Machine Learning for the Diatomic Molecular Spectroscopy Database: Gaussian process regression for predicting spectroscopic constants

Daniel Julian, Ethan Franco, Jesús Pérez-Ríos
Department of Physics and Astronomy, Stony Brook University

Introduction

The Diatomic Molecular Spectroscopy Database (DMSD) is a website that contains the spectroscopic constants of diatomic molecules. A new version of this website is being developed under the guidance of Dr. Jesús Pérez-Ríos in the theoretical AMO department. Unlike the previous iteration of the DMSD, the new version will host on-the-fly machine learning (ML) capabilities. With the new DMSD, users will be able to retrieve spectroscopic data for molecules already in the database and will be able to make predictions for molecules not in the database.

Spectroscopic constants refer to the measurable aspects of diatomic molecules found using spectroscopy. In spectroscopy, the light absorbed or emitted by atoms and molecules is used to measure the energy transitions between their quantum states. From their spectra, it is possible to determine the spectroscopic constants of molecules. This project will predict the values of these constants using an ML approach.

Methods

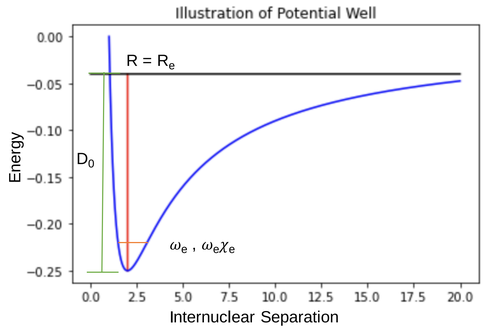
To predict spectroscopic constants, regression is the obvious choice for the ML model type. Gaussian Process Regression (GPR) is the model chosen for this project. GPR is an ML model that assumes the targets and data follow a specified distribution (typically multivariate gaussian). GPR doesn't assume a prior functional form of the target's dependence on the features. Two notable aspects of GPR are that it's a kernelized algorithm and it learns the mean and uncertainty of the target as a function of the features. GPR works well on small data sets, making it a good choice for this project. GPR uses kernel functions to compute inner products; the result of kernelization is that ML models can learn the target as a function of a high dimensional (or infinite dimensional) feature space, making it possible to find a linear dependance between a new higher dimensional feature vector and the target. Using, training data, kernelization, and hyperparameter optimization, GPR learns the likely distribution of target values. The kernel functions used in this project are:

$$k_{\text{Matern}}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} d\right)^{\nu} K_{\nu} \left(\sqrt{2\nu} d\right) \quad [1]$$

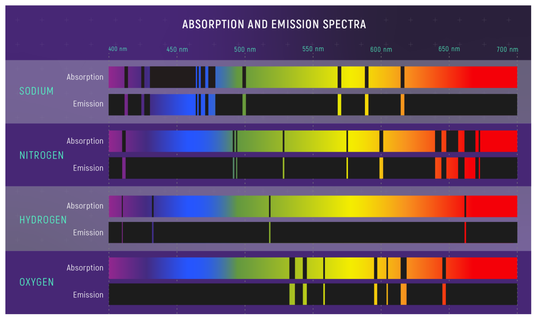
$$k_{\text{RBF}}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_2)^{\top} \Theta^{-2} (\mathbf{x}_1 - \mathbf{x}_2)\right) \quad [1]$$

The distribution expression of the GPR model is:

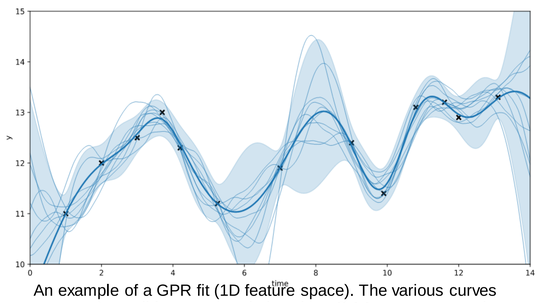
$$f(\mathbf{x}_i) \sim \mathcal{GP}\left(m(\mathbf{x}_i), K(\mathbf{x}_i, \mathbf{x}_j)\right) \quad [4]$$



Above is an illustration of a potential well similar to what would be seen in a diatomic molecule. The depth of the well is related to D_0, the location of the minimum is related to R_e, and the local parabolicity near the minimum is related to ω_e and $\omega_e\chi_e$.



Example of emission and absorption spectra [3]



An example of a GPR fit (1D feature space). The various curves represent possible means and show how GPR assumes no prior functional form. The shaded region can be viewed as uncertainty. [5]

Targets	Features
Equilibrium Internuclear Separation (Å)	Groups and periods of each constituent atom
Harmonic Angular Frequency (cm^{-1})	Groups and periods of the atoms, μ, R_e
First Anharmonic Correction (cm^{-1})	Groups and periods of the atoms, μ, R_e, $\ln(\omega_e)$
Binding Energy (eV)	Groups and periods of the atoms, R_e *(electronegativity difference)

R_e is the equilibrium internuclear distance, ω_e is the harmonic angular frequency, $\omega_e\chi_e$ is the first anharmonic correction, D_0 is the binding energy, and μ is the reduced mass. For D_0, the electronegativity difference is a feature still being tested.

Molecule	Prediction (Å)	Actual (Å)	Percent Error (%)
Fluorine	1.40778	1.41193	0.294
Chlorine	2.0001	1.987	0.66
RuC	1.6418	1.635	0.42
InBr	2.6444	2.54318	3.98
CoO	1.6532	1.6279	1.55
HCl	1.26829	1.27455	0.49

Prediction data for R_e

Molecule	Prediction (cm^{-1})	Actual (cm^{-1})	Percent Error (%)
Fluorine	788.02	916.64	14.03
Chlorine	560.201	559.7	0.09
RuC	1074.18	1030	4.29
InBr	236.291	221.0	7.03
CoO	816.281	862.4	5.35
HCl	3004.673	2990.946	0.46

Prediction data for ω_e

Molecule	Prediction (eV)	Actual (eV)	Percent Error (%)
Fluorine	2.519	1.602	57.24
Chlorine	2.4743	2.47936	0.204
RuC	6.168	6.6	6.55
InBr	3.728	3.9	4.41
CoO	3.5301	3.8 [2]	7.10
HCl	4.721	4.433	6.50

Prediction data for D_0

Training the Models

The features of each of the models were selected based on previous work [4], physical intuition, and model evaluation.

Our models used either the Matern kernel, RBF kernel, or products of the two.

The data was split between training and testing sets in either an 85:15 or 90:10 ratio.

With the kernels, number of training iterations, and training test-split decided upon, 500 or 1,000 models for each constant were trained and saved; this was done to combat the error of any single model by exposing the models to different splits of the data.

The performance of the models on molecules not included in the database was evaluated using percent error analysis.

Results and Discussion

The models were tested on 5-6 molecules that were omitted from the database. These molecules are molecular fluorine, molecular chlorine, indium bromide, ruthenium carbide, cobalt (II) oxide, and hydrogen chloride. It should be noted that fluorine and chlorine are homonuclear molecules, but the models were trained only on heteronuclear molecules.

The R_e model performed quite well on all novel molecules. This result is noteworthy because two molecules are homonuclear whereas the training data consisted of only heteronuclear molecules. The ω_e model performed relatively well, and the outlier in accuracy is F_2, which is not surprising due to its homonuclear nature. Like the ω_e model, the D_0 model did well on all molecules except F_2, for which the error is exceptionally high. The $\omega_e\chi_e$ model performed well for all molecules except F_2, continuing the trend. To improve the models' accuracy going forward, it is suggested that more molecules be included in the training data set; including homonuclear molecules in the training data may improve the models' accuracy for such molecules.

Molecule	Prediction (cm^{-1})	Actual (cm^{-1})	Percent Error (%)
Fluorine	7.9412	11.236	29.32
Chlorine	2.5226	2.67	5.52
InBr	0.6455	0.65	0.69
CoO	5.3215	5.13	3.73
HCl	53.109	52.8186	0.55

Prediction data for $\omega_e\chi_e$

Conclusion

The promising results of the models indicate that we are heading in the correct direction for using machine learning techniques in the DMSD. The future of this project will focus mainly on integrating ML into the DMSD. The performance of the models will continue to be improved by increasing the size of the training data set, experimenting with the types of kernels and model parameters, and trying new features. This project will also expand the capabilities of the on-the-fly ML part of the DMSD by including models for predicting additional spectroscopic constants.

References

1. GPyTorch, "gpytorch.kernels." Accessed March 16, 2023. https://docs.gpytorch.ai/en/stable/kernels.html.

2. Huber, K.P., Herzberg, G.. *Constants of diatomic molecules*. In: *Molecular Spectra and Molecular Structure*. Boston, MA: Springer, 1979. https://doi.org/10.1007/978-1-4757-0961-2_2.

3. Hustak, Leah. The European Space Agency "Absorption and emission spectra of various elements." Accessed March 16, 2023. https://www.esa.int/ESA_Multimedia/Images/2022/08/Absorption_and_emission_spectra_of_various_elements.

4. Liu, Xiangyue, Meijer, Gerard, Perez-Rios, Jesus. "On the relationship between spectroscopic constants of diatomic molecules: a machine learning approach." *The Royal Society of Chemistry* 11, no. 24 (April 2021): 14552-14561. DOI:10.1039/D1RA02061G.

5. Newman, Thomas. "Gaussian Process in Regression." May 5, 2022. https://www.lancaster.ac.uk/stor-i-student-sites/thomas-newman/2022/05/05/gaussian-processes-in-regression/.

Acknowledgements

We would like to acknowledge our advisor, Dr. Jesús Pérez-Ríos, for his help and inspiration on this project. We would also like to acknowledge our other team members, Yueqian Wang, Saketh Bhattiprolu and Connor Chin, who are developing the new DMSD website. We would also like to thank the Institute for Advanced Computational Science and Stony Brook Medicine for providing us with computational resources.