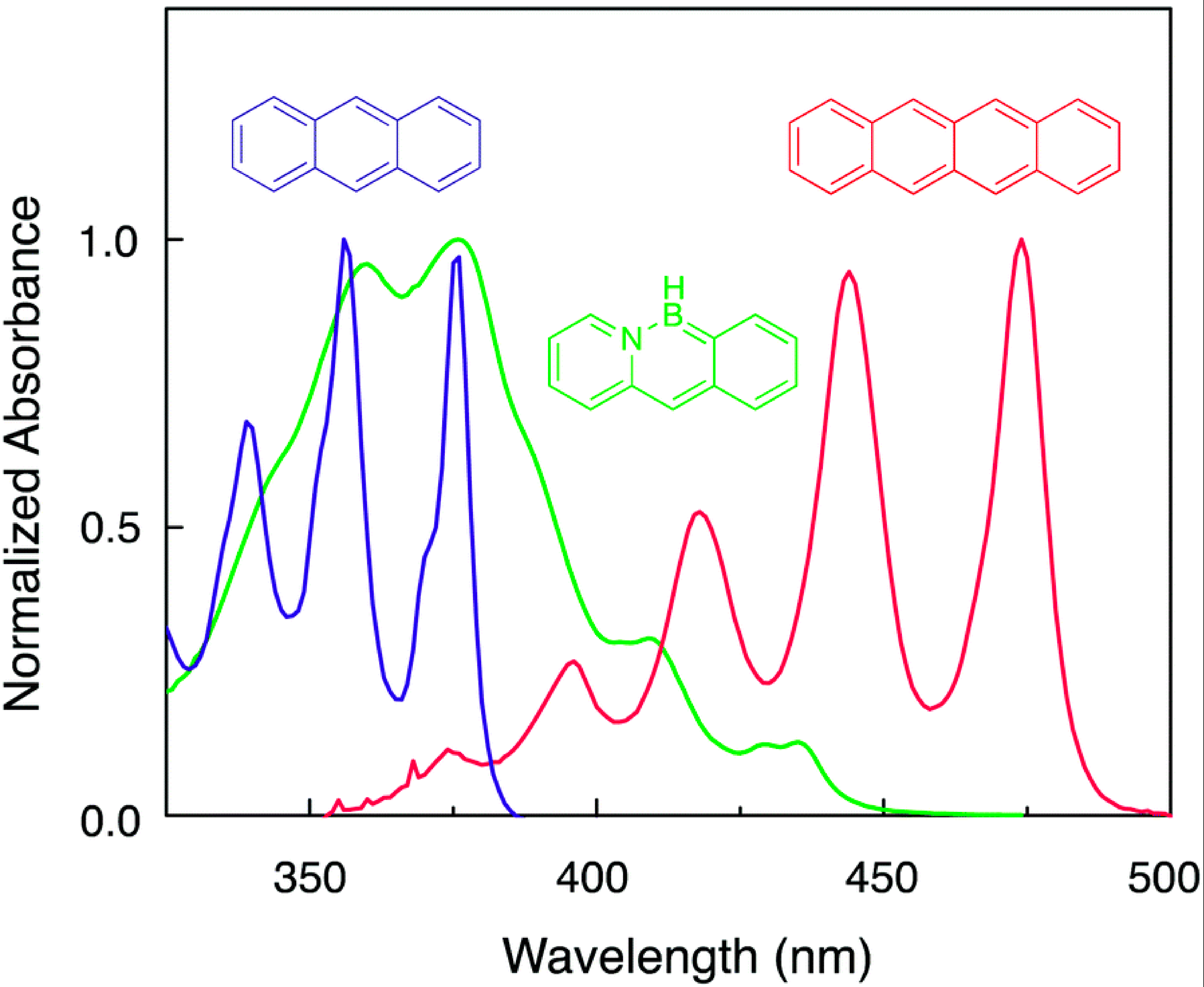


Machine Learning for Chemo-Informatics: Application to IR Spectroscopy Data

Advisor: Dr. Bhaskar Chaudhary

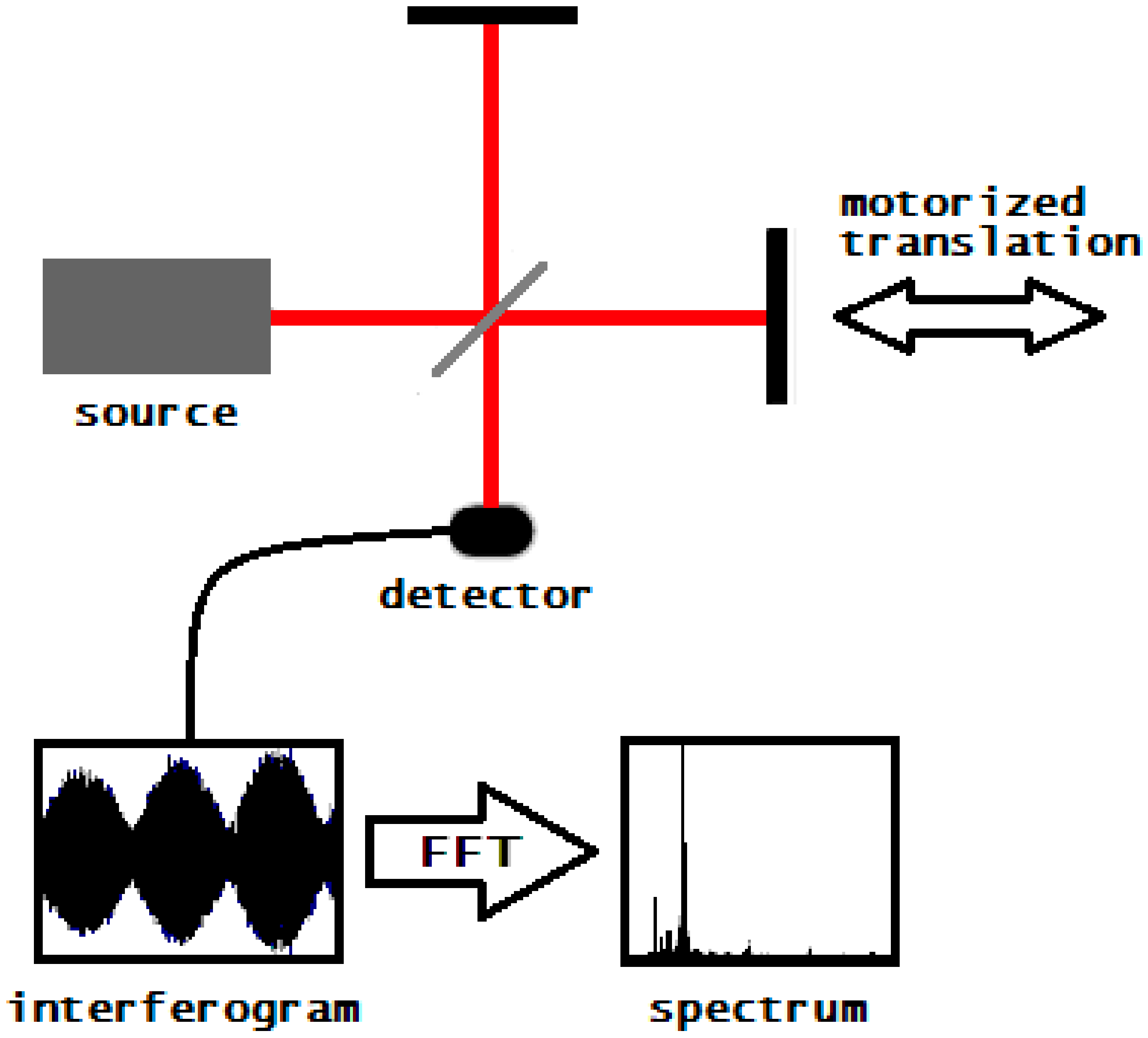
Presentation by
Tipsi Jadav: 201801091
Ujas Thakkar: 201801112

INTRODUCTION

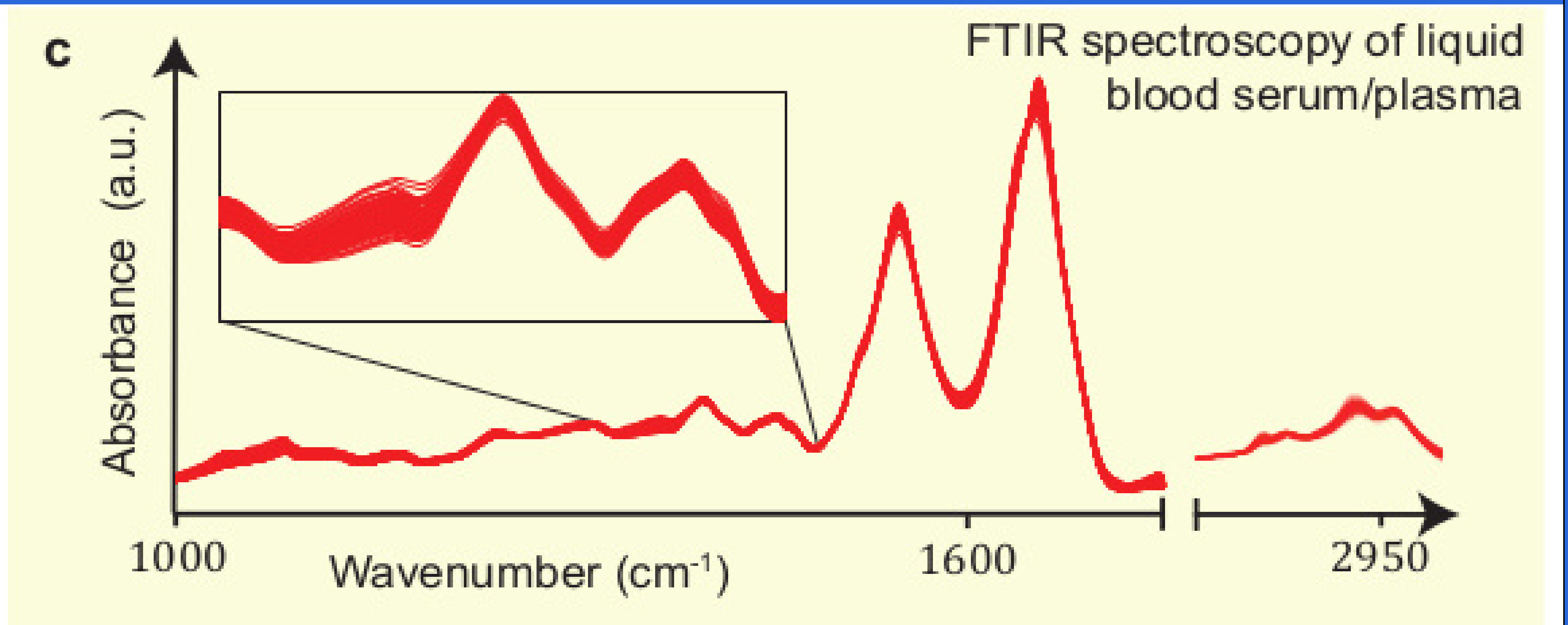
- Molecules exist as mixture in nature and not it their true form.
- e.g Human breathe contains hundreds of small VOCs (volatile organic compounds).
- VOC's represent different metabolic processes and carry vital information about diseases.
- Identifying certain biomarkers can lead to useful insights for clinical analysis.
- FTIR (Fourier Transform Infrared Spectroscopy) is a widely used method for determining if a chemical compound is present or absent.
- FTIR Spectroscopy gives absorption or emission spectrum of chemical molecules.

- Simultaneous identification of several molecules.
- FTIR Spectroscopy probes to vibrational characteristics.
- Traditional Method of Spectroscopy - Human Inspection.
- Disadvantage of Human Inspection- time-consuming and error-prone.
- IR spectral patterns overlapping leads to the spectral features losing their uniqueness.
- e.g Human breathe, concentration of target VOC is low, giving tiny spectral changes.
- Spectral envelops: produced by a superposition of several molecular species.
- Growing need for fast identification of chemical compounds => ML techniques for doing spectroscopic analysis.

FTIR SPECTROSCOPY



PROBLEM STATEMENT



What does our project solve?

- Test efficiency of ML algorithms in spectral envelope analysis.
- Generate synthetic data due to lack of availability of high resolution spectra.
- Two Problem Statements: 1) Classification IR spectra 2) Detecting presence or absence of molecule in envelope.

LITERATURE REVIEW

1. Identification of Chemical Structures from Infrared Spectra by Using Neural Networks - Applied Spectroscopy, 2001
2. Priority based functional group identification of organic molecules using machine learning - ACM, 2018
3. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models - ChemRxiv, 2021
4. Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples - Journal of biophotonics, 2013

Formulae for calculating absorption spectra:

$$I(\omega) = \sum_{k=1}^{N} \mu_k^2 F_k(E_k - \omega)$$

Ek: Influences location
uk: Influences amplitude

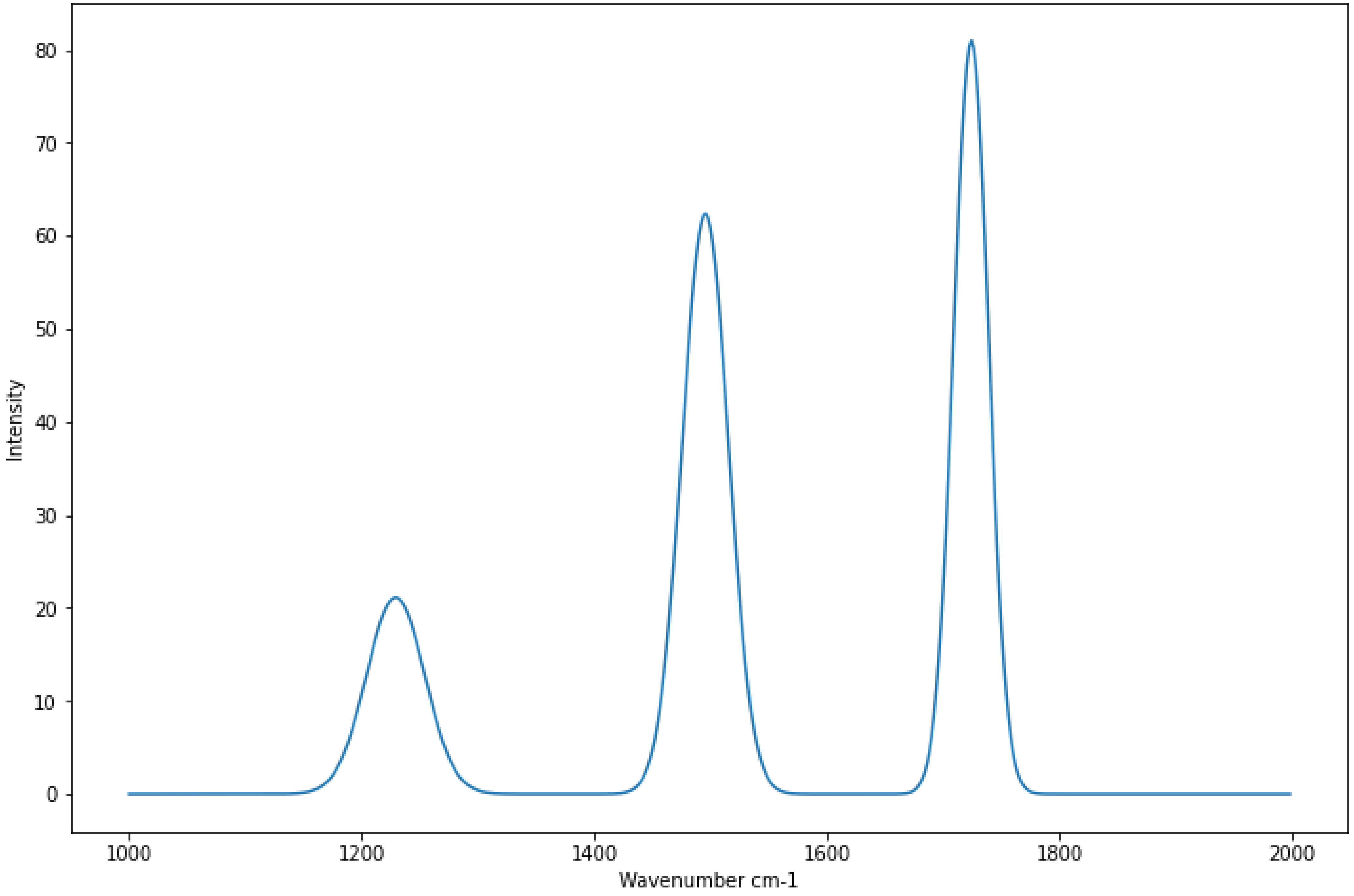
Fk is gaussian line-shape function, which is given by:

$$F_k(\omega) = e^{\frac{-\omega^2}{2\sigma_k^2}}$$

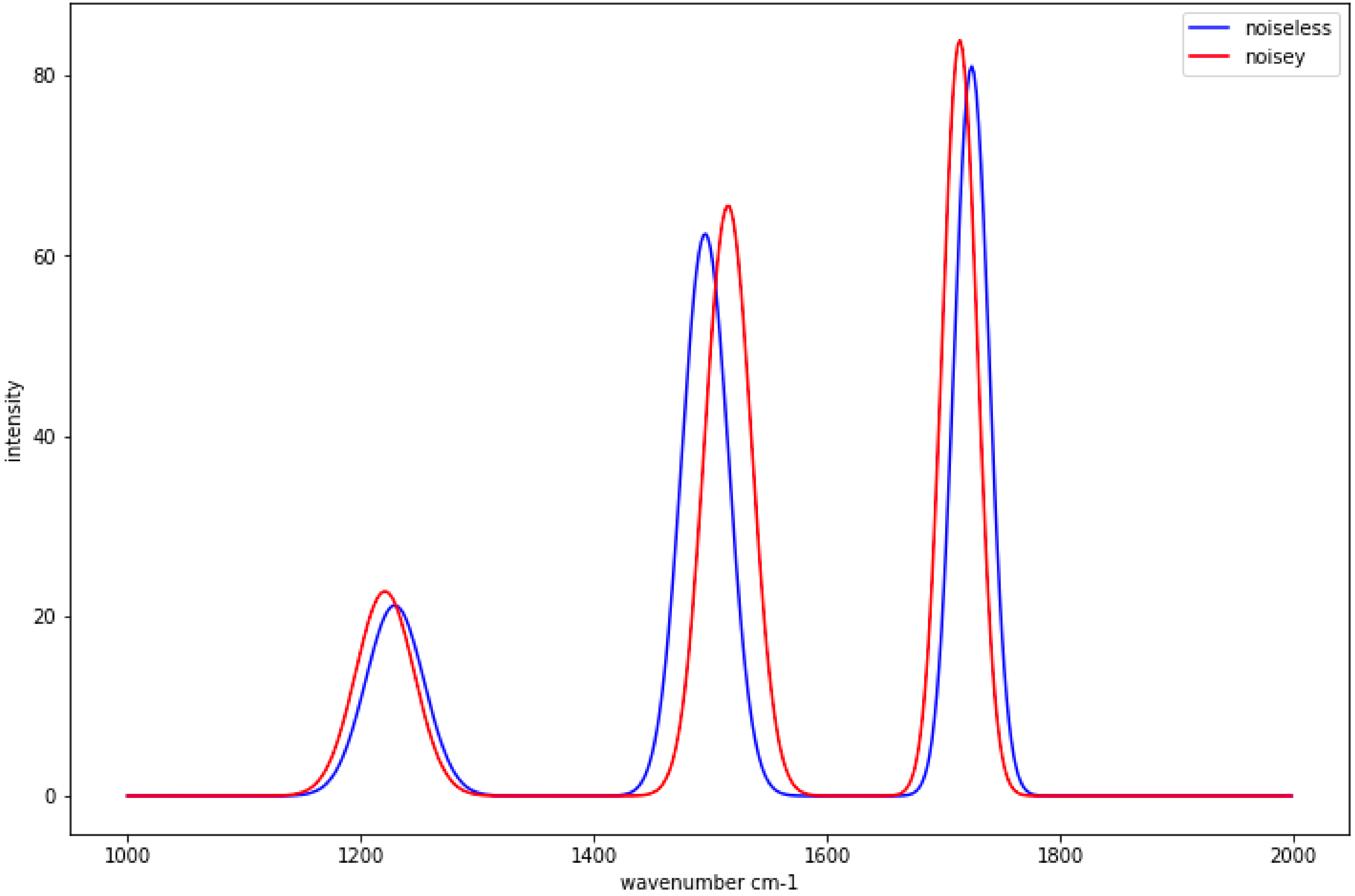
So our data generation model is thus defined by parameters:

$$(E_k, \mu_k), \quad k = 1, 2, 3 \ldots N$$

DATA GENERATION



Noise in Parameters



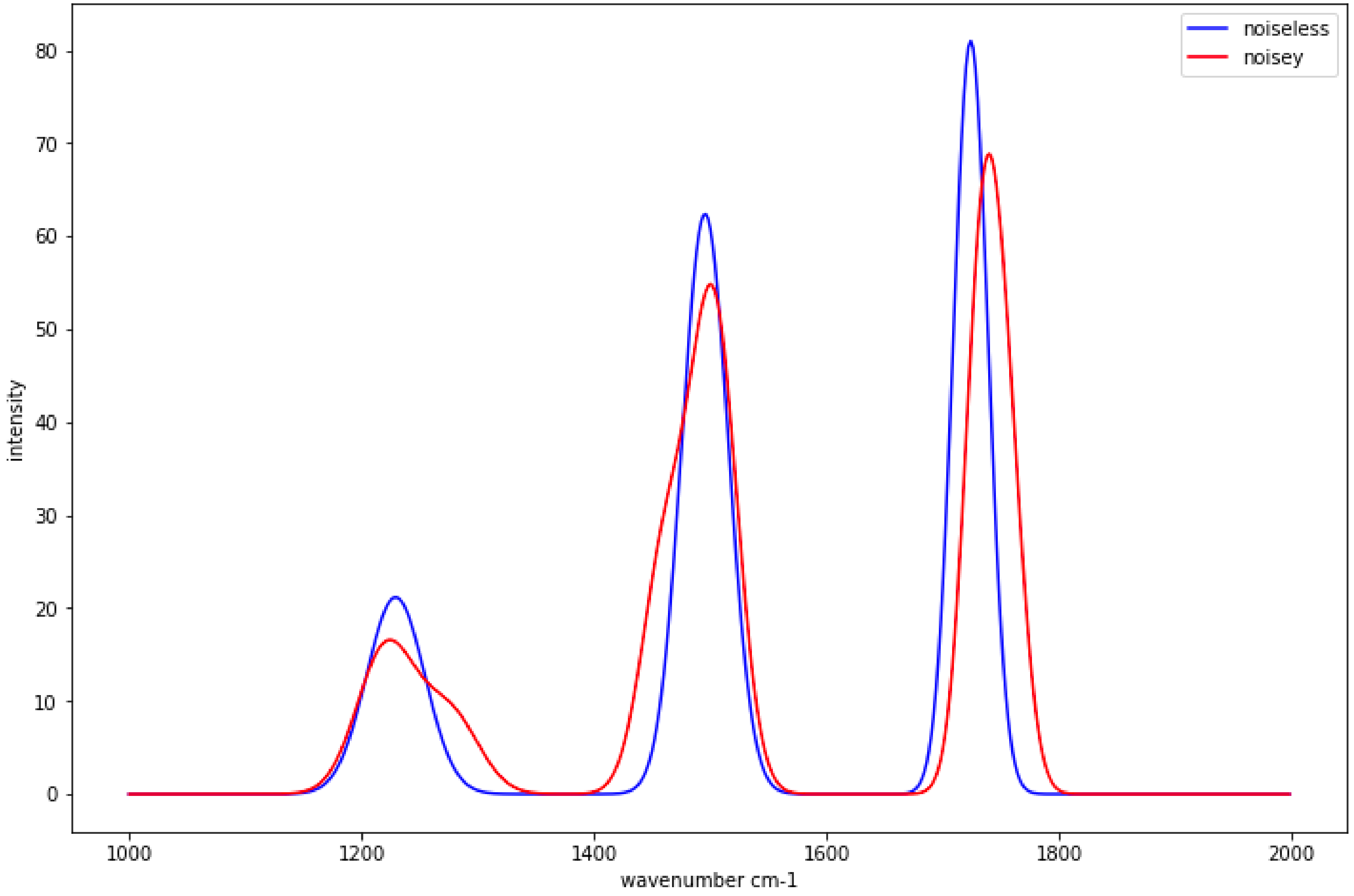
$$E_n^{(a)} \rightarrow E - \delta_n^{(a)}$$

Where δn is a random number whose absolute value is in the range δn ≤ 20.

$$\mu_n^{(a)} \rightarrow \mu \cdot (1 + \delta_n^{(a)})$$

Here δn is a random number whose absolute value is in the range δn ≤ 5% of μ.

Asymmetry in Peaks



Introducing of asymmetry in peaks.

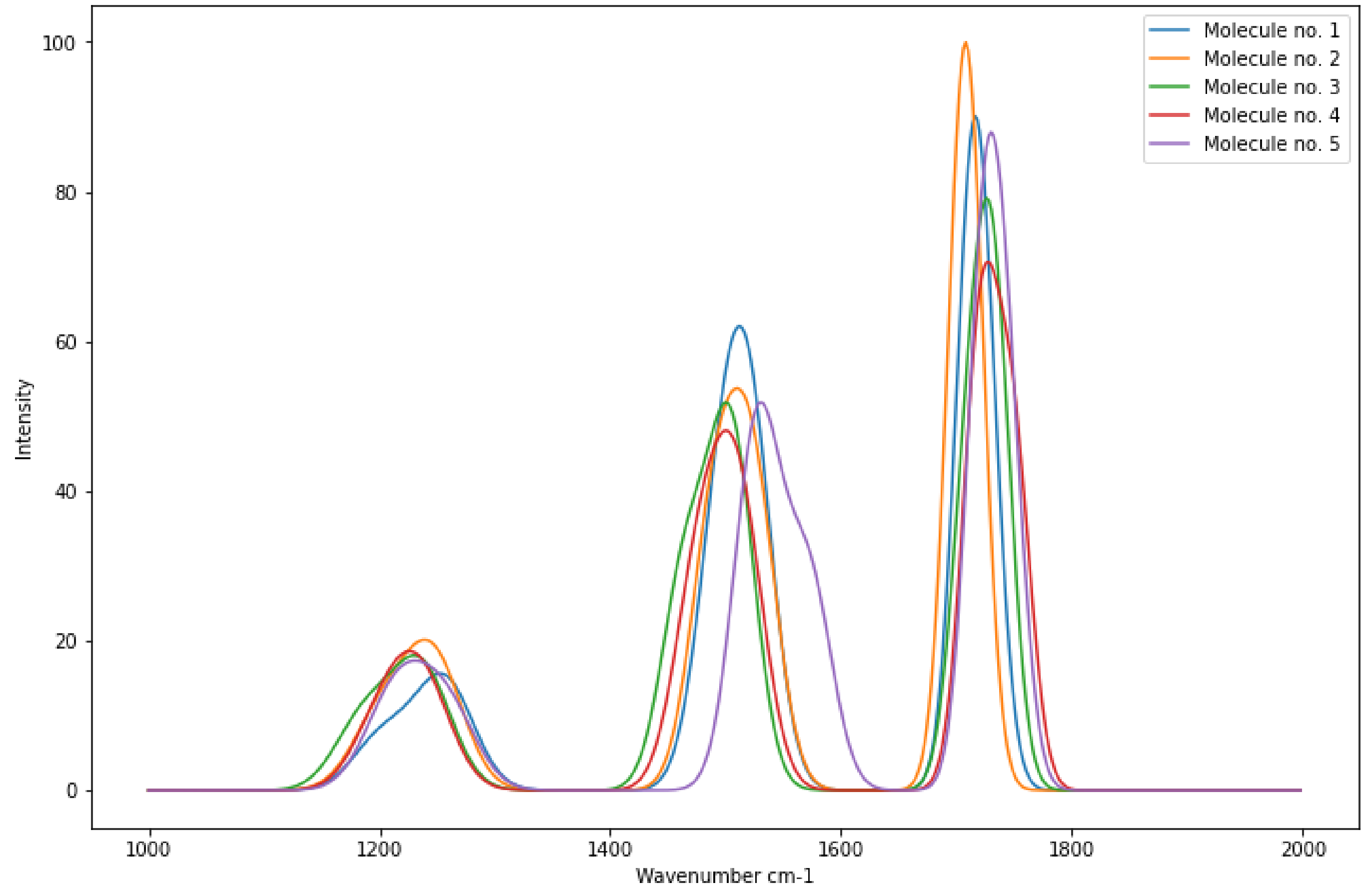
Adding left and right shifted version of the spectrum with the actual spectra, but overall shift will not exceed some value T, mathematically speaking.

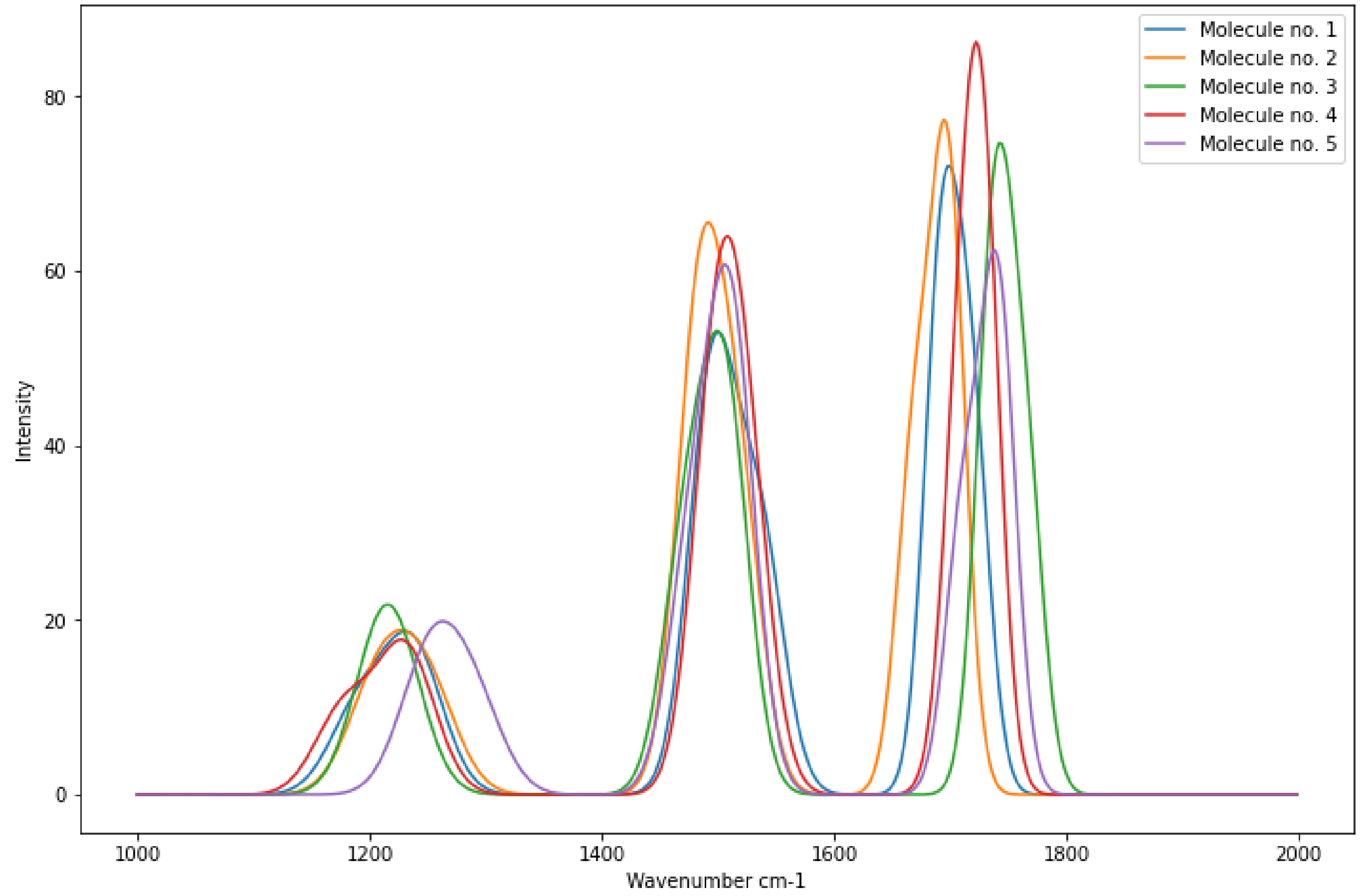
$$I(\omega) = A_l \cdot I(\omega + S_l) + I(\omega) + A_r \cdot I(\omega + S_r) \div 3$$

where Sl + Sr ≤ T

FINAL SPECTRA

- Five molecules.
- 2500 IR spectra each molecules.
- Wave-numbers: 1000-2000 cm-1, at a step of 2.
- Spectra dimension: 500x1





ENVELOPE GENERATION

Envelope: Resultant spectrum of different combinations of molecules

$$\left(\sum_{x}^{X} d^{x} \right) \div M$$

- M is the number of molecules used to form the envelope
- X is a combination of molecules used
- x is used to iterate on a specific combination

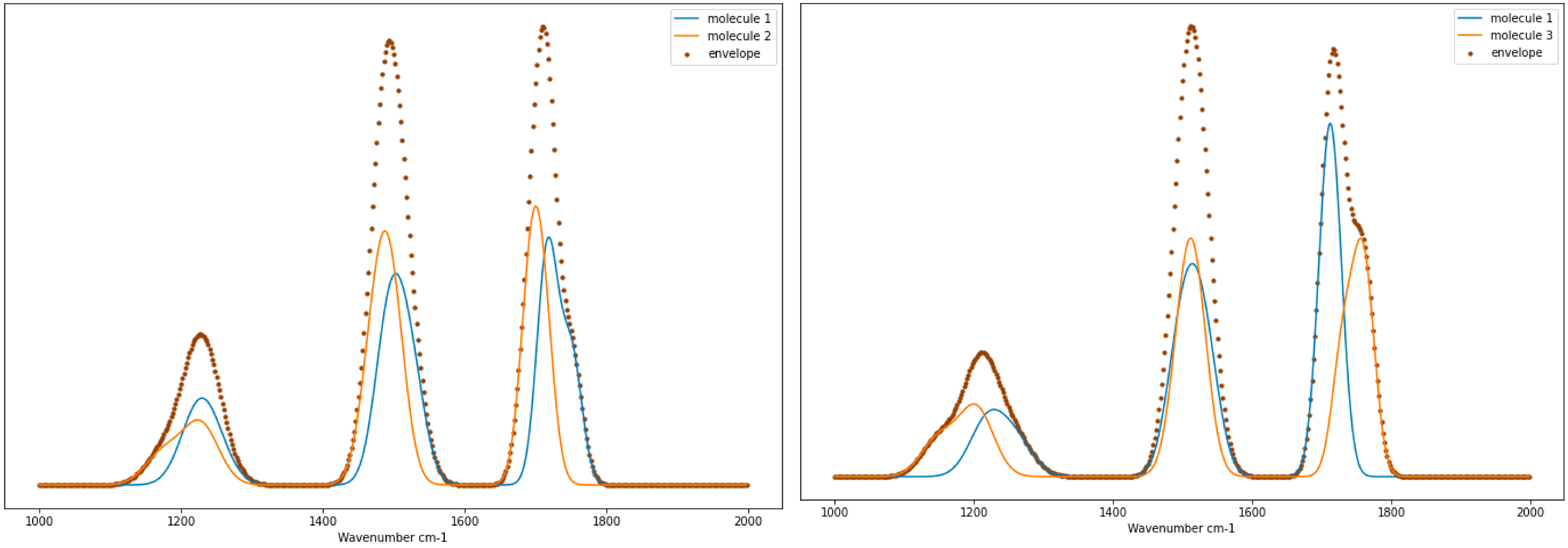
Labelling of Envelopes

Molecules used	Label
3,4	0
1,4,5	0
2, 3, 4	1
1, 5	0
1, 2, 3, 5	1
1 ,2, 3, 4, 5	1

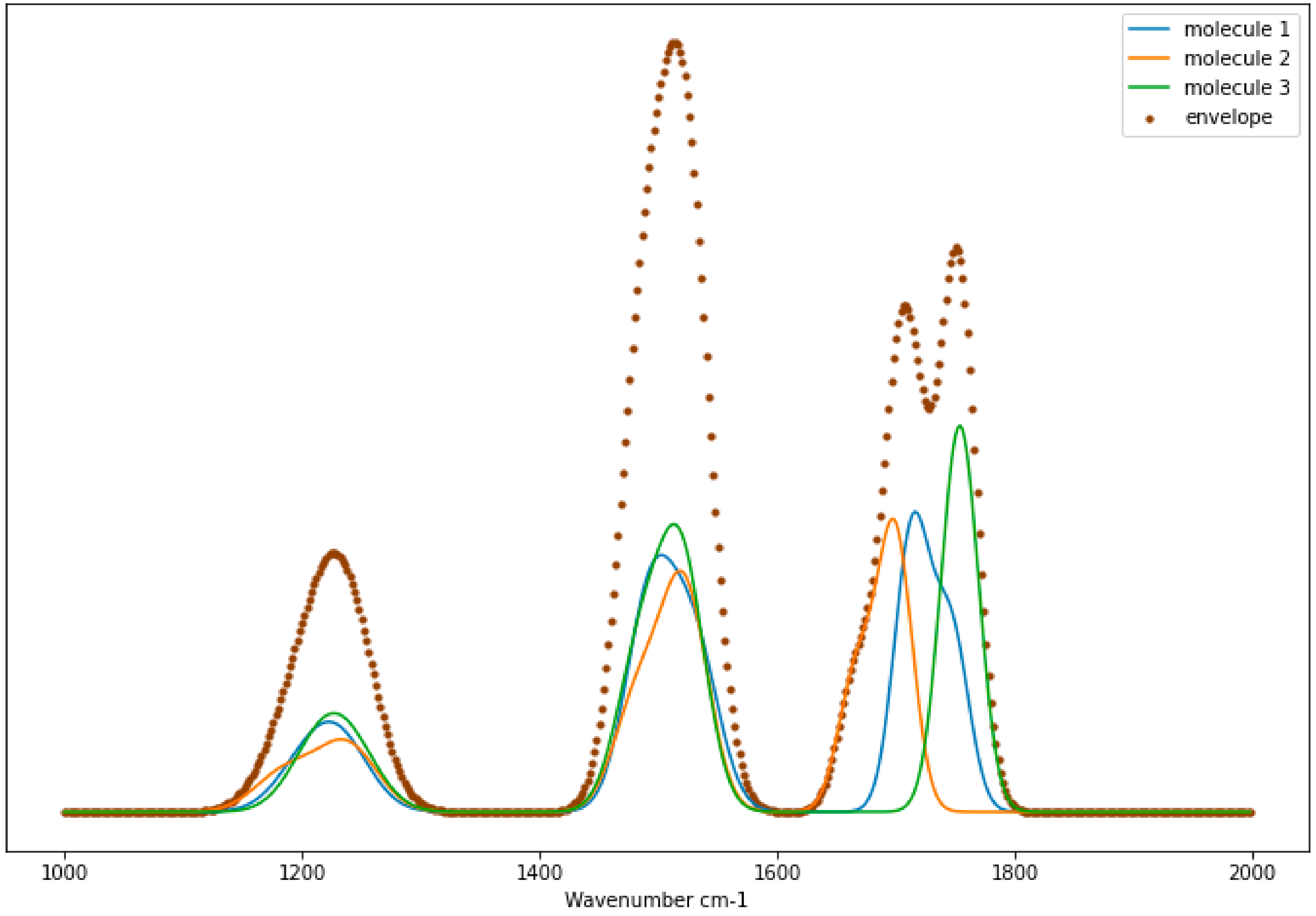
The set of molecules containing molecule 2 is labeled 1 and the rest as 0.

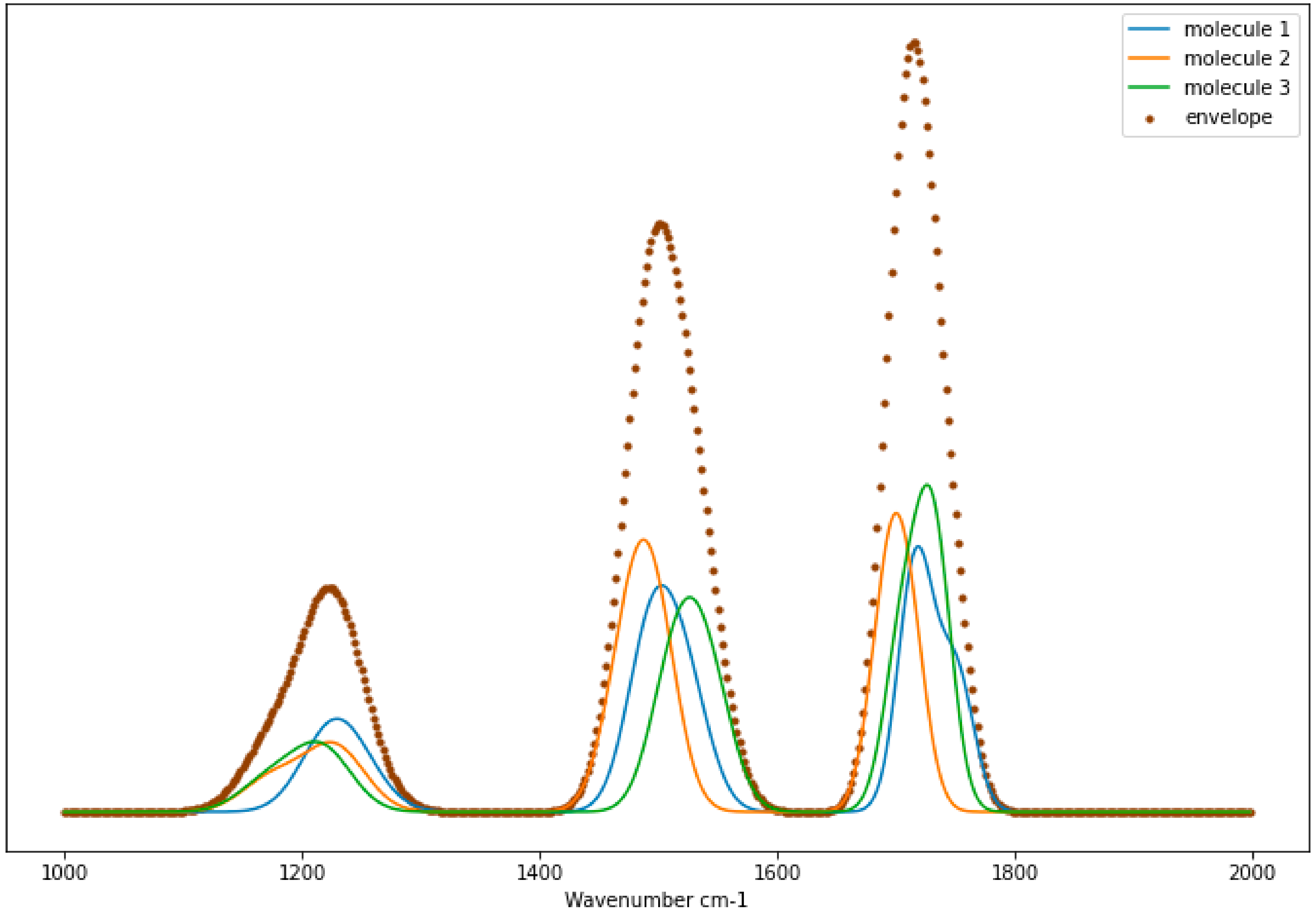
EXAMPLE ENVELOPES

Envelope of two molecules

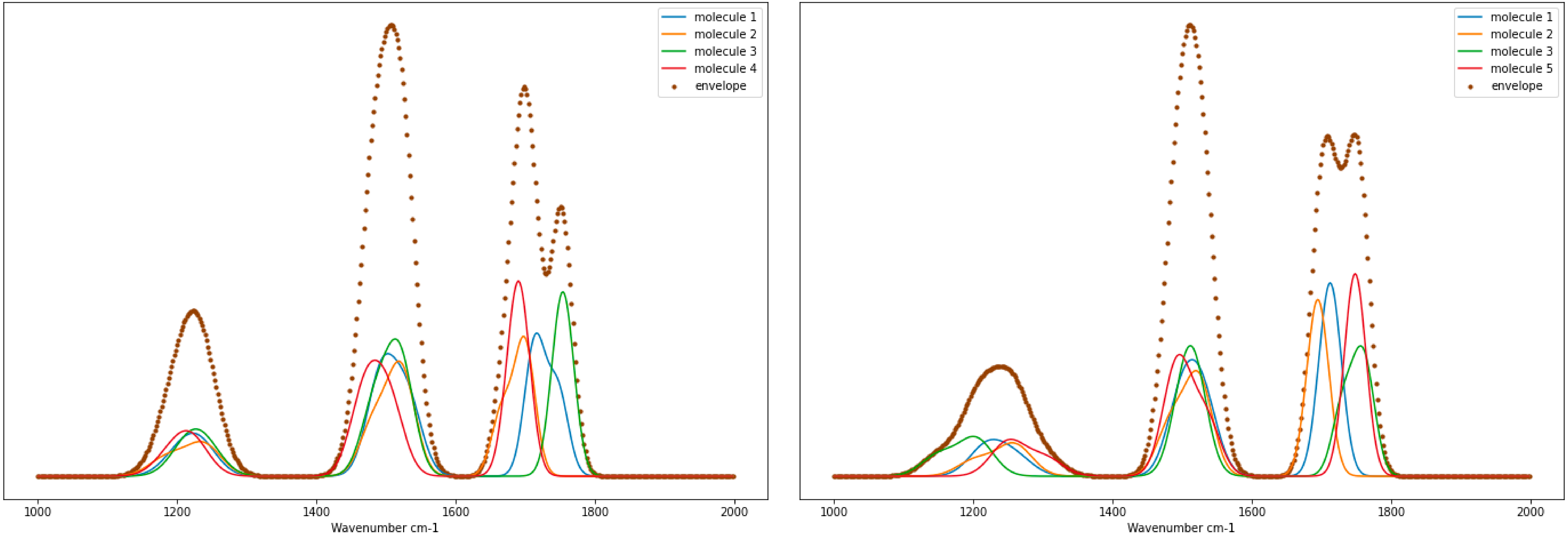


Envelope of three molecules

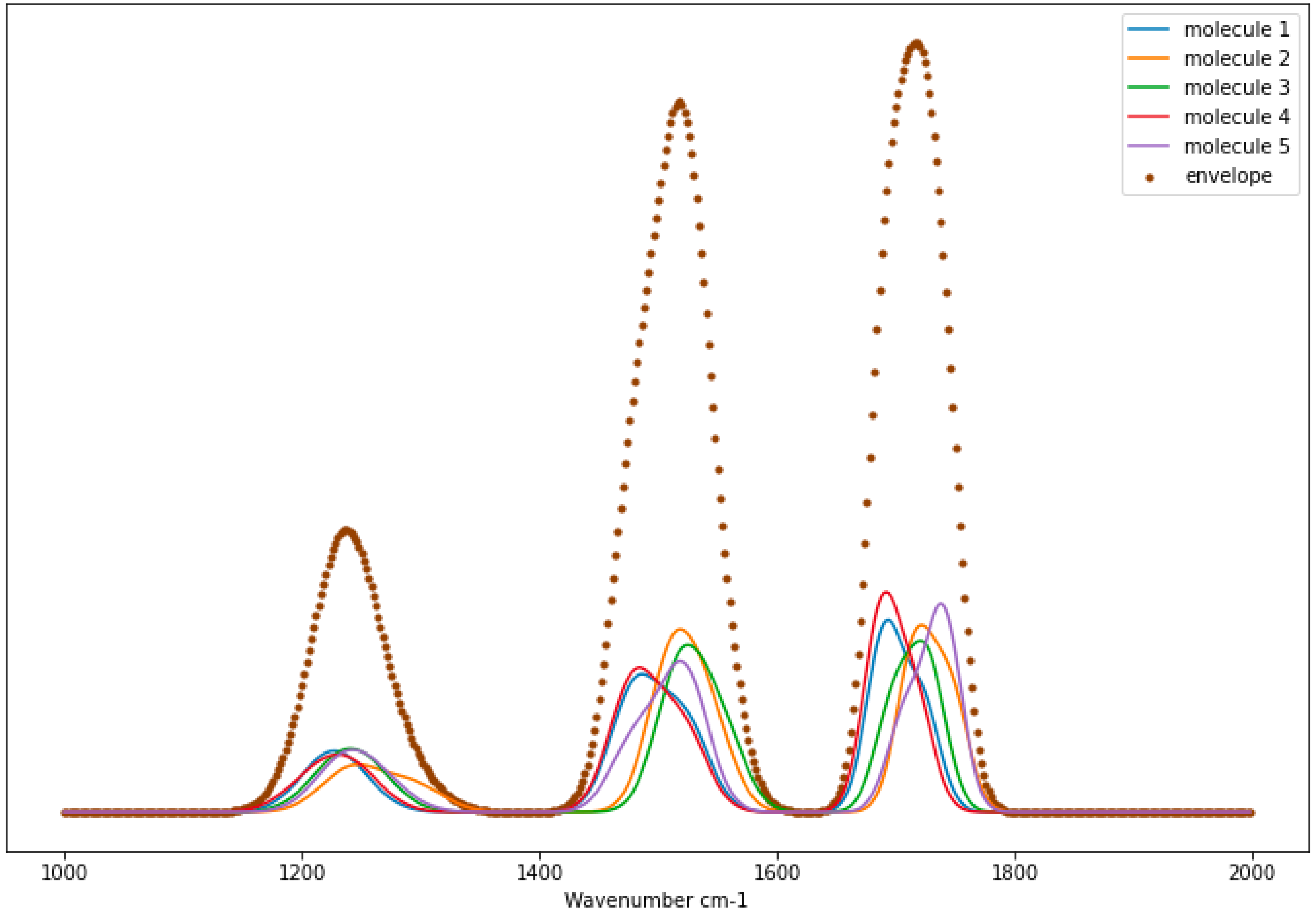


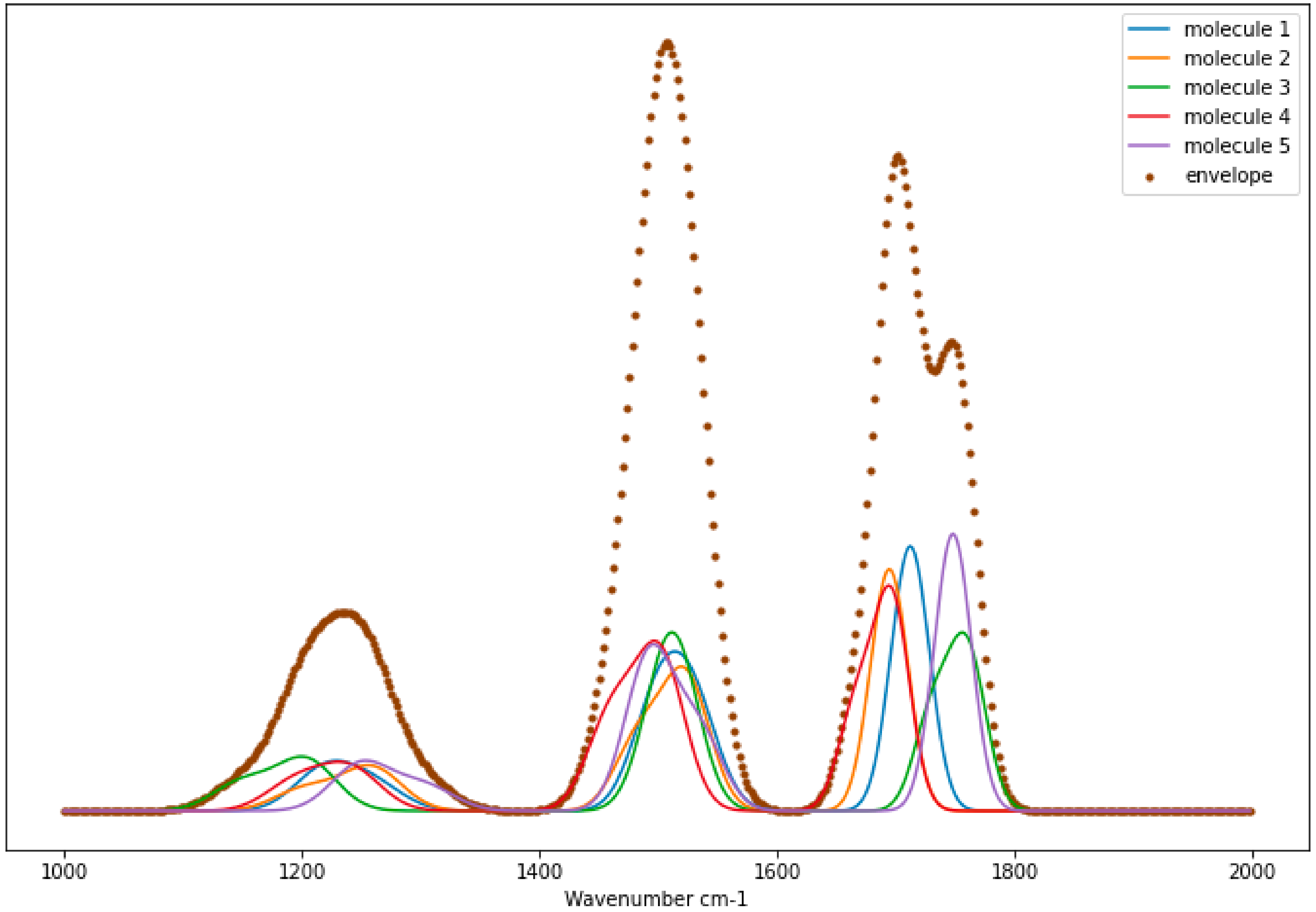


Envelope of four molecules

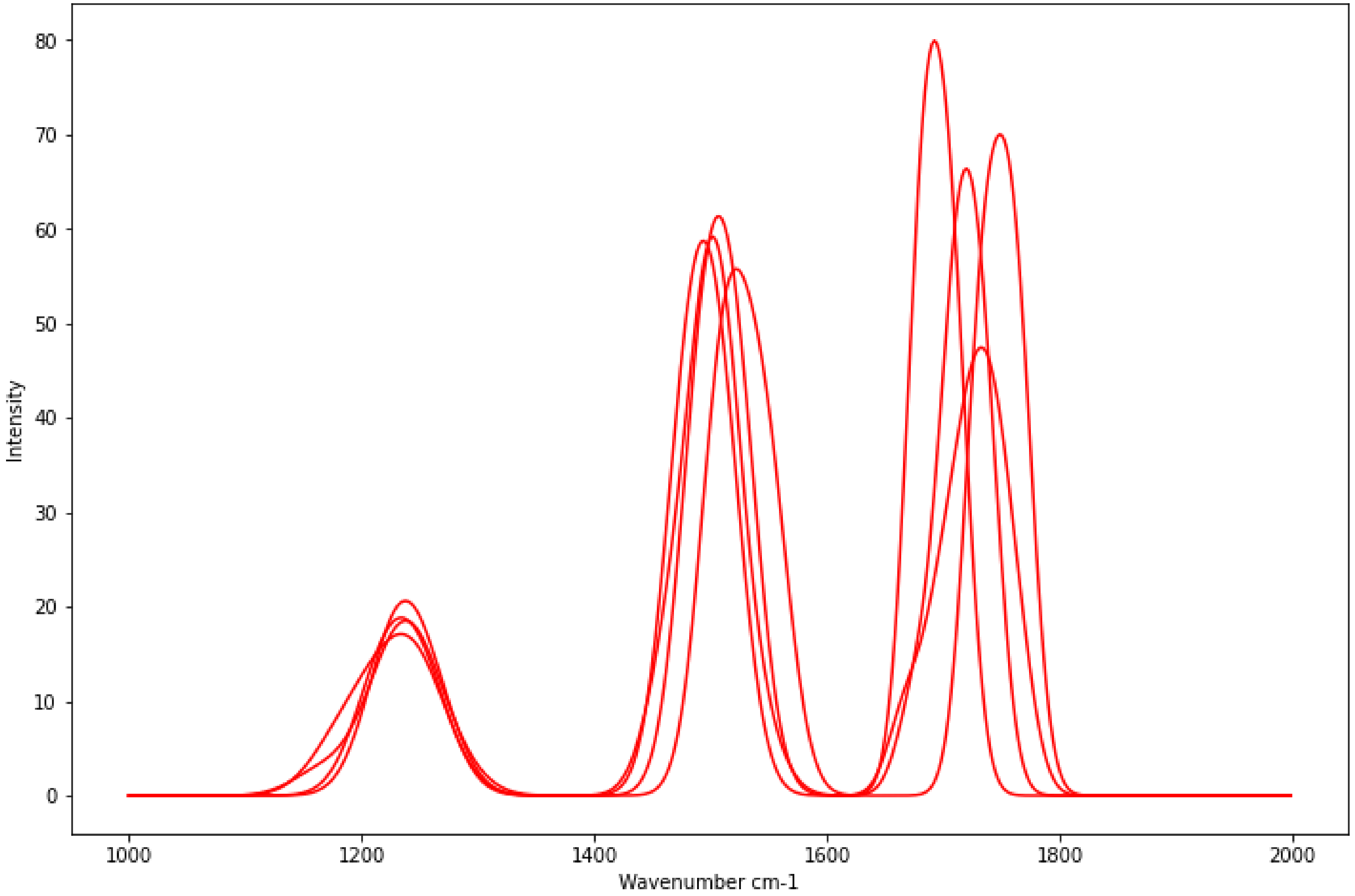


Envelope of five molecules

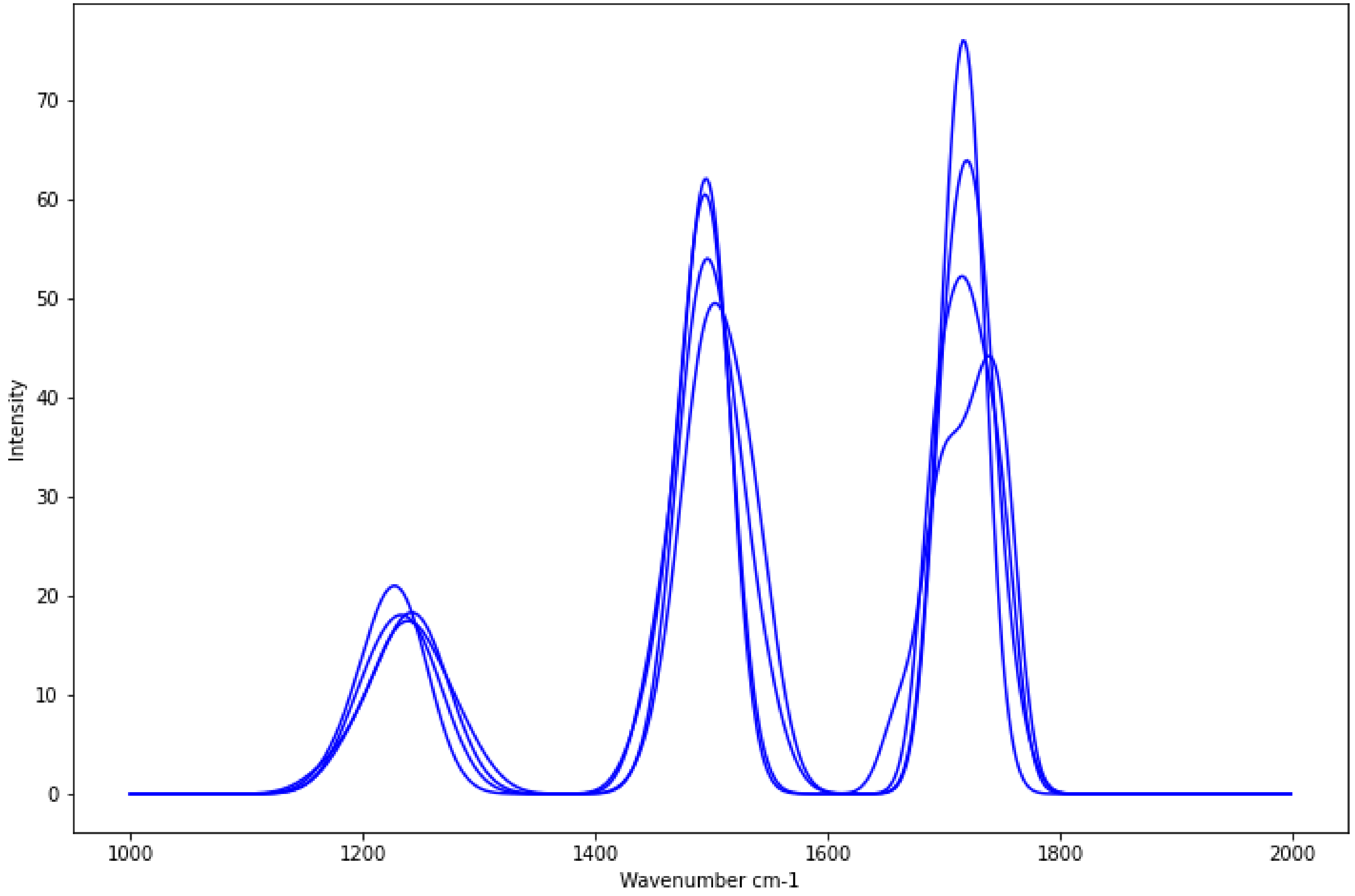




Label 1



Label 0



Computational Methodology

- Model a function $f : X \to Y$.
- For all of our experiments we consider X_i as the IR spectra/envelope and Y_i as the output of the model.
- Output of the model varies as per the experiment conducted
- Project Divided into two phases.

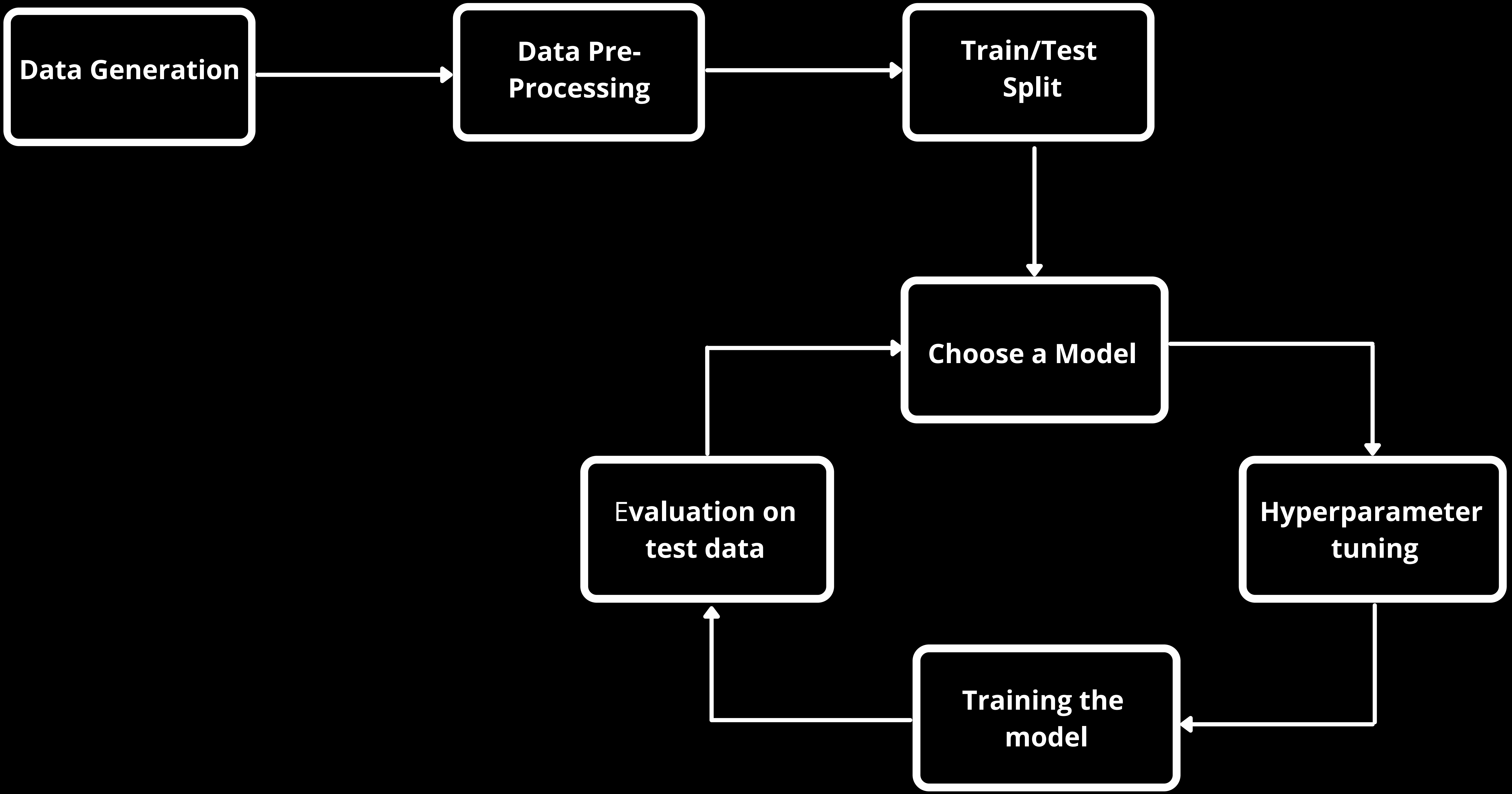
Computational Methodology

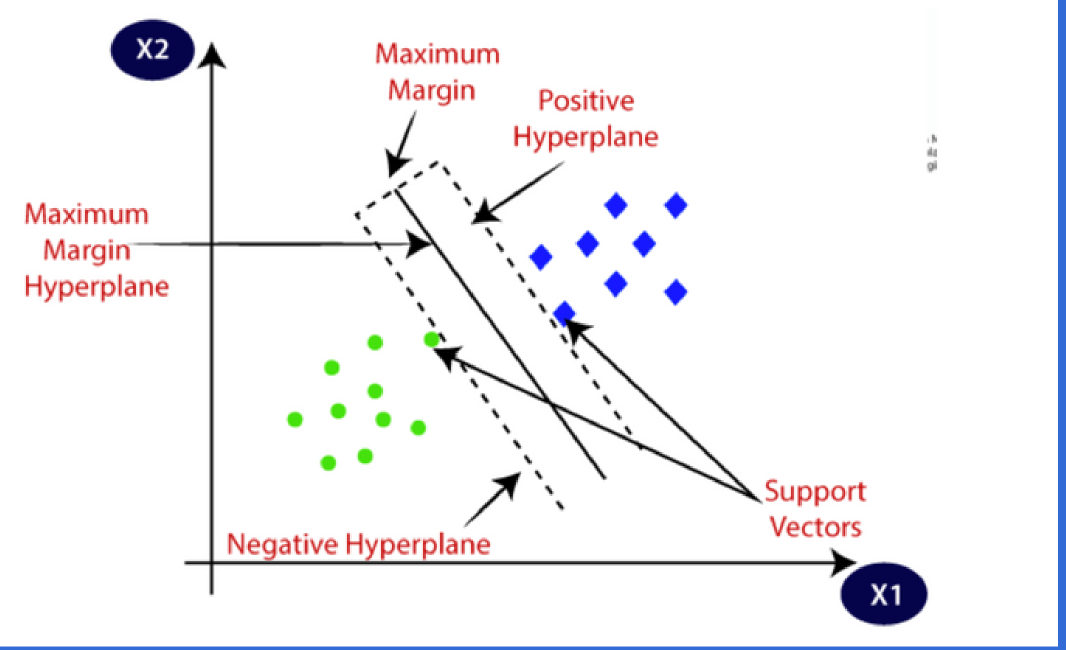
Phase-I

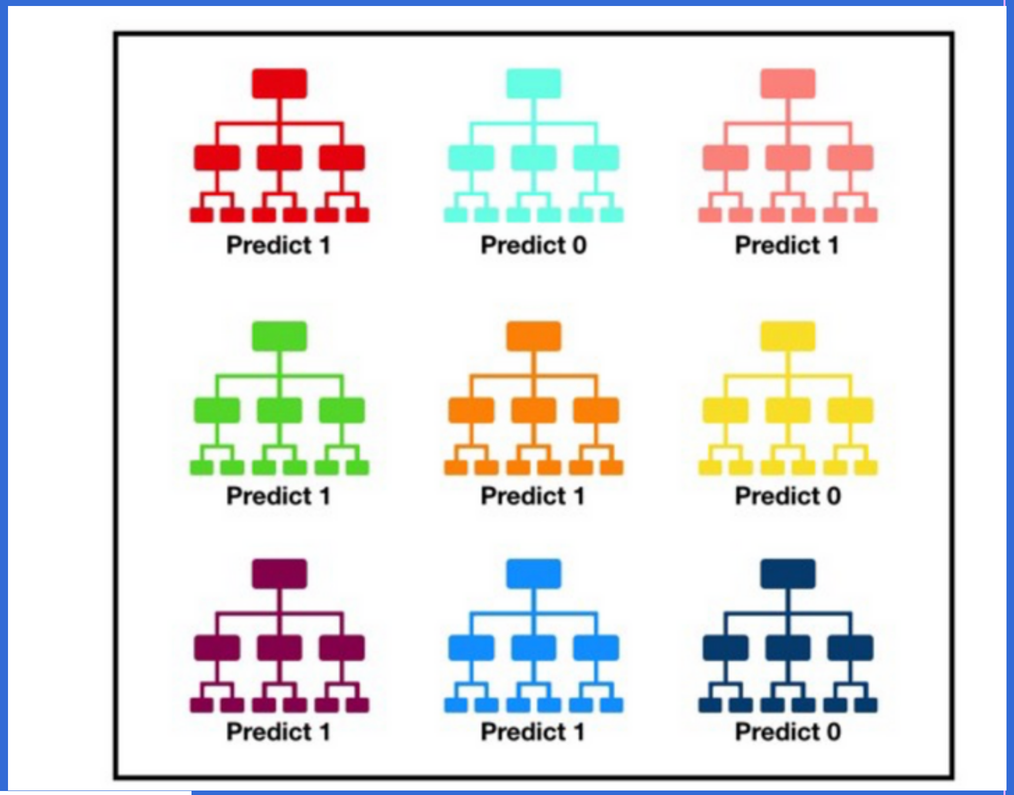
- X is IR spectra
- Y is one of the molecular class M(0,1,2,3,4).
- Worked on classification of IR spectra of pure chemical molecules

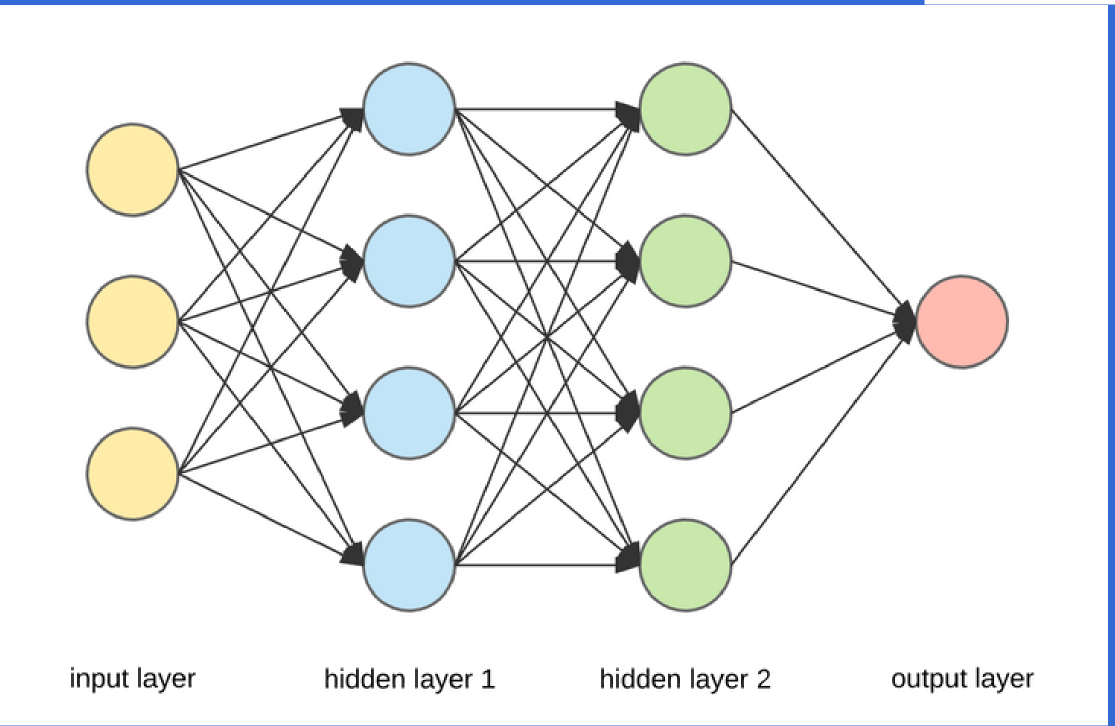
Phase-2

- X is spectral envelope
- Y is a binary number B(1 or 0)
- Worked on identifying presence or absence of molecule 2, in an envelope.


Data Generation
Data Pre-Processing
Train/Test Split
Choose a Model
Hyperparameter tuning
Training the model
Evaluation on test data







ML MODELS

Support Vector Machine

Random Forest Classifier

Multilayer Perceptrons 1

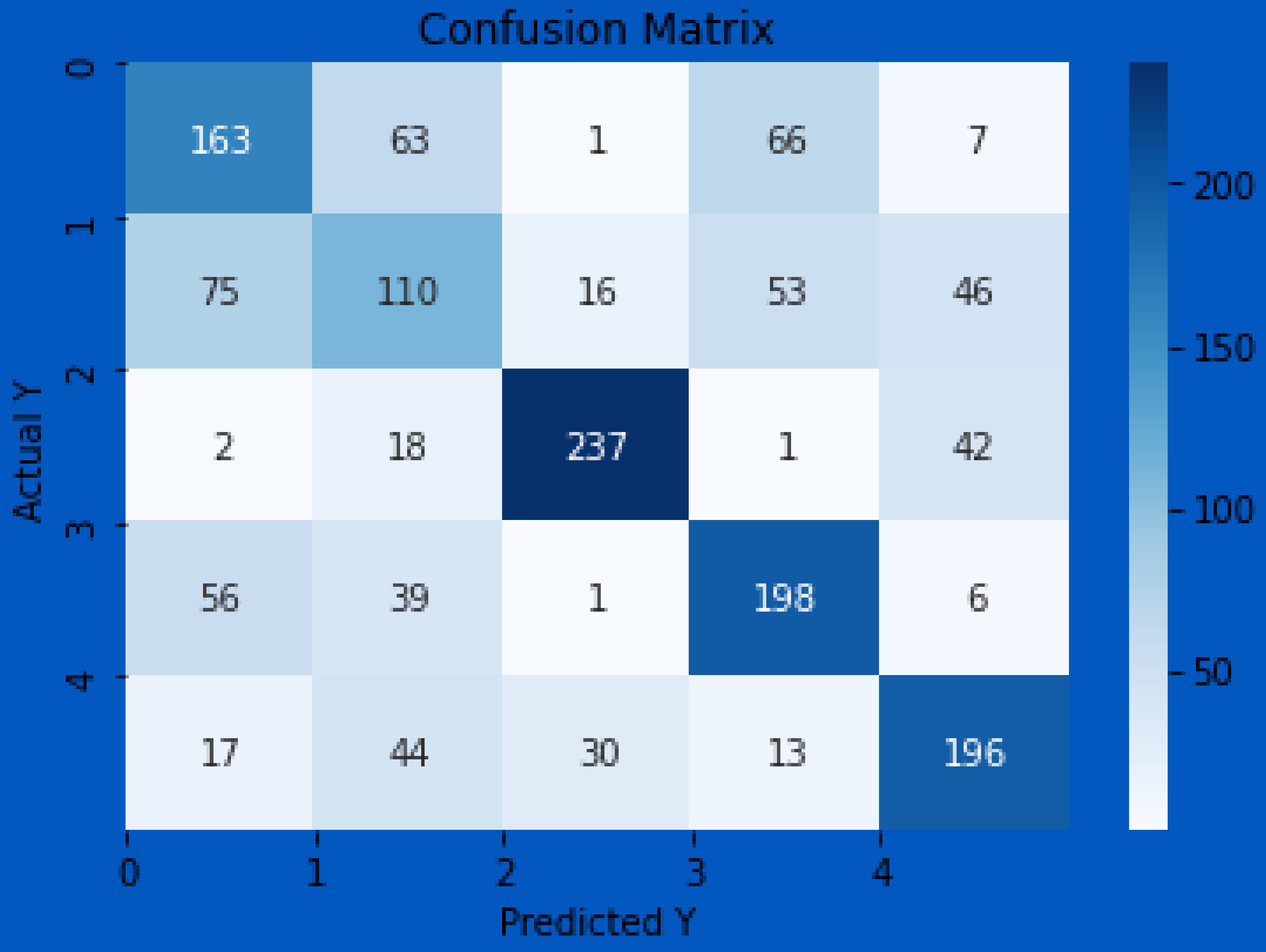
Multilayer Perceptrons 2

PHASE1 RESULTS

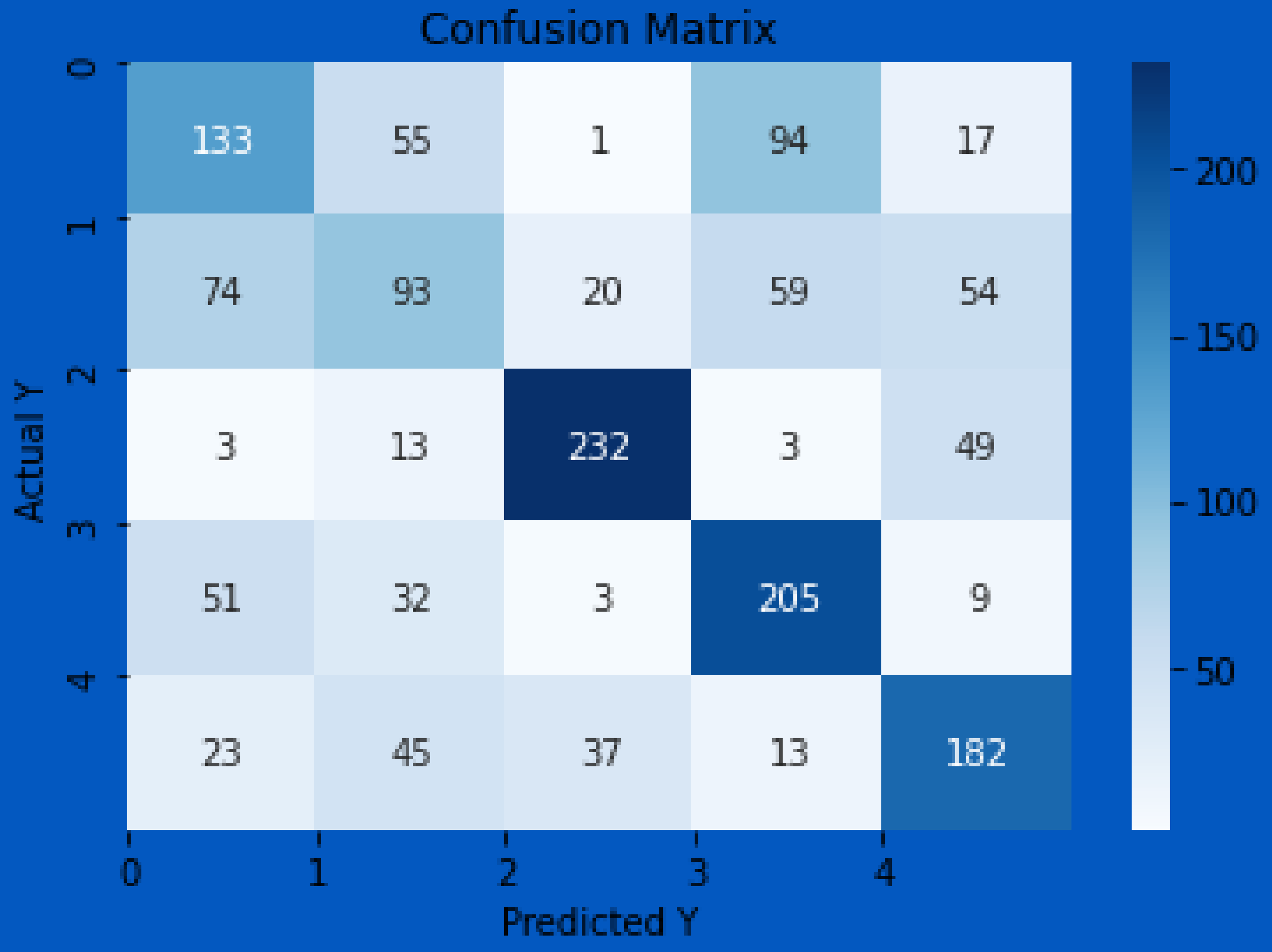
MODELS	TRAIN	TEST
Support Vector Machine	65%	60%
Random Forest Classifier	100%	56%
Multilayer Perceptrons 1	74%	58%
Multilayer Perceptrons 2	80%	56%

Confusion matrices

SVM

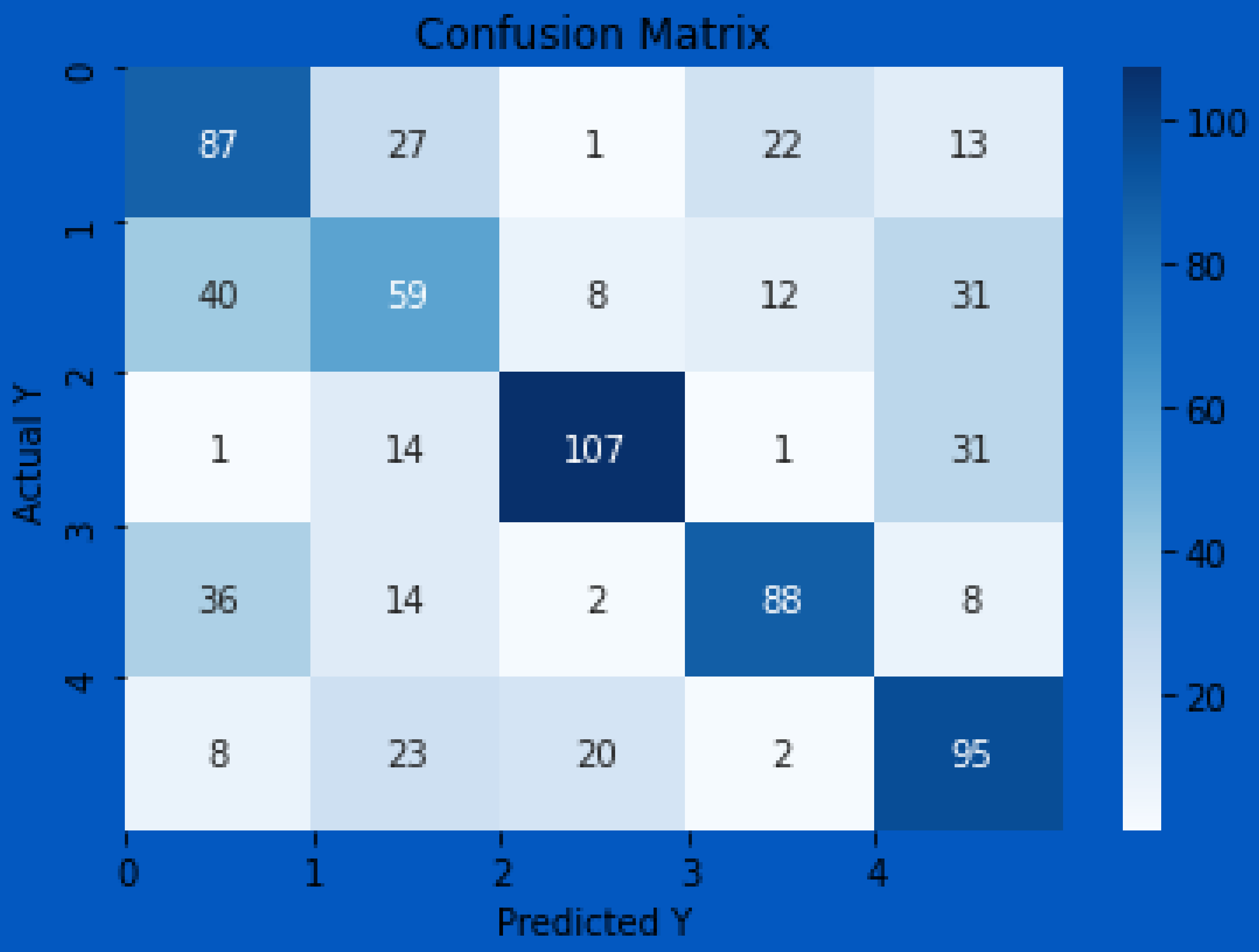


RFC

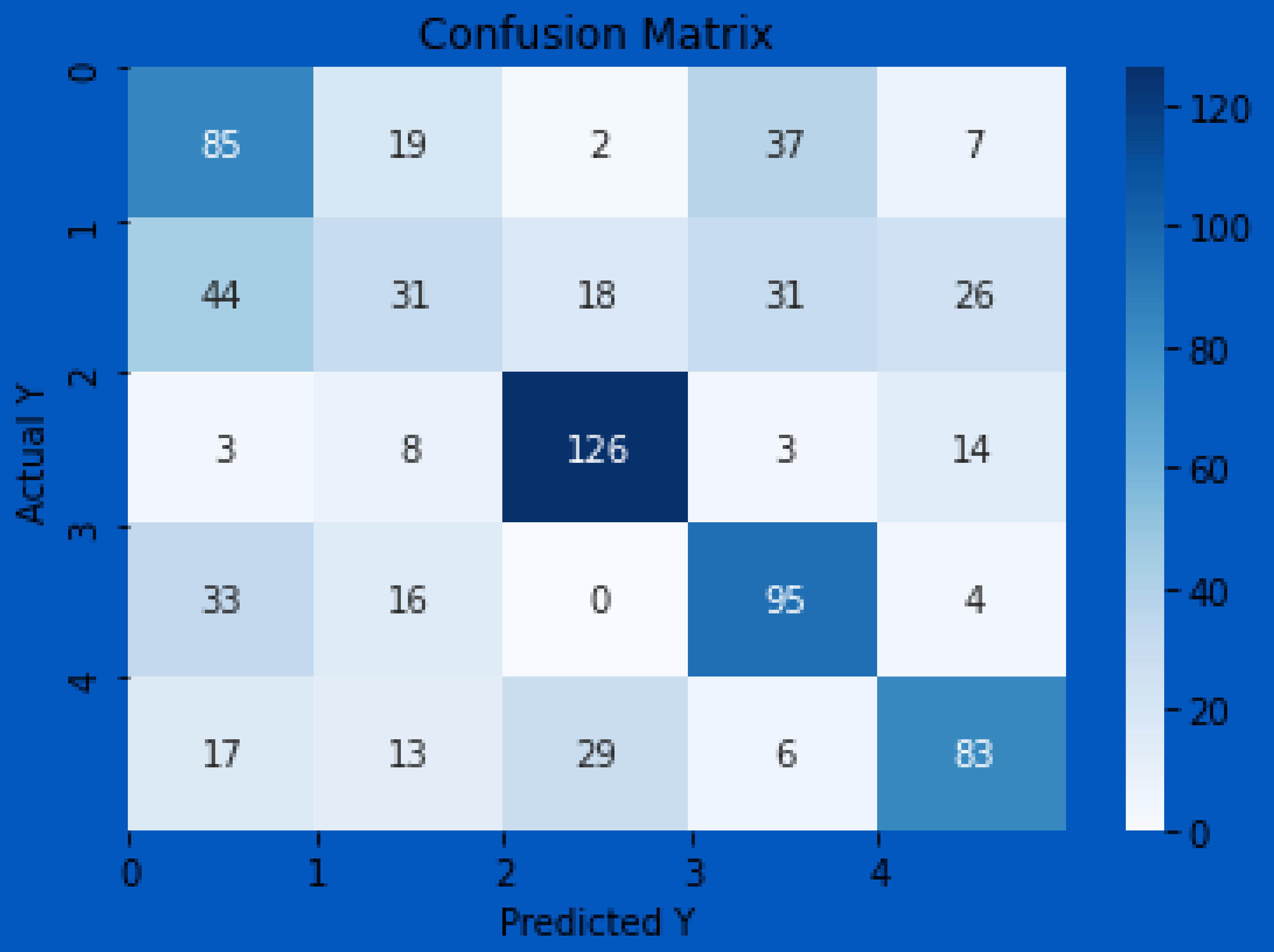


Confusion matrices

NN1



NN2

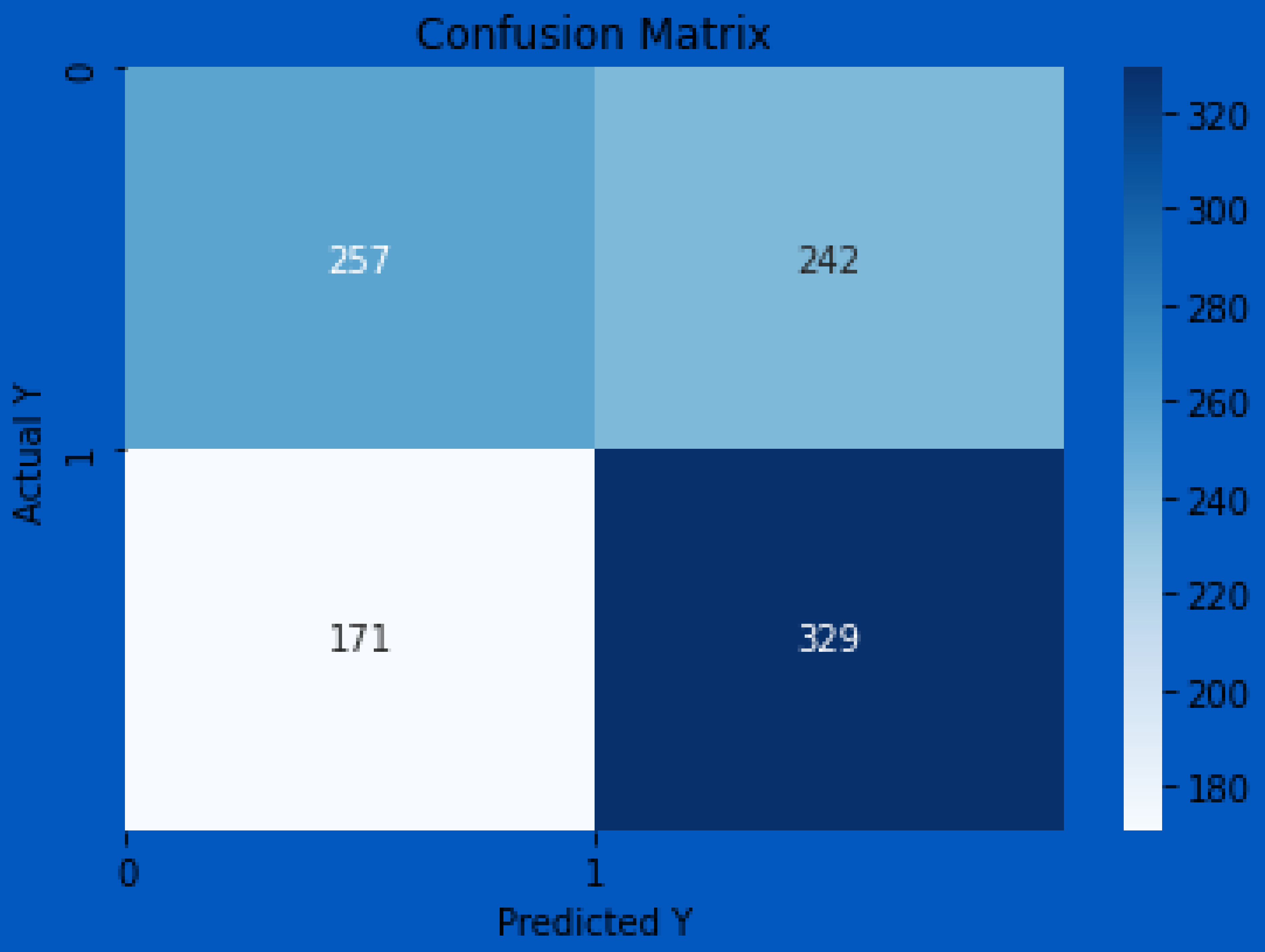


PHASE2 RESULTS

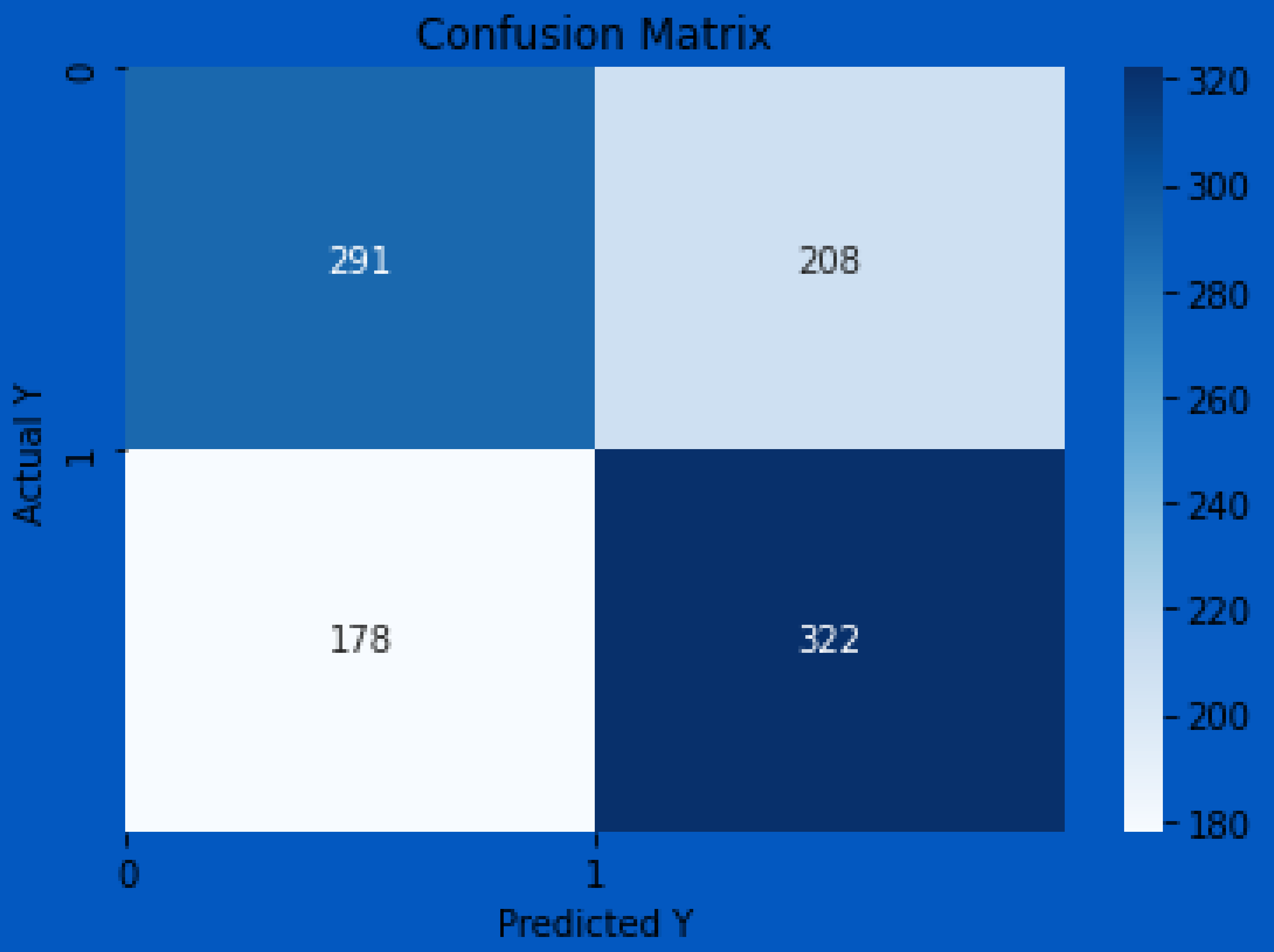
MODELS	TRAIN	TEST
Support Vector Machine	63%	59%
Random Forest Classifier	96%	61%
Multilayer Perceptrons 1	72%	57%
Multilayer Perceptrons 2	76%	58%

Confusion matrices

SVM

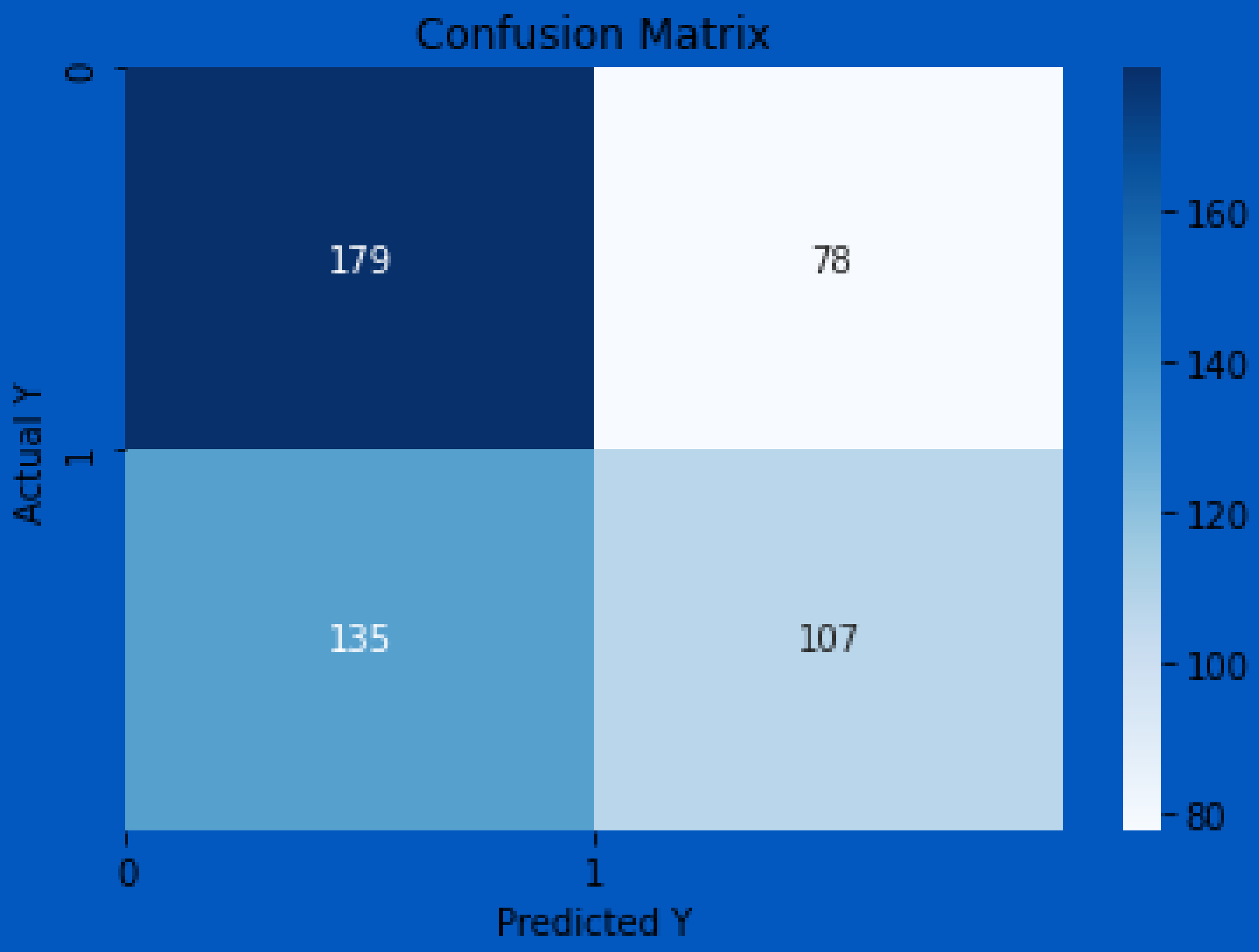


RFC

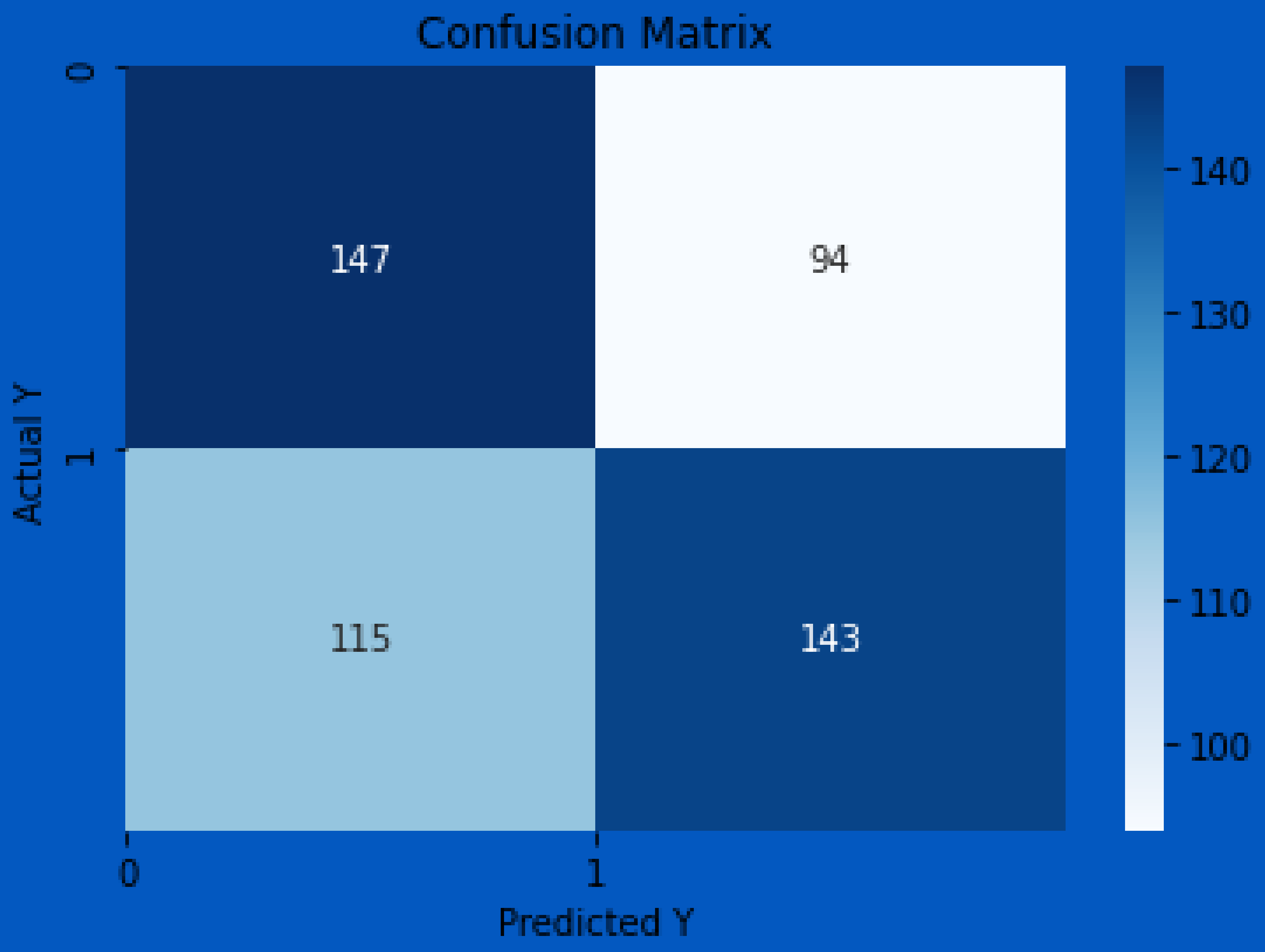


Confusion matrices

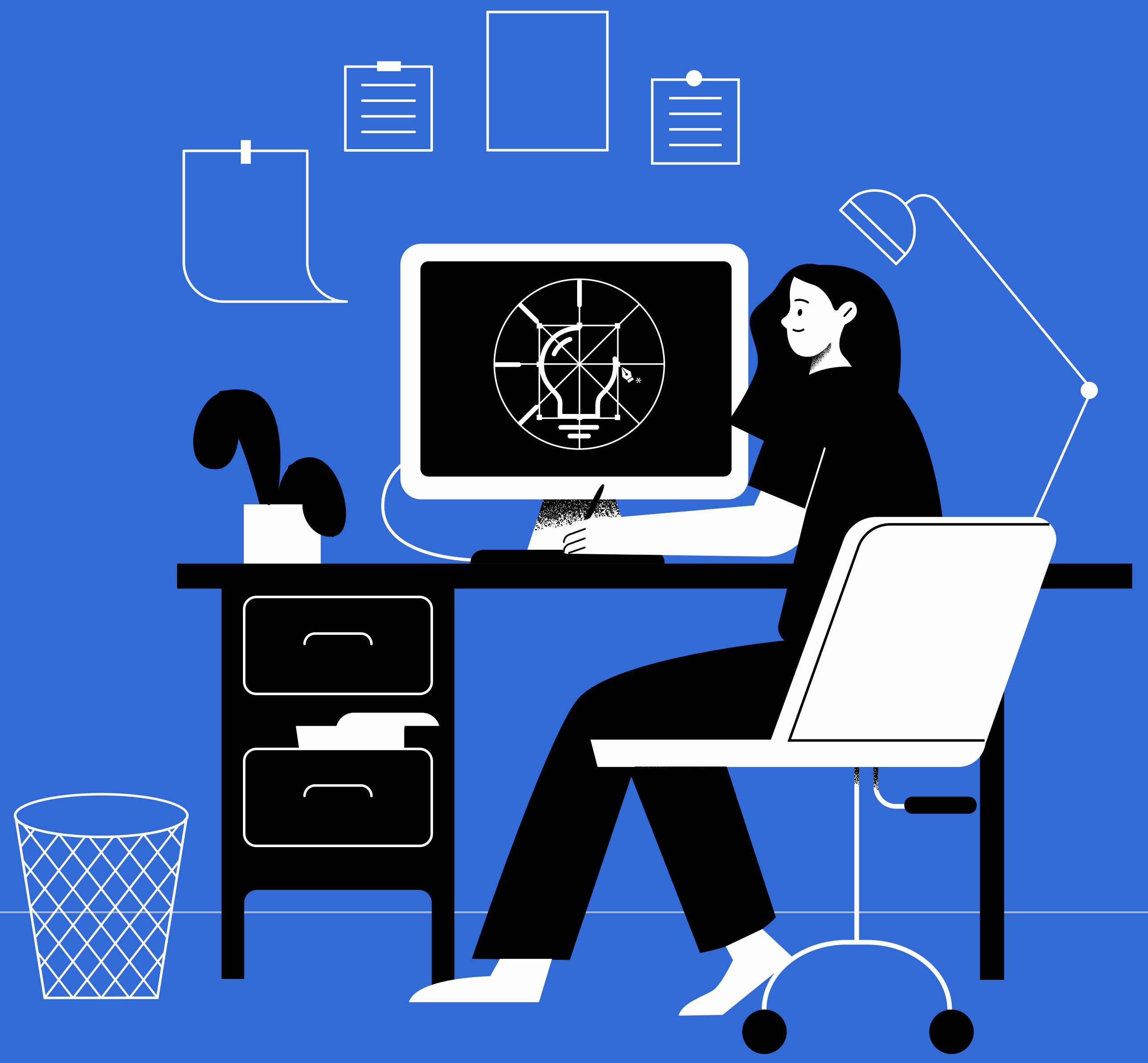
NN1



NN2



PHASE2 RESULTS



MODELS	Envelope	FTIR
Support Vector Machine	59%	69%
Random Forest Classifier	61%	68%
Multilayer Perceptrons 1	57%	68%
Multilayer Perceptrons 2	58%	72%

Comparison of FITR-based and Envelope-based molecule 2 classification

ACKNOWLEDGEMENTS

We would like to express our gratitude and appreciation to Dr Bhaskar Choudhary who gave us the opportunity to do this research under him. Special thanks is due to Prof. Maxim Gelin and Prof. Kiran Maiti (Max-Planck-Institut fu¨r Quantenoptik, Germany) without whom this research would not have been possible, the discussions with them led to a healthy conclusion to our project.

REFERENCES

- https://chemrxiv.org/engage/chemrxiv/article-details/60c75603469df44d0df45250
- https://onlinelibrary.wiley.com/doi/10.1002/jbio.201300149
- https://dl.acm.org/doi/10.1145/3152494.3152522
- https://journals.sagepub.com/doi/10.1366/0003702011953531
- https://opg.optica.org/ao/abstract.cfm?uri=ao-60-14-4217
- https://pubs.acs.org/doi/10.1021/jp052324l

THANK YOU !!