A data-driven approach to predicting Chemical Reaction Kinetics

Maneet Goyal, Keren Zhang

Feb 02 2018

Background

The products of chemical reactions are dependent on various factors, such as temperature, pressure and molecular properties, which underscores the complexity of this process. Meanwhile, it is imperative for us to figure out the impact of these factors, and to be able to predict the results of these reactions based on varying reaction conditions. However, often times it is extremely difficult to elucidate the underlying mechanisms and foretell what occurs in a reaction system. The tide of using data science to tackle such problems has risen recently, thanks to the advent of many large on-line databases and tools that facilitate a wide scale acquisition of relevant data in time efficient manner.

Goals

At the outset, the objectives of our project are to:
1. Successfully retrieve reaction and molecular data via web APIs, web scraping, or databases and then integrate them into feature vectors.
2. Use data science and machine learning approaches to provide quality predictions of:
a. Reaction Feasibility b. Reaction Order c. Activation Energy
3. And then finally make an attempt to validate our results using chemical engineering knowledge.

Definition of Success

Achieving over 90% accuracy in predicting reaction feasibility, over 75% in reaction order, and a relative error of under 20% in activation energy. Goals 1 and 2 are primary and critical to success. We have already scraped over 28000 NIST web pages and transfered the data into a SQLite database for future use. Goal 3 is secondary but quite important in that it helps integrate chemical engineering knowledge with our results.

Deliverables

The *3 models/ensembles* (Python functions, hyper-parameter JSON, if applicable) which predict the proposed parameters and a *comprehensive dataset* on gas phase reactions. Others include plots depicting the accuracy of our models, an OOP based Python package for reproducing our data collection/processing results, and a schematic which presents the main work-flow of our complete analysis.

Challenges

Being able to predict the proposed parameters via a general and robust model may require years of experimental, theoretical or even computational (simulation based) research. Data analytics, on the other hand, provides us with tools to process gigantic datasets and bestows opportunities of using statistical methods to build surrogate models that help correlate the input descriptors/parameters and outputs, and recognize patterns in reactions. However, the scale of data we may encounter and the speed & variety of analysis required could be overwhelming for traditional data analysis tools like MS Excel. Moreover, traditional tools assume that the data is already available in a tabular format. Clearly, this is not the case here where we had to resort to using Python to scrape the required data off the HTML sources. Although Excel also allows scripting to process data programatically, but the off-the-shelf Python libraries make the development time shorter.

Database Description

Our kinetics dataset is scraped off the NIST Kinetics Database which features essentially all the gas phase kinetics studies involving elementary reactions reported in over 12,000 papers through early 2000s. The host website reports over 11,700 unique reactant pairs in their pool of more than 38,000 reaction records. Each reaction record includes include one or more of the following (explicit) descriptors: reactants, products, rate Parameters: $A, n, \frac{E_a}{R}, k$, where $k = A(\frac{T}{298})^n e^{\frac{-E_a}{RT}}$. Here, k is rate constant, $A(\frac{T}{298K})^n$ can be seen as the frequency factor, R is the universal gas constant and equals $8.314472 * 10^{-3} \frac{kJ}{mol-K}$, T is temperature in Kelvin, and E_a is activation energy in $\frac{kJ}{mol}$. For rate constant, the units are: a. First Order: s^{-1} b. Second Order: $\frac{cm^3}{molecule-s}$ c. Third Order: $\frac{cm^6}{molecule^2-s}$

Uncertainty in $A, n, \frac{E_a}{R}$, temperature range of experiment or of validity (in case of review/theoretical studies), pressure range and bulk gas of the experiment, record type - experimental, theoretical, modeling-based, relative-rate, etc., and experimental procedure, including excitation technique are also reported.

For our study, we scraped reactants, products, rate constant, activation energy, A, temperature range, reaction order, the record's Squib code and the URL to further details on that record.

In addition to above, we also plan to extract additional (implicit) descriptors in the form of molecular properties, such as size, bond strength, and functional groups. Combining explicit and implicit inputs, we hope to acquire a relatively comprehensive set of features of interests to be mapped towards outputs.