

This folder contains Python software to train and predict  $\chi$  using Gaussian Process Regression  $\chi$  is a correction to the Eckart transmission coefficient for thermal rate constants as described in refs. 1 and 2. Users are urged to read ref. 2.

Refs.

1. Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. J. Phys. Chem. Letts 2019, 10, 5250–5258.

2. A Machine Learning Approach for Rate Constants II: Clustering, Training, and Predictions for the O(3P)+HCl -> OH+Cl Reaction, Apurba Nandi, Joel M. Bowman, and Paul L. Houston J. Phys. Chem. A 2020, XXXX, XXX, XXX-XXX

Publication Date: June 16, 2020, <https://doi.org/10.1021/acs.jpca.0c04348>

\*\*\*\*\*

There are three equivalent Python codes

"Chi\_GP\_pyth2.py" --> This is Python 2.7 code

"Chi\_GP\_pyth3.py" --> This is Python 3.0 or higher version.

"Chi\_GP.ipynb" --> This is the same code but it runs on Jupyter Notebook.

\*\*\*\*\*

This is a two-step Machine Learning process, training and prediction. Training on  $\chi$  is done using one of two clusters of data that is selected by the user. Prediction of  $\chi$  is done for the user's reaction of interest, namely using the descriptors supplied by the user in. Generally, the training is done without user input. Expert users who are familiar with Gaussian Process Regression can alter the kernel and "noise".

\*\*\*\*\*

A schematic of the flow is given in the enclosed pdf.

\*\*\*\*\*

The user supplies a file (user-named, e.g., in the pdf **user\_predict.dat**) of descriptors for the user reaction and is asked to select a cluster of accurate values to train on. Output is the set of user-inputted descriptors followed by the predicted  $\chi$  and standard deviation. (The SD probably should be taken semi-quantitatively and not a rigorous set or error bars.)

The two clusters for training are **small\_chi.dat** and **large\_chi.dat** files. Users are asked to select one for training. Both can be used if desired. Output from this training step consists of files **small\_chi\_train.out** and **large\_chi\_train.out**, respectively. (Other training output gp\_theta.dat and gp\_alpha.dat are seen, these can be ignored by ). Prediction is done using the trained GP and the **user\_predict.dat** file.

In the folder there are two sample inputs **Predict\_small\_chi.dat** and **Predict\_large\_chi.dat** and corresponding sample outputs are **Small\_chi\_predict.out** and **Large\_chi\_predict.out**. These are probably good to keep for testing purposes.

Details about these clusters and guidance on choosing one over the other or using both and then weighting the results are given in Ref. 2.

The descriptors are  $u^*$ ,  $\alpha_1$ ,  $\alpha_2$ ,  $\beta(\text{deg})$ / They are described in detail in the above references

.....

Note the training for the O+HCl reaction **O+HCl\_Data.dat** is not in either the **Predict\_small\_chi.dat** and **Predict\_large\_chi.dat**. This is because this reaction is an “outlier” as discussed in detail in ref. 2. The user may add some or all these data to the appropriate **small\_chi.dat** and **large\_chi.dat** file.

Important Note: All the training and test data files should be kept inside the same working directory.

\*\*\*\*\*

Contacts: Apurba Nandi, Dept. of Chemistry, Emory University, Atlanta, Georgia, USA.

Email: [apurba.nandi@emory.edu](mailto:apurba.nandi@emory.edu), [apurba.nandiju@gmail.com](mailto:apurba.nandiju@gmail.com)

Joel M. Bowman, Dept. of Chemistry, Emory University, Atlanta, Georgia, USA.

Email: [jmbowma@emory.edu](mailto:jmbowma@emory.edu)