10 Points)

Question D: Control the clock!

Consider the same set of reactions as in Question C. Assume that the temperature of the system, and the reaction rate constants cannot be changed. You can change the initial concentrations of the system.

The task is to find the initial conditions that halves the frequency found in Part 2 of Question C

Part 1: Qualitatively describe how you would search the space of possible solutions for a configura tion of initial concentrations that halves the frequency found Part 2 of Question C. (5 Points) Part 2: Implement your solution and provide the values for the initial concentration of molecules A, B, and C. (10 Points)