Estimation of Heat of Fusion and Acid Dissociation Constant using Machine Learning Techniques

**Abstract**

Three types of molecular representation -- Group Contribution, Mol2vec, and Mordred, were compared to determine their capabilities and limitations in molecular property prediction. Two molecular property datasets, acid dissociation constant (pKa) and heat of fusion (Hfus) were trained. The results from each calculation were fed into regression models as input. The regression model then gives out the predicted property value which was compared with the actual property value. The prediction accuracy of each property from different models was compared. The descriptors from each method were also visualized in two and three-dimensional plots against property value to check out if any grouping can be observed. Group contributors were able to predict heat of fusion much better than acid dissociation constant. Regression methods such as Support Vector regression (SVR) using RBF kernel, random forests and neural nets performed well in predicting both properties. However, in case of heat of fusion there was an overall higher R2 score across the board for all regression methods used. Grouping of molecular contributors can be observed from both properties visualization. Mol2vec showed better performance in case of Ridge regression and K- nearest neighbors Regression methods. Mordred presents better prediction accuracy with regression methods using SVR and random forests for pKa but not for Hfus. No grouping of Mordred descriptors can be seen from the visualization.

**Introduction**

Chemoinformatics is “the application of computer sciences or informatics for solving chemical problems has expanded into various areas of chemistry” by Thomas Eagle [1]. This Field had expanded into six aspects, which are (1) representation of chemical compounds, (2) chemical datatypes, (3) chemical databases and data sources, (4) search methods, (5) calculation of descriptors, and (6) data analysis [1]. In this report, we are going to be focused on the first and last two aspects of Chemoinformatics to predict molecular properties that were given. The use of computational property prediction allowed chemists to find relationships between molecular properties and molecule structure to reduce the time of experimental analysis.

To give the best prediction results, three types of calculation methods were compared which are group contribution, mol2vec, and Mordred methods. Each method was used to predict two different types of molecular property, acid dissociation constant and heat of fusion, to compare the suitability and limitations of the calculation when dealing with different properties. The calculation result from each method was fed into regression algorithms and visualized to determine which methods could better capture the performance of the property prediction model. Multiple regression models were used and trained, and the determination coefficient score was used as a scalar to represent the regression’s performance. From the results, group contribution was found to have better predictions for heat of fusion, while Mordred is better at predicting acid dissociation constant. Mol2Vec also predicts the heat of fusion much better than acid dissociation constant.

**Methods:**

**Representation**

Group Contribution

GC-based class of methods have been widely used in a variety of applications due to their invertibility, easy incorporation within mathematical models, and reasonably accurate estimates at an affordable computational cost. Under this class of models, pure component properties of chemicals are calculated from the contributions of the functional groups (also called molecular fragments), that represent the molecular structure of the chemical

Mol2vec

Mol2vec, which is an unsupervised machine learning approach to learn vector representations of molecular substructures. Mol2vec learns vector representations of molecular substructures that point in similar directions for chemically related substructures. Compounds can finally be encoded as vectors by summing the vectors of the individual substructures and, for instance, be fed into supervised machine learning approaches to predict compound properties.

Rdkit

Rdkit is an open-source cheminformatics software that stores basic molecular functionality. It can be installed on to python. It was used to convert SMILES string into series for Mordred calculation.

Mordred

Mordred is a molecular descriptor calculator, and was installed to fit SMILES that were transformed by the Chem function in the Rdkit package into the Mordred model and descriptors values for each molecule. The results were calculated. The raw descriptor columns were adjusted to remove invalid numbers.

t-SNE

The descriptors or contributors were dimensionally reduced to a low degree and data points were color coded according to property values using t-distributed stochastic neighbor embedding (t-SNE). T-SNE was imported from the python sklearn package.

**Regression**

Two property datasets were fit into Mordred and Mol2vec model to calculate different representations of molecule structures. Group contributors were provided so it does not need to be calculated. A total of six sets of data were visualized and trained to compare its performance at predicting specific molecule properties. Several regression methods were incorporated to utilize the three different representations used in this project. These methods are Linear regression, Polynomial regression, Support Vector regression, Decision tree, Ridge regression, Kernel Ridge regression, Random forests, KNN, Gaussian Process regression and Neural nets. All these methods were trained on 80% of the given dataset for Heat of Fusion and Acid dissociation constant, hyper parameters were tuned on 10% of the dataset and tested on the remaining 10% of the dataset with the score calculated on this portion of the dataset.

Linear regression and Polynomial regression was performed in coordination with Elastic Net regularization. The degree was set to 2 for Polynomial regression. When tried with the degree 3 we were facing session crash due to complete usage of available ram. Therefore, we had to settle with the degree 2. L1 ration was also a hyper parameter tuned to give optimal result.

Support Vector Regression was performed with the help of linear kernel, polynomial kernel and radial basis function kernel. The hyper parameters tuned include regularization parameter and epsilon value. Epsilon value stands for the region where no penalty is issued in training loss function with points predicted within a distance epsilon from the actual value.

Decision tree was performed with the squared error criterion selected. Max depth was the hyper parameter tuned for optimal result. Ridge regression was performed with the hyper parameter associated with regularization tuned

Kernel Ridge regression was performed incorporating linear, polynomial and radial basis function kernels with the regularization parameter tuned for each kernel and the degree set to 2 for polynomial kernel.

Random forests was utilized with hyper parameters such as n\_estimators, max\_depth and ccp\_alpha tuned for best score. n\_estimators is associated to number of trees in the forest. Max\_depth is associated to the maximum depth of a tree. ccp\_alpha is a complexity parameter used for Minimal Cost-Complexity Pruning. The criterion selected was squared\_error.

KNN was also used with the weights set to uniform and metric set to minkowski. The number of neighbors was tuned to achieve the best performance on the validation set.

Gaussian Process Regression was used with the DotProduct() and WhiteKernel() used. Hyper parameters associated to regularization and number of restarts were tuned.

Finally neural nets were utilized by setting up a sequential model with the first set of inputs being the values from group contribution. Tensor flow and keras libraries were imported. The activation function utilized in three layers was relu. Adam optimizer was utilized with hyper parameters such as learning rate, beta\_1 and beta\_2 were tuned upon validation set. While fitting the training data epochs was set to 50, batch size to 32 and validation split as 0.15. The parameters for compiling model were mean\_squared\_error for loss, RMSprop for optimizer and mean\_absolute\_percentage\_error for metrics.

The process for optimizing the hyper parameters for all these methods was by training the model on training dataset and check the coefficient of determination(score) for a range of set of hyper parameters on the validation set followed by plotting the score against the values used for hyper parameters and then selecting the best set of hyper parameters.

**Visualization**

All representation methods were visualized in 2D and 3D plots. The effect of perplexity was studied using the group contributors dataset. As perplexity increases, the data points are more likely to split into different groups rather than scattered around the middle. Usually, higher-dimensional datasets need higher perplexity so that perplexity at 100 was chosen [5]. The colors of the data points represent different property values. For the heat of formation property plot, some trends can be seen as higher heat fusion molecules were at the tip and lower heat of formation molecule scattered around and diverged (Figure 1). Furthermore, the 3-dimensional t-SNE plot was made to see if any trends of pKa can be found (Figure 2).

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| Figure 1. The two-dimensional plot of group contributions (left: pKa, right: Hfus). | |
| https://lh5.googleusercontent.com/ZePO1sZ7uBSyr4p6qITC2oy8Q16KdbLApUHoYy1xJpfNuyMK8-ENYhLuh_WKzb65guoeTbI5VeJnzsuq6dBuFsrqyJsVlKC1H-7_ow3dmA05Bmt5_DxbCR7ALseADx8uLnBR2poc | https://lh6.googleusercontent.com/cQyAtN2Dp8iAwMS-bai5c22tXCTrt3hUycH8AsURIFuu34a5m8hDuRIsgSRLfVhYIBY1e413k50kZmvlf4-NXkneuaH-6T3J0AcRAFikvqrmCewksI8l_D7rhlhgeN6k-XFzNwAB |
| Figure 2. The three-dimensional plot of group contribution (left: pKa, right: Hfus). Dynamic plot are included in the Mordred code. | |

Color clustering can be observed from both properties datasets in the 3D plots. In the pKa plot, the data points presented a curved sheet, where positive pKa molecules are on the inside and negative pKa molecules are on the outer side. This shows that the group contributors can capture similarities in chemical structures with regard to pKa values. The heat of fusion graph also shows a trend of lower heat of fusion is more toward one direction (up left in this view in Figure 2), suggesting that group contributors could capture some similarities of chemical structure that affect the heat of fusion.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| Figure 3. The two-dimensional plot of Mordred descriptors (left: pKa, right: Hfus). | |
| https://lh5.googleusercontent.com/QfUbca2tnMOHC0KtNrAO9XJz3uoD4NQ8WNDEixOfYBlHZZlnfK79QIYaYsQCRxRF7PQVXy-3d5QoLisk4yDhHWAsmeV5iYTluiAvpkKcyKMalHUFgoO-ywwz1ftDv4skey1s6GdN | https://lh4.googleusercontent.com/uRyNoTtiHIWdexL8nvvnUIaJp7ZnnpCXwOnLBUyV0AbbTH1zsZcDY1-BuAFZavZP0oUeUyczZ3ufRat9cwW4tYzkUMNvHLuJt3y6J3rCnmGLMDBGlambqGUn4nKXYwPPUzYzZOkT |
| Figure 4. The three-dimensional plot of Mordred descriptors (left: pKa, right: Hfus). Dynamic plots are included in the Mordred code. | |

The visualization of Mordred descriptors regarding properties was shown in Figure 4. No trends between properties and molecular representation could be observed from the 2 dimensional plot as well as the 3 dimensional plot. Colored dots seemed to be randomly scattered throughout the space. Some more improvements to find trends that are not obvious include dividing the property value into smaller groups and classifying data points.

|  |  |
| --- | --- |
|  |  |
| Figure 5. 2D TSNE for Mol2Vec Based Representation (pKa data) | Figure 6. 2D TSNE for Mol2Vec Based Representation (HFus data) |
|  |  |
| Figure 6. 3D Visualization of feature space using Mol2vec Representation (pKa data) | Figure 7. 3D Visualization of feature space using Mol2vec Representation (HFus data) |

In the 2D TSNE, the data points are somewhat randomly arranged and can be classified into few separate clusters. However in the 3D visualization, the data appears to be a single large cluster of data points arranged in an oval shape for pKa, in the case of Hfus the irregularity in shape is noticeable.

**Results**

Using Group Contribution, Mol2vec and Mordred representations the results of the regression methods are evaluated with help of coefficient of determination, also known as R2 score. The methods along with their score on validation and test datasets for Acid dissociation constant (pKa) and Heat of fusion (Hfus) are as follows.

Table.1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Group contribution** | | | | **Mol2vec** | | | | | **Mordred** | | | |
| **pKa** | | **Hfus** | | **pKa** | | | **Hfus** | | **pKa** | | **Hfus** | |
| Validation set | Test set | Validation set | Test set | Validation set | Test set | | Validation set | Test set | Validation set | Test set | Validation set | Test set |
| Linear regression | 0.21 | 0.24 | 0.49 | 0.89 | 0.31 | 0.40 | - | | - | 0.51 | 0.49 | 0.50 | 0.75 |
| Polynomial regression | 0.05 | 0.26 | 0.62 | 0.86 | 0.006 | 0.32 | 0.69 | | 0.41 | - | - | - | - |
| Support Vector regression-Linear kernel | 0.10 | 0.28 | 0.61 | 0.90 | - | - | - | | - | - | - | - | - |
| Support Vector regression-Polynomial kernel | 0.18 | 0.16 | 0.57 | 0.86 | 0.42 | 0.03 | 0.56 | | 0.70 | 0.33 | 0.26 | 0.19 | 0.65 |
| Support Vector regression-RBF kernel | 0.37 | 0.37 | 0.69 | 0.90 | 0.59 | 0.46 | 0.69 | | 0.85 | 0.61 | 0.62 | 0.14 | 0.82 |
| Decision tree | 0.24 | 0.30 | 0.44 | 0.81 | 0.15 | 0.13 | 0.67 | | 0.68 | 0.38 | 0.42 | 0.52 | 0.69 |
| Ridge regression | 0.17 | 0.25 | 0.55 | 0.89 | 0.46 | 0.33 | 0.73 | | 0.83 | - | - | - | - |
| Kernel Ridge regression- Linear kernel | 0.13 | 0.28 | 0.61 | 0.89 | 0.21 | 0.14 | 0.67 | | 0.68 | - | - | - | - |
| Kernel Ridge regression- Polynomial kernel | 0.29 | 0.34 | 0.52 | 0.87 | - | - | 0.51 | | 0.15 | 0.45 | 0.60 | - | - |
| Kernel Ridge regression- RBF kernel | 0.11 | 0.31 | 0.61 | 0.89 | 0.17 | 0.26 | 0.72 | | 0.43 | 0.56 | 0.49 | 0.21 | 0.81 |
| Random forests | 0.45 | 0.40 | 0.57 | 0.88 | - | - | - | | - | 0.54 | 0.71 | 0.13 | 0.81 |
| KNN | 0.37 | 0.27 | 0.50 | 0.84 | 0.32 | 0.36 | 0.63 | | 0.82 | 0.44 | 0.33 | 0.82 | 0.29 |
| Gaussian Process regression | 0.15 | 0.25 | 0.54 | 0.86 | 0.38 | 0.39 | 0.53 | | 0.73 | - | - | - | - |
| Neural nets | 0.47 | 0.37 | 0.09 | 0.89 | 0.37 | 0.42 | 0.56 | | 0.74 | - | - | - | - |

Group contribution

Group contribution method of representation has worked well for Heat of fusion but not as well for acid dissociation constant. We can observe high score for test case of Hfus across the board accompanied by relatively good validation case scores for the same property except for neural nets which performed poorly on the validation case but scored one of the highest values in test case. Prediction of Acid dissociation constant didn’t go that well. We can see scores lower than 0.5 across the board with the best performers being neural nets, random forests and Support Vector regression with RBF kernel. This does hint at the possibility that certain representation methods provide better prediction capabilities for certain properties.

Mol2vec

Mol2Vec was able to predict the Heat of fusion data much better than the Dissociation constant data. The methods which performed very well include support Vector regression, K Nearest neighbors regression, ridge Regression and Neural networks which is clearly visible through the high R2 values. However for the Dissociation constant data, the accuracy was pretty low and this is reflected by the low R2 values almost all of them are below 0.5, in some cases the hyper parameters could not be tuned to its potential given the computational restriction in terms of RAM available on disk. However the model can be better used to predict the values of heat of fusion and dissociation constant too (after tuning the parameters well). Also its more of a mathematical model rather than a chemical one, hence the accuracy on real life might not match with that shown by the R2 value.

Mordred

According to the R2 value, the Mordred representation has a relatively better prediction on pKa values using SVR RBF kernel and random forests. The mean squared value of these regression models is 4.2 and 3.2 respectively which indicates proper tuning of the regression model. The heat of fusion values is hard to predict according to diverge R2 value using Mordred descriptors (Table 3), nevertheless, the regression model might not be tuned to the best hyper parameters. This could be due to the fact that Mordred is better at characterizing molecular components and structure so that values like pKa could be determined from proton attaching or not attaching to the molecule. While the heat of fusion is largely affected by intermolecular interactions, studying a single molecular structure might not be able to capture such characteristics.

**Discussion and Conclusion**

So as we can see from the results of various models developed using Group contribution, Mol2vec and Mordred representations, we can see that out of the three the first two perform better on Heat of Fusion dataset while the last one performs the best on the dissociation constant.

One of the key learnings here is how important representation of data is to perform property prediction. The results section is a good example of results varying for prediction of pKa and Hfus depending on the type of representation used even though the same regression method was used.

In case of Mol2Vec, a lot of tuning of hyper parameters is required to get the desired results, which could be less than that for Group contribution representation. Also Mol2Vec is quite sophisticated to use unlike group contributions, however it would require less storage space as only the SMILES strings are required along with a pertained model. The same is the case for Mordred which also takes data from the SMILES representation.

Mordred descriptor calculator presents good acid dissociation constant prediction using proper tuned regression model, but less accurate with heat of fusion. Thus, it is necessary to choose the correct molecular representation of molecule structure, to achieve correct prediction of properties.

**Author Contributions**

The project comprised of two main categories of tasks. Coding and report writing. The split of tasks was even and is as follows.

Kunal Jalindar Gaikwad: Opted to work on Mol2vec representation and perform regression analysis on it. He also worked upon t-SNE visualization for Mol2vec representation. He also worked on composing the discussions and conclusions part of this report

Hrishikesh Arvind Talasila: Opted to work on Group Contribution representation and perform regression analysis on it. He also worked on deciding which regression methods shall be used in this project and provided the exoskeleton code used for regression for all representations which was then tailored by each member to suit the representation they had opted. He also worked on composing the methods and results part of this report.

Ying Tong Yue: Opted to work on Mordred representation and perform regression analysis on it. She also worked upon t-SNE visualization for Mordred and Group Contribution representation. She also worked on composing the abstract and introduction part of this report.

These are the links to the various Google collab notebooks used.

Mordred t-SNE Visualization and Regression: <https://colab.research.google.com/drive/1jqQ-K6DGMd89Kh1zMbQOqO-R77u5MIaV?usp=sharing>

Group Contributors Visualization:

<https://colab.research.google.com/drive/1f4hYDI9Wi-OFUbX2wvdHklhm0La8Vwhm?usp=sharing>

Group Contributors Regression:

<https://colab.research.google.com/drive/1uoOMMrue2N4OwpB953rKlnhpUYqxMYTC?usp=sharing>

Mol2vec Regression:

<https://colab.research.google.com/drive/1IQI7C6ZZ4dKh_vBuWld_hEoYsCz73VV5?usp=sharing>

Mol2vec t-SNE Visualization:

<https://colab.research.google.com/drive/1ddaR7LV6e6lOLG0SeDfwxugRLWg7hu0F?usp=sharing>

**References**

1. Engel, T. (2006). Basic overview of Chemoinformatics. *Journal of Chemical Information and Modeling*, *46*(6), 2267–2277. https://doi.org/10.1021/ci600234z
2. Alshehri, A. S., Tula, A. K., You, F., & Gani, R. (2021). Next generation pure component property estimation models: With and without machine learning techniques. *AIChE Journal*. https://doi.org/10.1002/aic.17469
3. Jaeger, S., Fulle, S., & Turk, S. (2018). Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, *58*(1), 27–35. https://doi.org/10.1021/acs.jcim.7b00616
4. Moriwaki, H., Tian, Y.-S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1). https://doi.org/10.1186/s13321-018-0258-y
5. Maaten, L. van der, & Hinton, G. H. (2018). Visualizing Data using t-SNE. *Journal of Machine Learning*, *9*. Retrieved from https://lvdmaaten.github.io/publications/papers/JMLR\_2008.pdf.