Chapter 1

Organic Compound Solubility

1.1 Introduction

At the molecular level, solubility is controlled by the energy balance of intermolecular forces between solute-solute, solvent-solvent and solute-solvent molecules. Intermolecular forces come in different strengths ranging from very weak induced dipole – induced dipole interactions to much stronger dipole-dipole forces (including the important special case, hydrogen bonding). Generally, Polar molecules dissolve in polar solvents (e.g. water, alcohols) and non-polar molecules in non-polar solvents (e.g. the hydrocarbon hexane). The polarity of organic molecules is determined by the presence of polar bonds1 due to electronegative atoms (e.g. N, O) in polar functional groups such as amines (-NH2) and alcohols (-OH).

The solubility of hydrocarbons in water are of special interest, given its application in fields like drug delivery, dyeing and the assessment of the environmental impact of a contaminant in the soil and groundwater. While it is certainly possible to determine the solubility of a molecule experimentally, it would certainly be more beneficial to be somehow able to correlate the solubility with the other easily obtainable properties of the molecule like the molecular structure, and general physical properties like the boiling point or enthalpy of fusion. The ability to predict the solubility only from the molecular structure is certainly extremely desirable, as it would help to determine

the water solubility of even unsynthesized molecules whose theoretical structure is known.

1.2 Related Work

Literature Survey Attempts to predict solubility of organic compounds have been made in the past. Huibers and Katritzky (1998) tried to correlate the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structures. This uses a 'best multiple-linear relationship' method to find the best set of descriptors. Klopman and Zhu (2001) tried a group contribution approach to predict the solubility, while Tolls et al. (2002) studied the aqueous solubility vs molecular size relationship.

Research gaps While the extant literature looks at our target objective of trying to correlate and predict the aqueous solubility of hydrocarbons, even going to the extent of using only the molecular structure, a large caveat is that the cited work uses linear relationships between the features. There has been no attempt made to determine new features given the 3D molecular structure of the compound.

Objective There are two objectives to the project. First, we will try and fit a predictive model to predict the aqueous solubility of hydrocarbons, using features that can be easily obtained, such as boiling point and heat of fusion. This study will assume that the compound has been synthesized and exists, and assumes that the solubility is related to the other physical properties of the said hydrocarbon.

The second objective is slightly ambitious. Given only the molecular structure of the compound, we will try and come up with a way to predict the solubility. This has the huge benefit of predicting the solubility of compounds which have not yet been synthesized, and if solubility is of prime concern (in cases like drug delivery), then getting an estimate of the solubility of the molecule even before the synthesis holds significant utility.

Chapter 2

Solubility Data

2.1 Data Source

Any attempt to build a model of such a nature requires extensive and verifiable amount of data. The widespread relevance of solubility data to many branches and disciplines of science, medicine, technology and engineering, and the difficulty of recovering solubility data from literature mandated the need for a centralized experimentally obtained solubility database. We used the IUPAC Solubility Project database (Kertes (1989)) for our data.

2.1.1 Solubility Data Series

For the purpose of our data, we decided to use the Solubility Data Series from IUPAC, specifically **Hydrocarbons with Water and Seawater**. The data was available in two parts, the first volume consisting of C5 to C7 hydrocarbons, and the second volume consisted of C8 to C36 Hydrocarbons. For the sake of uniformity, the McAuliffe Solubility at 298K was taken for all the compounds.

2.2 Features

For predicting the solubility, a total of 16 elementary physical properties were considered. They are as follows:

1. **MW**: Molecular Weight of the hydrocarbon
2. **CPG**: Specific heat of hydrocarbon in gas phase
3. **CPL**: Specific heat of hydrocarbon in liquid phase
4. **TB**: Boiling point
5. **TMelt**: Melting point
6. **TT**: Triple point
7. **TC**: Critical Temperature
8. **PC**: Critical Pressure
9. **VC**: Critical Volume
10. **RHOC**: Critical Density
11. **PSAT**: Saturation Pressure
12. **HVAP**: Enthalpy of vaporization
13. **HFUS**: Enthalpy of fusion
14. **HF**: Enthalpy of formation
15. **Vm**: Molecular Volume
16. **Visc**: Viscosity at 298K

These properties were obtained from the ***Thermo*** python package. The combination of features was decided upon by a combination of easiness of availability and intuitive impact on solubility. There were a lot of other parameters, but they were either empirical constants, or completely unrelated to solubility, or both.

2.3 Data Cleaning and Preprocessing

The raw data was collected and a *Pandas dataframe* was generated out of the data as follows:

	Name	MF	MW	CPG	CPL	TB	TMelt	TT	TC	PC	VC
0	1,3-Cyclopentadiene	C5H6	66.10114	1140.819957	1802.292349	314.15	182.05	182.05	500.0	4699960.0	0.000232
1	Cyclopentene	C5H8	68.11702	1194.705857	1960.689484	317.35	138.15	138.13	506.5	4800000.0	0.000245
2	2-Methyl-1,3-butadiene	C5H8	68.11702	1508.531750	1960.689484	307.15	127.15	127.15	483.3	3739997.0	0.000266
3	1,4-Pentadiene	C5H8	68.11702	1443.650821	1960.689484	299.05	125.15	125.15	478.0	3789555.0	0.000276
4	1-Pentyne	C5H8	68.11702	1507.232005	1960.689484	313.05	167.15	167.15	493.4	4053000.0	0.000278
5	Cyclopentane	C5H10	70.13290	1183.596267	2085.071702	322.35	179.15	179.70	511.7	4510000.0	0.000259
6	2-Methyl-2-butene	C5H10	70.13290	1497.059793	2085.071702	311.65	139.15	139.42	470.0	3420000.0	0.000299
7	3-Methyl-1-butene	C5H10	70.13290	1632.299430	2085.071702	293.25	104.90	104.72	452.7	3530000.0	0.000305
8	1-Pentene	C5H10	70.13290	1542.189146	2215.479533	303.15	108.15	108.15	464.8	3560000.0	0.000298
9	2-Pentene	C5H10	70.13290	1408.641678	2085.071702	310.05	127.55	121.80	475.0	3690000.0	0.000301

RHOC	PSAT	HVAP	HFUS	HF	Vm	Visc
284.918707	59226.323229	405076.739586	NaN	130800.0	0.000085	0.000253
278.028653	50440.182509	408584.752841	49326.878950	32930.0	0.000089	0.000332
256.079023	73379.808091	361991.935213	72375.450365	75750.0	0.000097	0.000251
246.800797	97975.483765	360428.662730	89845.386660	106380.0	0.000104	0.000195
245.025252	58084.265066	413392.656845	NaN	144350.0	0.000103	0.000353
270.783398	42319.067142	404672.441280	8697.772372	-77240.0	0.000095	0.000415
234.558194	62147.076804	388900.424471	108365.688571	-42550.0	0.000107	0.000206
230.019351	120187.537332	339283.484960	76426.327729	-28950.0	0.023570	0.000007
235.029826	84980.631649	361893.854967	84696.340804	-20920.0	0.000110	0.000213
232.999668	65970.039962	369640.183678	101378.953387	-28070.0	0.000108	0.000216

FIGURE 2.1: Raw feature Data Set

The data had certain missing values (NaN). This data required to be imputed. We used a technique called Iterative SVD (Singular Value Decomposition), described below

2.3.1 Iterative SVD Method (Troyanskaya et al. (2001))

In this method, we employ singular value decomposition to obtain a set of mutually orthogonal matrices that can be linearly combined to approximate the feature values of all compounds in the data set.

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$$

Once k most significant eigenvalues from V^T are selected, we estimate a missing value j in row i by first regressing this value against the k eigenvalues and then use the coefficients of the regression to reconstruct j from a linear combination of the k eigenvalues. The j^{th} value of row i and the j^{th} values of the k eigenvalues are not used in determining these regression coefficients.

It should be noted that SVD can only be performed on complete matrices; therefore we originally substitute row average for all missing values in matrix A, obtaining A . We then utilize an expectation maximization method to arrive at the final estimate, as follows. Each missing value in A is estimated using the above algorithm, and then the procedure is repeated on the newly obtained matrix, until the total change in the matrix falls below the empirically determined threshold of 0.01.

The dataset, after Iterative SVD, had the missing values imputed as given below. The matrix stopped updating after 4 iterations.

```
[IterativeSVD] Iter 1: observed MAE=16436.427256
[IterativeSVD] Iter 2: observed MAE=10700.613090
[IterativeSVD] Iter 3: observed MAE=1040.697022
[IterativeSVD] Iter 4: observed MAE=12.631030
```

	Name	MF	MW	CPG	CPL	TB	TMelt	TT	TC	PC	VC
0	1,3-Cyclopentadiene	C5H6	66.10114	1140.819957	1802.292349	314.15	182.05	182.05	500.0	4699960.0	0.000232
1	Cyclopentene	C5H8	68.11702	1194.705857	1960.689484	317.35	138.15	138.13	506.5	4800000.0	0.000245
2	2-Methyl-1,3-butadiene	C5H8	68.11702	1508.531750	1960.689484	307.15	127.15	127.15	483.3	3739997.0	0.000266
3	1,4-Pentadiene	C5H8	68.11702	1443.650821	1960.689484	299.05	125.15	125.15	478.0	3789555.0	0.000276
4	1-Pentyne	C5H8	68.11702	1507.232005	1960.689484	313.05	167.15	167.15	493.4	4053000.0	0.000278
5	Cyclopentane	C5H10	70.13290	1183.596267	2085.071702	322.35	179.15	179.70	511.7	4510000.0	0.000259
6	2-Methyl-2-butene	C5H10	70.13290	1497.059793	2085.071702	311.65	139.15	139.42	470.0	3420000.0	0.000299
7	3-Methyl-1-butene	C5H10	70.13290	1632.299430	2085.071702	293.25	104.90	104.72	452.7	3530000.0	0.000305
8	1-Pentene	C5H10	70.13290	1542.189146	2215.479533	303.15	108.15	108.15	464.8	3560000.0	0.000298
9	2-Pentene	C5H10	70.13290	1408.641678	2085.071702	310.05	127.55	121.80	475.0	3690000.0	0.000301

RHOC	PSAT	HVAP	HFUS	HF	Vm	Visc
284.918707	59226.323229	405076.739586	59596.026999	130800.0	0.000085	0.000253
278.028653	50440.182509	408584.752841	49326.878950	32930.0	0.000089	0.000332
256.079023	73379.808091	361991.935213	72375.450365	75750.0	0.000097	0.000251
246.800797	97975.483765	360428.662730	89845.386660	106380.0	0.000104	0.000195
245.025252	58084.265066	413392.656845	55025.113638	144350.0	0.000103	0.000353
270.783398	42319.067142	404672.441280	8697.772372	-77240.0	0.000095	0.000415
234.558194	62147.076804	388900.424471	108365.688571	-42550.0	0.000107	0.000206
230.019351	120187.537332	339283.484960	76426.327729	-28950.0	0.023570	0.000007
235.029826	84980.631649	361893.854967	84696.340804	-20920.0	0.000110	0.000213
232.999668	65970.039962	369640.183678	101378.953387	-28070.0	0.000108	0.000216

FIGURE 2.2: Imputed feature Data Set

2.3.2 Feature Normalization

The dataset, thus consists of 15 features, which we will refer to a $\mathbf{X}$, and the solubility vector, which is referred to as $\mathbf{Y}$. As a preprocessing step, the values in $\mathbf{X}$ were normalized to be between 0 and 1. We used RobustScaler for this purpose, as it is shown to deal the best with outliers.

Robust Scaler Normalization step:

$$\frac{x - Q_1(x)}{Q_3(x) - Q_1(x)}$$

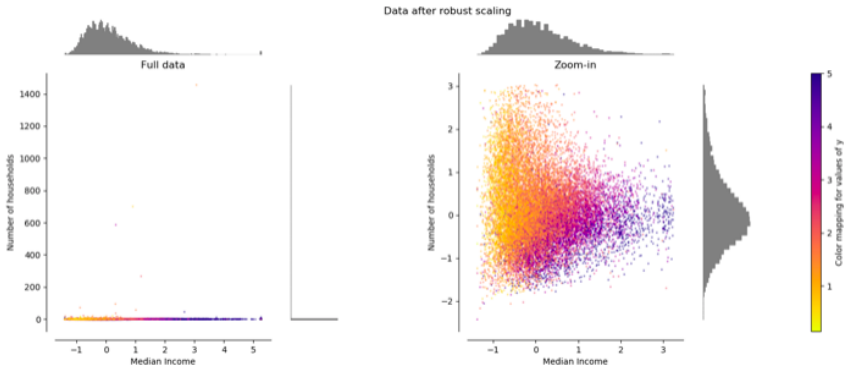


FIGURE 2.3: RobustScaler Performance

2.3.3 Conversion to a classification problem

The solubility values in $\mathbf{Y}$ are mostly values lying between 0 and 1, with the data heavily skewed towards 0. To make the prediction easier, the solubility values were bucketed into classes from 0 to 500, with 0 being the minimum and 500 being the maximum. This ensured enough fine tuned result, as well as allowed us to approach it as a classification problem which is easier to deal with.

Since the data was heavily skewed towards 0, class weights were assigned. More penalty was awarded to the optimization algorithm if the predicted class was 0, and the penalty reduced for higher classes.

Chapter 3

Baseline Model: Linear Regression

3.1 Requirement of a baseline model

Before delving into the machine learning aspects, we decided to look into the nature of the data and run some simple statistics on it. Multiple Linear Regression was selected as a baseline model for the accuracy and for examining the dependence of solubility on the chosen parameters.

3.2 Multiple Linear Regression

Code:

```
result = sm.ols(formula="SOL ~ MW + CPG + CPL + TB + TMelt + TT + TC + PC
    + VC + RHOC + PSAT + HVAP + HFUS + HF + Vm + Visc", data=df).fit()
result.summary()
```

FIGURE 3.1: Linear Regression Results

Dep. Variable:	SOL	**R-squared:**	0.320
Model:	OLS	**Adj. R-squared:**	0.241
Method:	Least Squares	**F-statistic:**	4.051
Date:	Thu, 02 Nov 2017	**Prob (F-statistic):**	2.71e-06
Time:	06:47:31	**Log-Likelihood:**	402.49
No. Observations:	155	**AIC:**	-771.0
Df Residuals:	138	**BIC:**	-719.2
Df Model:	16		
Covariance Type:	nonrobust		

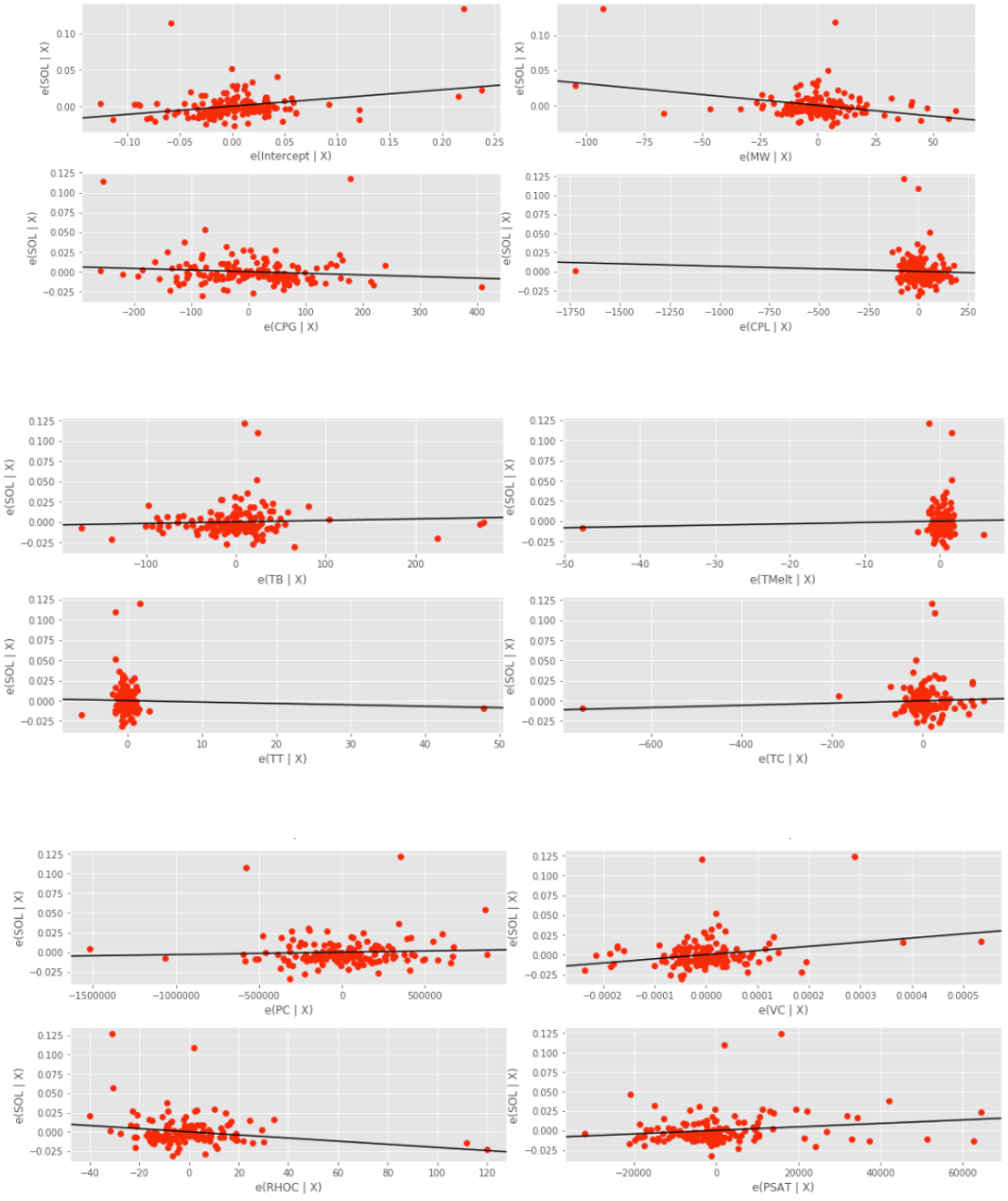
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1132	0.032	3.588	0.000	0.051	0.176
MW	-0.0003	7.99e-05	-3.849	0.000	-0.000	-0.000
CPG	-2.041e-05	1.6e-05	-1.280	0.203	-5.2e-05	1.11e-05
CPL	-6.709e-06	1.01e-05	-0.667	0.506	-2.66e-05	1.32e-05
TB	1.856e-05	2.9e-05	0.640	0.523	-3.87e-05	7.59e-05
TMelt	0.0002	0.000	0.417	0.677	-0.001	0.001
TT	-0.0002	0.000	-0.456	0.649	-0.001	0.001
TC	1.415e-05	2.18e-05	0.650	0.517	-2.89e-05	5.72e-05
PC	3.026e-09	4.91e-09	0.616	0.539	-6.69e-09	1.27e-08
VC	52.3788	17.509	2.992	0.003	17.759	86.999
RHOC	-0.0002	8.37e-05	-2.416	0.017	-0.000	-3.67e-05
PSAT	2.243e-07	1.09e-07	2.055	0.042	8.51e-09	4.4e-07
HVAP	-4.089e-08	3.12e-08	-1.309	0.193	-1.03e-07	2.09e-08
HFUS	5.16e-10	4.8e-08	0.011	0.991	-9.43e-08	9.54e-08
HF	3.47e-08	1.71e-08	2.023	0.045	7.92e-10	6.86e-08
Vm	-1.5877	0.812	-1.955	0.053	-3.194	0.018
Visc	-2.5083	3.933	-0.638	0.525	-10.286	5.269

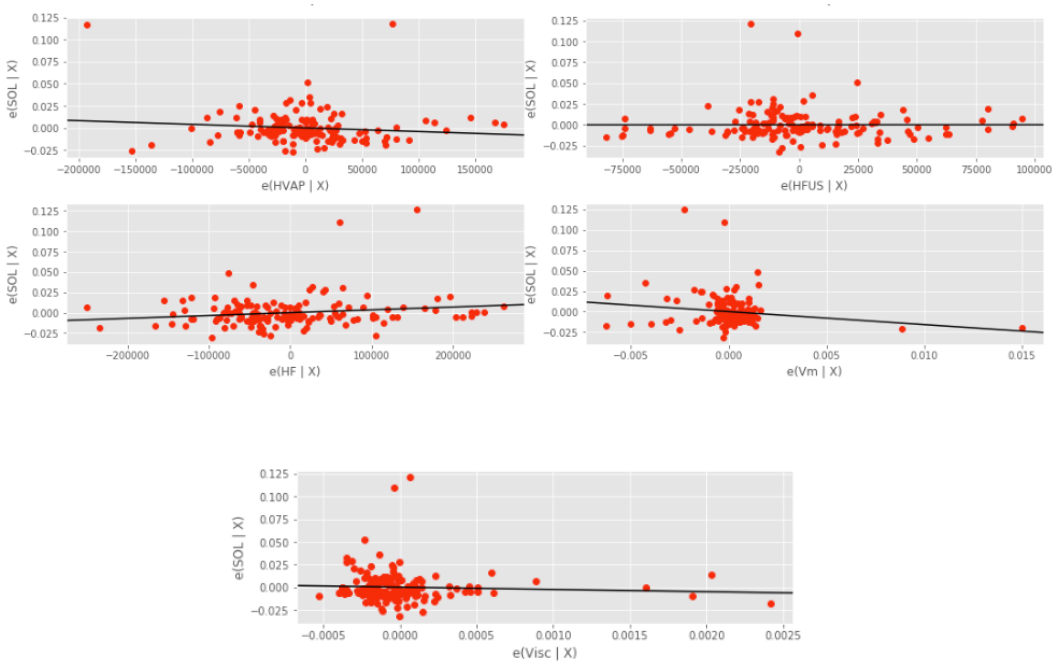
Omnibus:	151.448	**Durbin-Watson:**	1.908
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2834.611
Skew:	3.560	**Prob(JB):**	0.00
Kurtosis:	22.703	**Cond. No.**	3.26e+10

We can observe from the P value that the features MW, VC, RHOC, PSAT, HF and Vm are signifant, while the other features are either less significant or insignificant.

However, the R^2 value is only 0.320, which signals that the underlying data may not be linearly separable, infact, the system informs us that there is a lot of multi-collinearity involved as well.

FIGURE 3.2: OLS Partial Regression Plots





Now that it is established that the data is not linearly separable, we will move on to using a Neural Network to predict the solubility data.

Chapter 4

Artificial Neural Network

4.1 Requirement of the Neural Network

Artificial Neural Networks (ANN) are a sophisticated family of Machine Learning techniques which can represent strong non-linearity. The data, as described earlier, is not separable linearly and hence a method to represent the non-linearity is required.

4.2 Structure of the Neural Network

The network was built using *Keras*, and Google's *Tensorflow* backend. The exact number of nodes was arrived at using trial and error.

It consists of the input layer (16 nodes for the 16 features), followed by two hidden layers (10 nodes and 5 nodes respectively), and the final output layer. The first hidden layer has a *tanh* activation function, while the second hidden layer and the output layer have *relu* activation function. All the layers are fully connected.

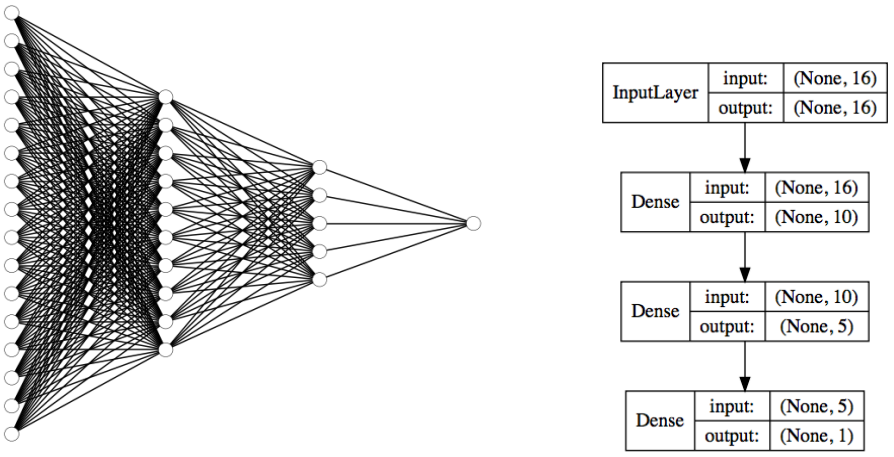
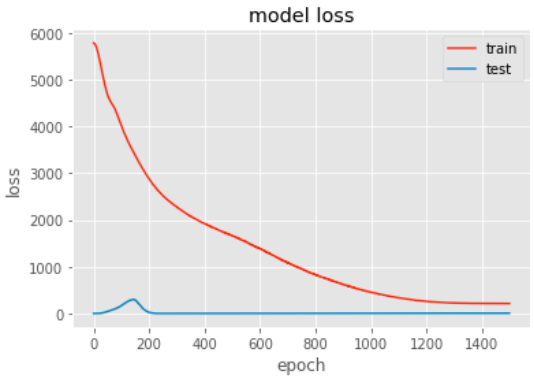


FIGURE 4.1: Structure of the ANN

4.3 Performance of the Network

The train:test split ratio was 9:1, and the network was trained for 1500 epochs with a batch size of 5. *RMSProp* was used as the optimizer and *Mean Squared Error* was used as the loss function.

FIGURE 4.2: Loss over time



The final performance metrics were as follows:

	Loss	Accuracy
Train	212.4339	0.2230
Test	5.3718	0.8750

TABLE 4.1: Network performance

4.4 Validation of the Model

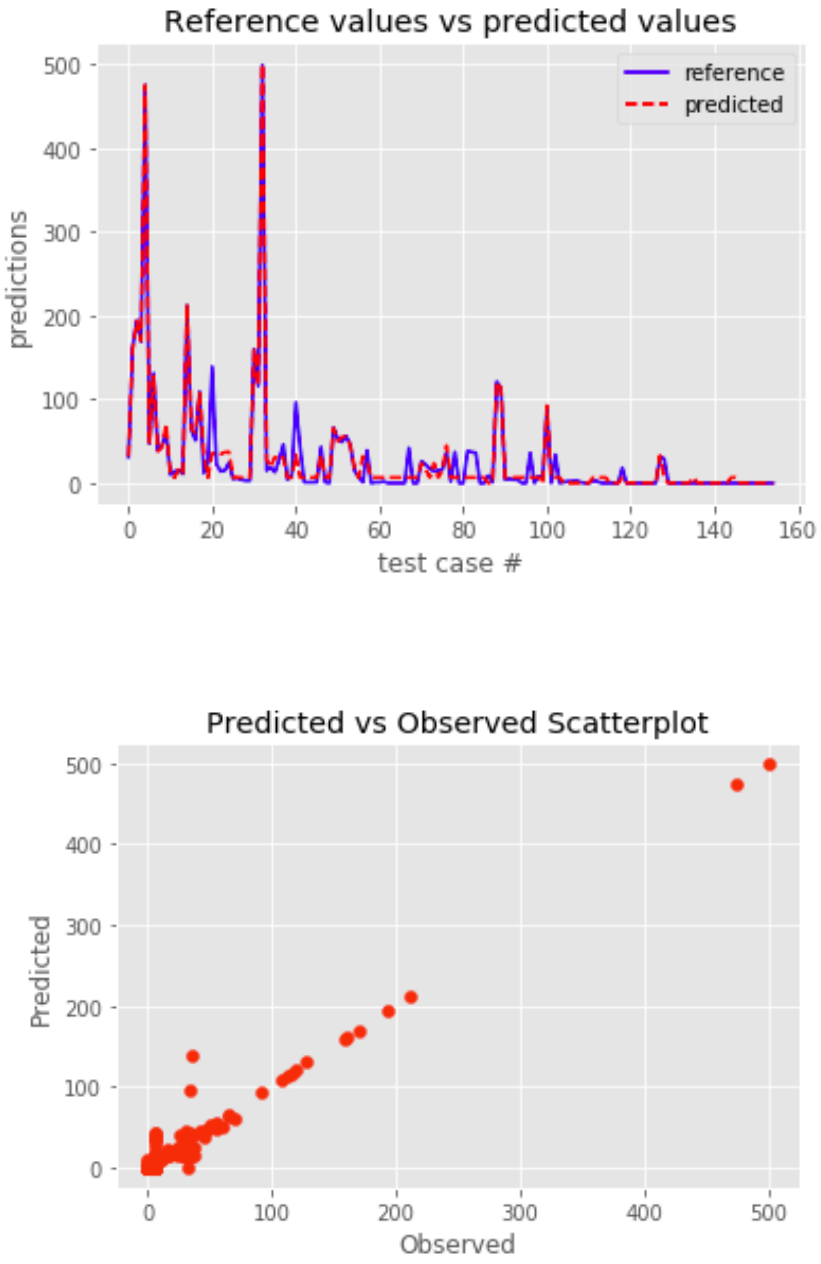


FIGURE 4.3: Validation of the model

Chapter 5

Conclusion and Further Work

5.1 Conclusion

It can be seen that the expectedly non-linear relationship between the physical properties and the aqueous solubility can be estimated within reasonable error using an ANN. The accuracy is 87.5%, which is a significant improvement over the Linear Model (32%). Since this is a data driven model, the predictive power of the Network will only improve with more data, since a small training dataset may not generalize well. However, the current model is a proof-of-concept of the predictive ability of the network.

5.2 Further Work

There is a lot of scope of improvement here. Firstly, not all physical properties are easy to obtain, especially ones like Saturation Pressure, Heat of Vaporization, etc. In fact, if getting these features were possible, then getting the solubility directly itself is easier than building a model and feeding obscure features to it.

A solution to this that we will be exploring comes from the field of Natural Language Processing.

5.2.1 Vector Space Model and 3D Molecules

Vector Space Model is an algebraic model used to represent text documents, and finds wide applications in the field of Natural Language Processing and Information Retrieval, where entire text documents are processed, and represented as a higher dimensional vector. There are a lot of different approaches to generate these vectors, like the bag of words model, cosine distance, latent dirichlet allocation to represent the document as a vector of topics, and the like. This is a powerful technique, because this vector representation can now be used as a feature set for machine learning applications.

Each molecule's 3D structure can be represented, in a standardized manner, as a text document (SDF file).

For example, this is what the 3D structure of Benzene looks like, as an SDF file:

```
12 12  0  0  0  0  0  0  0  0999 V2000
    3.2917    3.3940    0.2349 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9023    3.5389    0.2241 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8618    2.1207    0.1613 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0830    2.4105    0.1396 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0425    0.9922    0.0768 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6531    1.1372    0.0660 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.9292    4.2719    0.3006 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.4588    4.5296    0.2813 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.9429    2.0079    0.1699 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0019    2.5232    0.1310 H   0  0  0  0  0  0  0  0  0  0  0  0
    3.4861    0.0016    0.0196 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.0156    0.2592    0.0002 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  3  1  0  0  0  0
  1  7  1  0  0  0  0
  2  1  2  0  0  0  0
  2  8  1  0  0  0  0
  3  5  2  0  0  0  0
  3  9  1  0  0  0  0
  4  2  1  0  0  0  0
  4 10  1  0  0  0  0
```

```
  5  6  1  0  0  0  0
  5 11  1  0  0  0  0
  6  4  2  0  0  0  0
  6 12  1  0  0  0  0
M  END
```

If we can use document vector models upon such structure files, it is theoretically possible to use only the structure to predict solubility instead of the features. More importantly, it provides us the huge benefit of not having to actually synthesize a hydrocarbon in order to estimate the solubility. This opens up unforeseen opportunities in the estimation of physical parameters on simulated compounds, and drastically cut down on the time and resources involved in the synthesis of new compounds.

Bibliography

Huibers, P. D. T. and Katritzky, A. R. (1998). Correlation of the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structure. *Journal of Chemical Information and Computer Sciences*, 38(2):283–292.

Kertes, A. S. (1989). Hydrocarbons with water and seawater. *International Union of Pure and Applied Chemistry Analytical Chemistry Division Commission on Solubility Data*, 37,38.

Klopman, G. and Zhu, H. (2001). Estimation of the aqueous solubility of organic molecules by the group contribution approach. *Journal of Chemical Information and Computer Sciences*, 41(2):439–445. PMID: 11277734.

Tolls, J., Van Dijk, J., Verbruggen, E. J. M., Hermens, J. L. M., Loeprecht, B., and Schüürmann, G. (2002). Aqueous solubility-molecular size relationships: A mechanistic case study using c10- to c19-alkanes. *Journal of Physical Chemistry A*, 106(11):2760–2765. Cited By :47.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.