A Data driven approach to predicting Chemical Reaction Kinetics

Maneet Goyal, Keren Zhang

April 24 2018

1 Abstract

In this report we examine the relationship between chemical bonds formed and broken during reactions and Arrhenius kinetic parameters, including activation energy and reaction order, using machine learning approaches. We utilized the NIST kinetic database to acquire kinetic data and the perform thorough data cleaning to convert HTML-based data to more accessible dataframe format. We constructed the feature vector that contains broken and formed bonds and utilized it to train our models to predict activation energy and reaction order. The Random Forest based classification model was chosen because it is an ensemble method and is more robust, has a more intuitive basis behind its working, and is also resilient to overfitting. Huge class imbalance was encountered in favor of second order reactions and oversampling of the minority classes was done by random oversampling and synthetic minority over-sampling technique (SMOTE). A high prediction score of around 95% was achieved. In regards to regression for predicting activation energy, the results obtained call for the engineering of more comprehensive/informative feature vectors. Finally, we assessed our results based on statistical and chemical engineering knowledge.

2 Introduction

2.1 Cheminformatics Overview

Cheminformatics is the use of computational and informational techniques to understand problems of chemistry. The rapid growth of cheminformatics indicates its emerging power in multiple areas, such as drug deliveries and kinetics [1]. In many scenarios, traditional computational chemistry is unable to correlate compounds' properties and structures due to their extreme complexity. This shortcoming, however, inspires people to take advantages of in-silico data-driven approaches to tackle the problems that traditional methods cannot deal with [2].

Information technology in scientific research often utilizes the comprehensiveness and the high dimensionality of the data accumulated by the continuous efforts of both computational and experimental communities [3]. This data may be gradually shared with open access, as in the case of NIST Kinetics database. Data science can enable integration of many such data sources and help researchers identify underlying patterns in the domain of interest and provide reliable predictions of the information which was earlier harder to obtain.

2.2 Machine Learning Methods and Kinetics Study

Machine learning (ML) plays an important role in cheminformatics. It seeks to derive insights from data by integrating computer science, statistics and information theory [4]. Many of the ML approaches are quite scalable and adaptive, and hence can effectively deal with large volumes and variety of data.

Recently, many ML models have been identified as effective and accurate tools in elucidating convoluted chemical reaction systems. Here, the ML models that researchers built often used molecular representations as input, such as specific functional groups, physical and chemical properties [5]. Either experimental or simulated data can be used as input to support the training of our ML models to

finally be able to predict the information of interest. *Ramakrishnan et al.* [6] combined machine learning models and quantum chemistry approximation to accurately predict the thermochemical changes during isomerization reactions. *Glielmo et al.* [7] found that the composition of gas combustion waste can be predicted by an artificial neural network with the temperature and concentration profiles as inputs. Quite recently, the chemistry analogue of AlphaGo developed by *Segler et al.* [8] surpassed human being's performance in designing organic chemistry synthesis routes.

Machine learning methods can flexibly cope with different representations of molecular information. Evidently, they have enormous potential in exploring chemical space and facilitating the process of discovering patterns. In this paper, we evaluate different machine learning methods to study chemical kinetics, in particular, to explore the strength of the relation between the types of bonds formed and broken during a reaction with the reaction order, via classification, and the activation energy, via regression.

3 Methods

3.1 Data Identification

The primary source of our kinetics data is the NIST Kinetic Database [9]. It contains a large amount of experimental and computational results of gas-phase reactions from various papers, wherein researchers report kinetic data for some specific reactions they study extensively. This kinetics data include: reactants and products, activation energy, reaction order, pre-exponential/frequency factor, the temperature range at which the data is measured and the original source paper title, among others.

Most of the compounds involved in the reported reactions are present in the PubChem database, another comprehensive database storing chemical molecules along with their properties. Through PubChem, we requested the molecular properties of the concerned species in JSON format and then filtered and parsed the bonding information required to construct our feature vectors.

3.2 Data Gathering

The entire workflow behind data gathering and cleaning is elucidated through figure 1. In the initial stages, we pulled an HTML page containing all the reactions and their URLS from the NIST database by performing an all-inclusive query (Reaction Order ≥ 0) and then performed web scraping to fetch the kinetic records of all those reactions. This was followed by data cleaning via Python, OpenRefine and manual means to tackle encoding issues and unexpected delimiter effect (offsetting of the data by a few columns). Thereafter, data augmentation was performed both within our datasets and with an external source, PubChem to give rise to feature vectors. Some issues encountered in this stage are discussed in the Appendix (6).

3.3 Feature Vector Construction

Our input is a vector that contains chemical bonds broken and formed during chemical reactions. We designed a process that dissects each molecule into the chemical bonds it has. First, we request data (JSON file format) from PubChem for each chemical compound, which has a record of the atoms associated with each bond and the corresponding bond order. This yields a list of all the chemical bonds a certain molecule has, which is to be compared to a heuristically constructed vector whose individual element is a type of chemical bond and is identified by its fixed index position. For example, "C-H" single bond is always the 13^{th} item . Such comparison allows us to build a consistent feature vector indicating both the type and the number of chemical bonds a molecule has. A "C_2H_6 molecule, for instance, should be decomposed into one "C-C" single bond and six "C-H" single bonds. Subtracting the sum of feature vectors of the products by that of reactants leads to a final feature vector of the corresponding reaction.

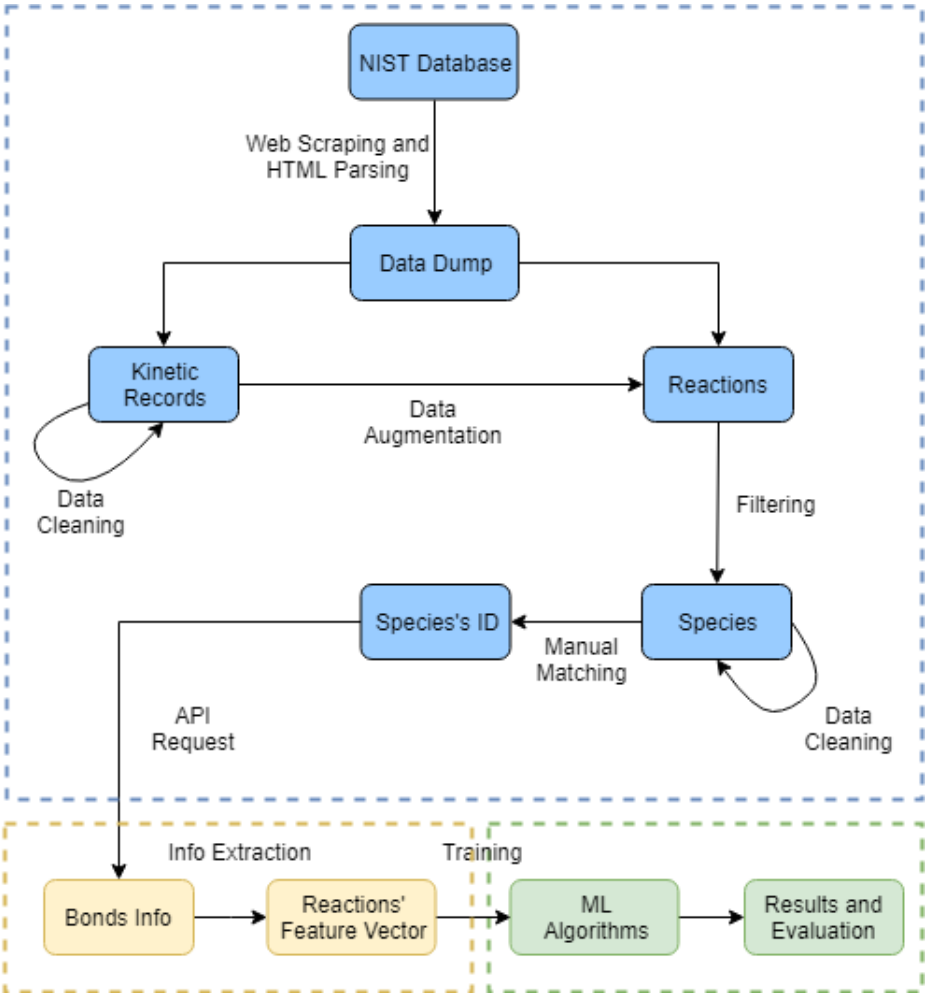

Figure 1: Main Workflows

4 Results and Discussion

4.1 Correlation of Features

The refined data set has dimension of 42, given by the number of the types of chemical bonds. The high dimensional nature of the data increases the complexity and difficulty of subsequent data analysis and visualization. Principle component analysis (PCA) is performed to evaluate the relative the "significance" of each feature and reduce the dimension. The results of the PCA is illustrated in the figure 2. There are no highly dominant features and the covariance between each feature is low. This is not surprising because each chemical bond is independent of each other. The most salient feature correspond to carbon-hydrogen single bond. However, it is insufficient to ascribe it to a principle component because it forming and breaking are quite common in small gas-phase molecule reactions, majority of which are small organic compounds. Although in the Scree Plot 95% can be explained by around 14 features, some "unimportant" features might contain critical information, such as rarer chemical bonds. Thus, we opt to retain the original dimension of the data set in the subsequent analysis.

4.2 Linear Regression for Activation Energy

For predicting the activation energy, we used 2 models: Linear Regression and Random Forest Regressor. Both were evaluated via 10 fold cross validation and it showed unsatisfactory performance as shown in Figure 3. The huge amount of variance in R^2 score in both a linear and a non-linear model suggests the absence of any strong correlation between the nature and type of bonds formed/broken and the activation energy. This seems counter intuitive, however. Engineering a better feature vectors

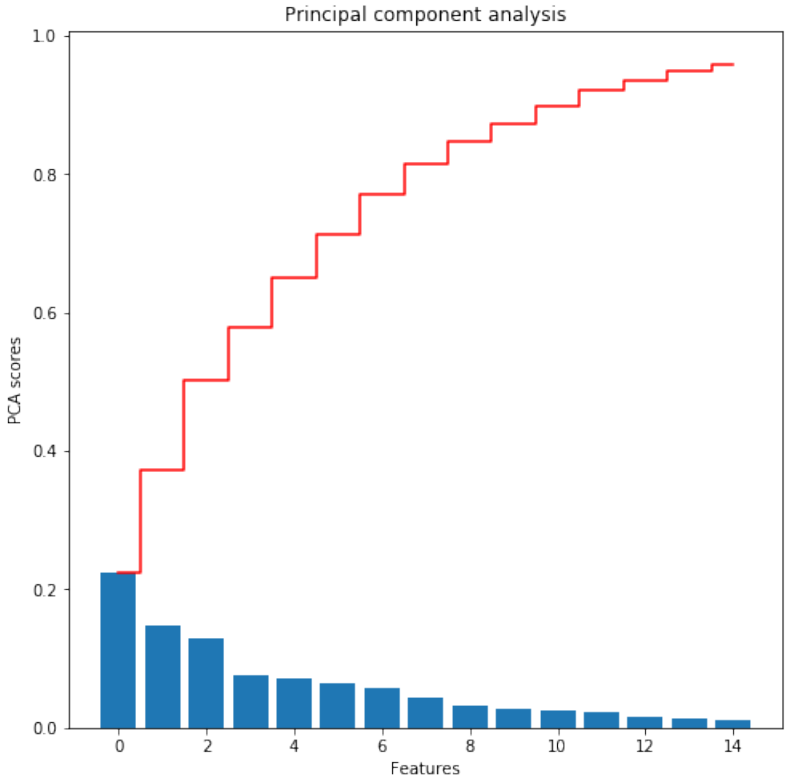

Figure 2: Principal Components

encompassing the stereochemistry of the molecules may have the potential of generating better results. Random Forest Regression shows a good fit when its fitness is evaluated over the same data set it is trained over; implying over-fit.

4.3 Classification

The reaction order usually is an integer between zero and three, which makes it an apt categorical variable for classification. Since the dimension of our feature data set is extremely high, we choose random forest classifier, a good candidate for dealing with high dimensional data. We split the data set and perform cross-validation via k-fold methods (k = 10). The results of the classification approached around 95% accuracy. Hyperparameter tuning of maximum depth and random state indicates that the default (automatic) parameters we choose already give us good predictions and variation of parameters does not affect the results too much. However, it was our contention that it is largely due to the class imbalance since 96% of our reactions are second order. Therefore, it was assumed that the classifier is trained to select reaction order of two, regardless of the variation in feature vector. The main classification errors, indicated by the confusion matrix, occur when first and third order are incorrectly predicted to be second order (false positive). Taking a quick look at our feature vectors, we found that a large amount of reactions take place between two radicals, and the quick combination of two radicals results in a second order reaction.

4.4 Dealing with Class Imbalance

The following measures were adopted to deal with it: Random Oversampling with Replacement (ROR) and SMOTE [10]. SMOTE differs from ROR in that it doesn't replicate the data points randomly until it restores the class imbalance. Instead, it creates synthetic data points via interpolation between minority class instances. One issue with SMOTE oversampling is that if we choose a naive interpolation, then the intra-class variance may be effected because the new data points will likely be place in the centroid on its chosen neighbours. In an attempt to reduce the alteration of this variance, various flavours of SMOTE are avaiable [11]. In this study, we also use one such flavour based on SVM.

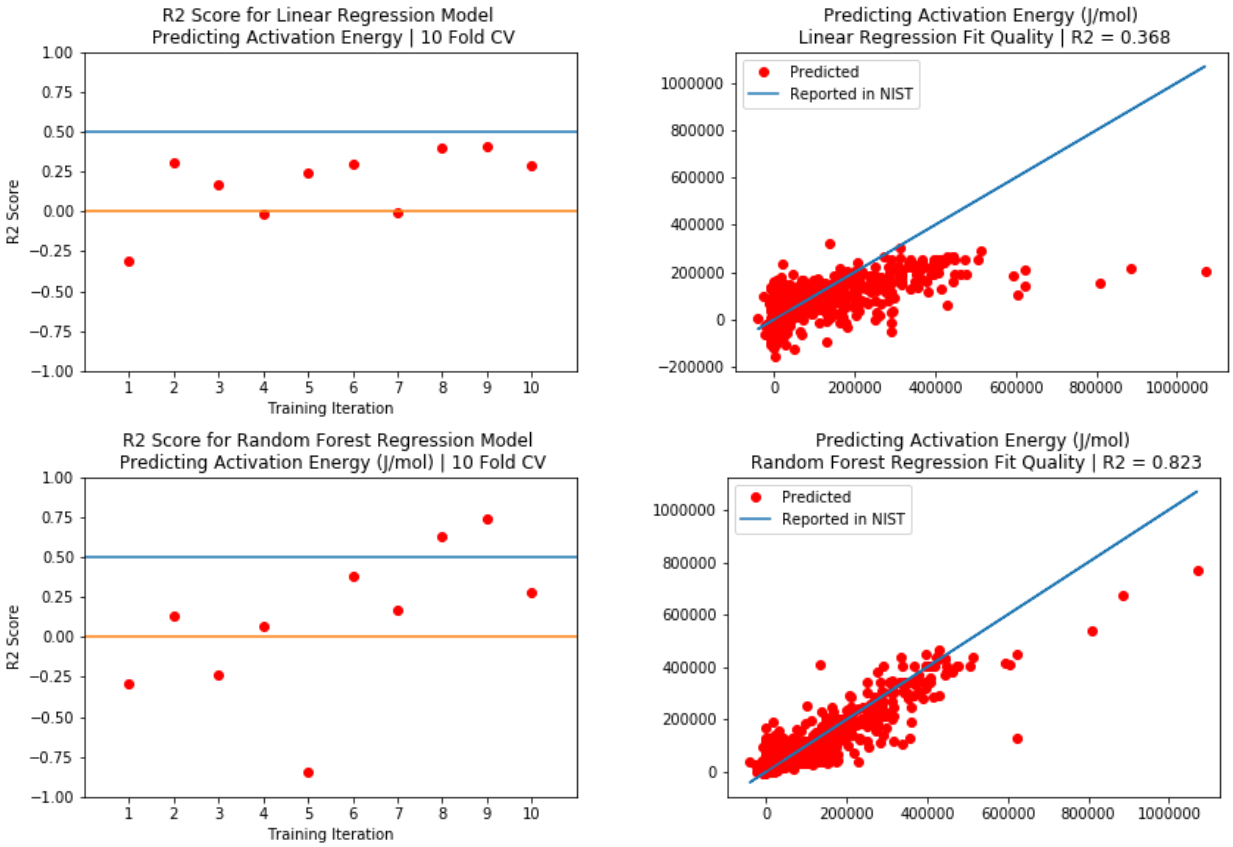

Figure 3: Regression Results for predicting Activation Energy

Both the cases gave a good classification accuracy: a 10 fold CV over oversampled data generated via both ROR and SMOTE gives a mean score of around 95% as suggested by Figure 4 also. On further investigation, we also noted that the random tree classifier gives around 92% accuracy even when the maximum tree depth is set at 2 which is highly surprising and point towards the existence of elements in the feature vector which strongly separates the reaction classes.

Given our datasize, it's hard to confidently conclude that the correlation between bond types broken and formed and reaction is really as strong as depicted by our results. Therefore, analysis with more data points is suggested as far as a data analytic approach to solving this problem is concerned.

5 Notes and Recommendations

In our process of data acquisition and cleaning, we found that the data available in the data base is not complete in that the products/activation enery/reaction order of many reactions are missing. The data handlers behind this database have gathered information from review/research papers to give rise to this gigantic database so it's very likely that the missing values weren't reported in the original papers in the first place. Further, while creating feature vectors, we realized that there are some complex chemicals whose information are not present in general purpose chemical databases like PubChem and ChemSpider. There is a large room for the NIST database's managers to organize data in a more user-friendly format, for instance, by presenting all the species in a uniform format, say by IUPAC names, instead of the current heterogeneous mix of common names, empirical formulae and IUPAC names.

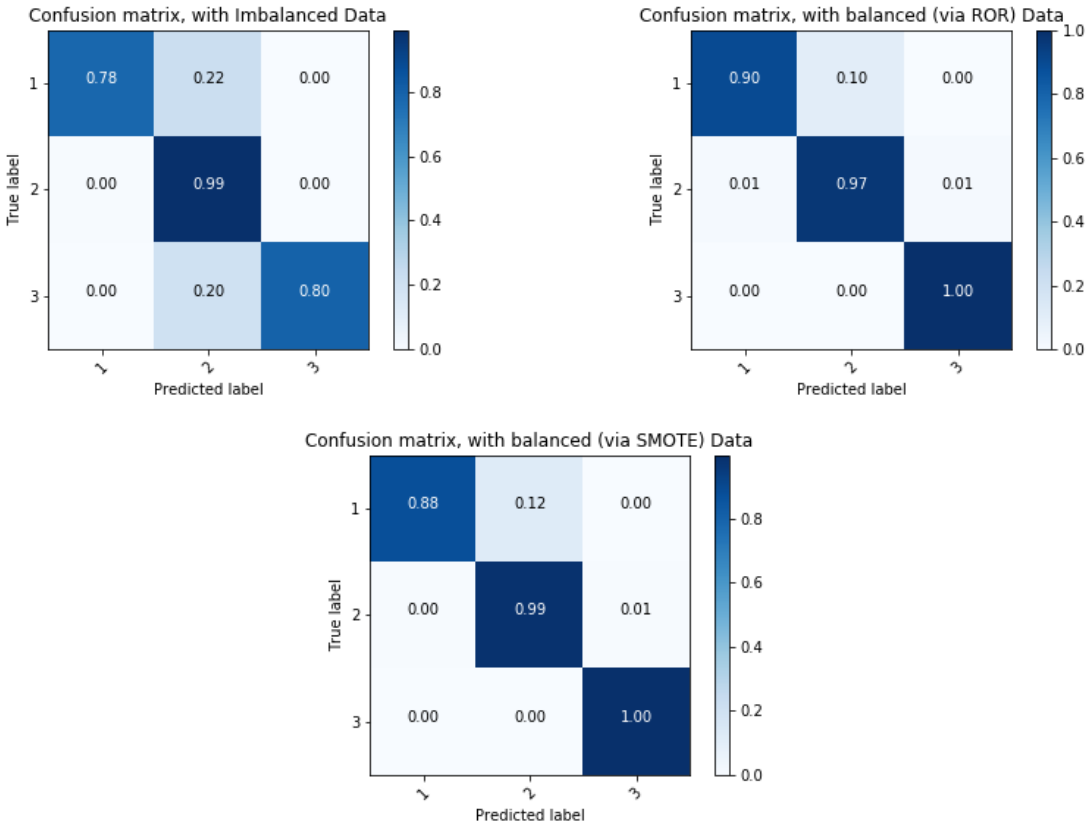

Figure 4: Confusion matrix. Top-Left: With Class Imbalance. Top-Right and Bottom: Without Class Imbalance

5.1 For Future Researchers

Appendix contains some important issues that we faced during data gathering and cleaning. Since, this exercise is already done, the future researchers can focus more on the machine learning aspects of this problem. One way more reliable results can be achieved is via manual fetching of the PubChem IDs of more species. We have designed a pipeline via which the users only need to update the Excel copy of the Species Dataframe with more PubChem IDs, and the information required for feature vector construction can then be augmented with just a few lines of code. They are also suggested to retrain the classification models with reduced dimensions and/or better feature vectors. Another pipeline has been included which allows the data augmentation with ChemSpider data in addition to PubChem data to give the researchers an easy access to more chemical/species information.

Link to GitHub Codebase: kineticsML

6 Appendix

This section presents the main issues we faced while gathering and cleaning data, and is meant mainly for supporting future extensions of this project. We started the process by querying the NIST Chemical Kinetics Database for all the reactions that it stored. NIST returned the results in a single HTML page which we then parsed using **BeautifulSoup** to extract all the reaction URLs/links present. We then scraped each of those links (around 28000) to extract information about different works which report some kinetic information on those reactions. The following issues were faced in scraping the data via those URLs:

1. Bandwidth Throttling: The URL requests were made via a Python module called **urllib**. In this process, the requests were initially made over a private internet connection which we hypothesize had a throttled bandwidth. Fetching around 500 pages took close to 1 hour. We then switched

our connection to the one provided by Georgia Tech via VPN and obtained a tremendous speed-up. All the remaining >27000 pages were fetched in around 1 hour.

2. Bad HTMLs: In a few iterations of the web scraping routine, we realized that our code broke. It was observed that the generalization we had assumed for all the web-pages do not hold. There were some pages which had a different DOM structure and therefore, a few modifications were made and a few **try-except** block were introduced to handle those cases. Those exceptions are categorized into 2 classes:

   (a) Reference Reactions: Some of the kinetic parameters in the NIST Chemical Kinetics Database were reported by comparing a reaction to some reference reaction. Such reactions were represented in a different DOM structure. Since our analysis doesn't consider such reactions, they were stored separately in another dataset.

   (b) Fake Links: Some of the HTML pages simply didn't have any data to display. These pages were ignored via a try-expect block.

The HTML parsing and web scraping routine discussed above resulted in 3 datasets:

1. **Dataset A**/Chemical Reactions: Contains all the reaction URLs, reactants and products, and the number of records available in the NIST database for each of those reactions.

2. **Dataset B**/Normal Reaction Records: Contains information on the paper which has reported some kinetic information on those reactions, the kinetic information itself (reaction order, activation energy, temperature range, etc.).

3. **Dataset C**/Reference Reaction Records: The URLs to the reaction records which correspond to reference reactions.

We then parsed Dataset A to create a child dataset, **Dataset D**/Species which house all the unique chemical species present in Dataset A. Dataset D was then subjected to data augmentation via ChemSpider and PubChem APIs. The following issues were faced:

1. The unique species data / Dataset D was highly heterogeneous. Some of the species were represented by common names, and others by molecular formula. Requesting data via the ChemSpider API using molecular formula as a search query resulted in multiple results since many molecules can have the same chemical formula. Selecting the right molecule from that list could not be done programatically.

2. Even those results which had only one record in the API response weren't guaranteed to be correct. For instance, the ChemSpider API confused carbon-monoxide (CO) with Cobalt (Co).

3. Some molecules (in the form they were represent in the Dataset D) didn't exist in the ChemSpider database.

4. Making API calls was a time expensive process given the number of species (around 7000) we had in our dataset.

We then designed a few metrics that helped us better plan the feature engineering stage and manually extracted the PubChem CIDs of around 100 chemicals which were a part of around 1600 reactions. The PubChem request were made programatically and the responses contained bond information which was used for feature construction. Our performance metrics suggest some important info about Dataset A:

1. Out of 28983 reactions, only 18680 had no missing products.

2. Out of the above 18680, only 1763 reactions had species that occurred in $\geq$100 reactions.

3. These 1760 reaction are currently considered for analysis. These too suffer from missing information:

(a) Only 1088 had activation energy reported.

(b) Only 1747 had reaction order reported.

4. **The above subsets of 1088 and 1747 reactions were chosen for regression and classification, respectively.**

References

[1] R A Lewis R D King, S Muggleton and M J Sternberg. Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceeding of the National Academic of Science of the United States of America*, 89(23):11322–11326, 1992.

[2] Alexandre Varnek and Igor Baskin. Machine learning methods for property prediction in chemoinformatics: Quo vadis? *J. Chem. Inf. Model.*, 52:14131437, 2012.

[3] William H. Green Gregory R. Magoon. Design and implementation of a next-generation software interface for on-the-fly quantum and force field calculations in automated reaction mechanism generation. *Computers and Chemical Engineering*, 52:35–45, 2013.

[4] Jonathan H. Chen Matthew A. Kayala, Chloe-Agathe Azencott and Pierre Baldi. Learning to predict chemical reactions. *J. Chem. Inf. Model.*, 51:2209–2222, 2011.

[5] John B. O. Mitchell. Machine learning methodsin chemoinformatics. *WIREs Computational Molecular Science*, 4, September 2014.

[6] Raghunathan Ramakrishnan, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The -machine learning approach. *JCTC*, 11:20872096, 2015.

[7] Luigi Glielmo, Michele Milano, and Stefania Santini. A machine learning approach to modeling and identification of automotive three-way catalytic converters. *IEEE/ASME TRANSACTIONS ON MECHATRONICS*, 5, June 2000.

[8] Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555, March 2018.

[9] J. A. Manion; R. E. Huie; R. D. Levin; D. R. Burgess Jr.; V. L. Orkin; W. Tsang; W. S. McGivern; J. W. Hudgens; V. D. Knyazev; D. B. Atkinson; E. Chai; A. M. Tereza; C.-Y. Lin; T. C. Allison; W. G. Mallard; F. Westley; J. T. Herron; R. F. Hampson; D. H. Frizzell. Nist chemical kinetics database, nist standard reference database 17, version 7.0 (web version), release 1.6.8, data version 2015.12. *National Institute of Standards and Technology, Gaithersburg, Maryland, 20899-8320.*

[10] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.

[11] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 2018.