STAT project

Predicting Acute Aquatic Toxicity of Chemicals Towards D. Magna by Linear Regression on Molecular Descriptors

1 Introduction

1.1 Background and Motivation

Aquatic animals are vital aspects to many ecosystems and as such it is important to understand how the presence or introduction of certain chemicals can effect their environment. One particular topic of interest is the acute toxicity of certain chemicals to the organisms as short term exposure of these chemicals can kill the animals and cause major damage and adverse effects to the aquatic system [1]. As such it is of interest to know how toxic certain chemicals are in order to protect and monitor susceptible ecosystems.

A common measure of chemical toxicity to animals in aquatic systems is the Lethal Concentration 50 (LC50) which is the concentration of a material (usually in air, but in this case water) in which 50% of a group of test animals is expected to die when administered to a single exposure [2]. While this is a great way to quantify the toxicity of a chemical to aquatic animals, finding this concentration for all chemicals and for all animals is not only unreasonable, but also highly unethical. As such it is far more appropriate and of best interest to try and find a model that can accurately predict the LC50 concentration of a chemical given a set of parameters.

This brings us to the purpose of this report where we explore models that try to predict LC50 concentrations of chemicals based on their molecular descriptors.

1.2 Data set

The data set we will be using to compute the models contains the LC50 for 546 different chemicals towards *Daphnia Magna*. The variables in the data set are:

name	**description**	**units**
TPSA(Tot)	topological polar surface area	$\mathring{A}^2$
SAacc	Van der Waals surface aea of atoms that are acceptors of hydrogen bonds	$\mathring{A}^2$
H-050	# of hydrogen atoms bonded to heteroatoms	N/A
MLOGP	octanol-water partition coefficient	N/A
RDCHI	topological index that encodes information about molecular size and branching	N/A
GATS1p	molecular polarizability	$\mathring{A}^3$
nN	# of nitrogen atoms present in the molecule	N/A
C-040	# of carbon atoms in specific electron-poor groups	N/A
LC50	lethal concentration 50	$-LOG(mol/L)$

Table 1: Variables of the data set with description. Obtained from [3]

From the data set we are given the molecular descriptors: TPSA(tot), SAacc, H-050, MLOGP, RDCHI, GATS1p, nN, and C-040 of each of the chemicals. We will be using them as the explanatory variables to try and predict the response variable LC50.

1.3 Pre-Analysis

We start by visualizing the scatter plots between the response variable LC50 and explanatory variables respectively.

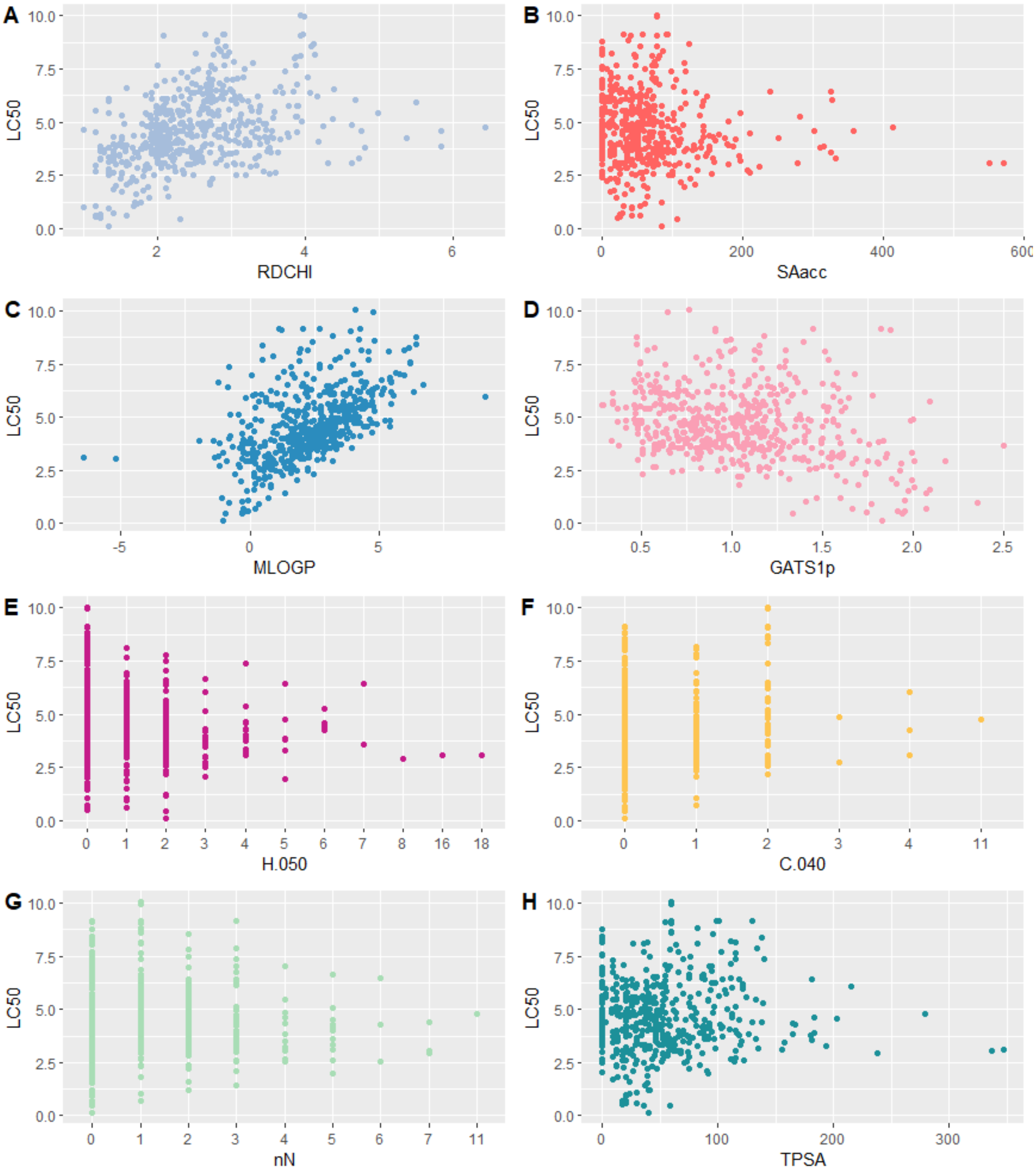


Figure 1: LC50 vs molecular descriptors scatter plot

1.3.1 RDCHI

When looking at RDCHI against LC50 (Fig 1.A) we can see a slight positive trend (p-cor = 0.1379211) suggesting that an increase in RDCHI will lead to an increase to LC50. Not only that but there also seems to be a slight concave pattern to the graph suggesting there might be a slight negative quadratic relationship between the two variables where the quadratic coefficient is small and negative, but the linear coefficient is large and positive. This should be taken into consideration when creating our models.

1.3.2 MLOGP

From the graph MLOGP (Fig 1.C) seems to have the strongest positive linear trend against LC50 (p-cor = 0.3124679) suggesting a linear positive relationship between the two variables to likely be significant.

1.3.3 GATS1p

GATS1p (Fig 1.D) shows a slight negative trend against LC50 (p-cor = -0.1554366) suggesting a linear negative relationship between the two variables to likely be significant.

1.3.4 TPSA, SAacc

TPSA (Fig 1.H) and SAacc (Fig 1.B) are quite interesting because while the graph shows no real pattern when visualized against LC50 the partial correlation says otherwise with TPSA having the highest positive partial correlation of 0.3659737 and SAacc having the lowest partial correlation of -0.3252569. This is interesting as both of the graphs show very similar visuals yet the partial correlation suggest each variable having the complete opposite relationship with LC50.

1.3.5 H-050, C-040, nN

From these graphs we can see that H-050 (Fig 1.E), nN (Fig 1.F), c-040 (Fig 1.G) behave like categorical variables, which makes sense as they represent the number of bonds of a certain type in each molecule. While treating them as categorical variables might make our models inaccurate or even invalid when used to predict molecules with a number of bonds that has not yet been seen in our data set used for training, we believe that viewing them as categorical variables will give us more accurate results.

2 Analysis

Since our categorical variables have a large number of levels, for compactness and clarity sake we define the following:

$$\begin{aligned} X_h^T &= \begin{bmatrix} x_{h,1} & x_{h,2} & \cdots & x_{h,8} & x_{h,16} & x_{h,18} \end{bmatrix} \\ X_c^T &= \begin{bmatrix} x_{c,1} & x_{c,2} & \cdots & x_{c,4} & x_{c,11} \end{bmatrix} \\ X_n^T &= \begin{bmatrix} x_{n,1} & x_{n,2} & \cdots & x_{n,7} & x_{n,11} \end{bmatrix} \end{aligned}$$

$$B_h = \begin{bmatrix} b_{h,1} \\ b_{h,2} \\ \vdots \\ b_{h,8} \\ b_{h,16} \\ b_{h,18} \end{bmatrix} \qquad B_c = \begin{bmatrix} b_{c,1} \\ b_{c,2} \\ \vdots \\ b_{c,4} \\ b_{c,11} \end{bmatrix} \qquad B_n = \begin{bmatrix} b_{n,1} \\ b_{n,2} \\ \vdots \\ b_{n,7} \\ b_{n,11} \end{bmatrix}$$

Where:

$$\begin{aligned} x_{h,i} &= \begin{cases} 1 & \text{if H-050} = i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2, ..., 8, 16, 18 \\ x_{c,i} &= \begin{cases} 1 & \text{if C-040} = i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2, ..., 4, 11 \\ x_{n,i} &= \begin{cases} 1 & \text{if nN} = i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2, ..., 7, 11 \\ b_{h,i} &= \text{ the slope parameter for the dummy variable } x_{h,i} \\ b_{c,i} &= \text{ the slope parameter for the dummy variable } x_{c,i} \\ b_{n,i} &= \text{ the slope parameter for the dummy variable } x_{n,i} \end{aligned}$$

We also define:

$$x_1 = \text{TPSA(tot)} \qquad x_2 = \text{SAacc} \qquad x_3 = \text{MLOGP} \qquad x_4 = \text{RDCHI} \qquad x_5 = \text{GATS1p}$$

2.1 Full Linear Model

Full linear model including all of the explanatory variables but without interactions is fitted as the starting point. This model will be our baseline model and future models will be using this as reference to try and get a better fit. After fitting the data we get the model:

$$y = 2.588161 + 0.02755x_1 - 0.015601x_2 + 0.470868x_3 + 0.47613x_4 - 0.518082x_5 + X_h^T B_h + X_n^T B_n + X_c^T B_c$$

Where:

$$\begin{aligned} B_h &= \begin{bmatrix} 0.0732 & -0.0826 & -0.0621 & 0.4806 & -0.0627 & -0.03754 & 2.0992 & -2.1366 & 2.3971 & 2.5271 \end{bmatrix}^T \\ B_n &= \begin{bmatrix} 0.280708 & -0.40056 & -0.47782 & -1.002553 & -1.119198 & -0.893993 & -1.467262 & -1.089241 \end{bmatrix}^T \\ B_c &= \begin{bmatrix} -0.173762 & 0.080182 & -1.161755.5 & 0.252138 & 0* \end{bmatrix}^T \end{aligned}$$

* parameter set to 0 due to singularities in our data

In this full model, only 51.8 percent of LC50 concentrations of chemical is explained by the explanatory variables with adjusted R2 value of 0.4929. In addition, the p-values of coefficients indicate that more than half of the explanatory variables are insignificant in explaining the LC50 concentrations. The variables which are significant in explaining the concentrations are, TPSA, SAacc, H-050 with 7 hydrogen atoms, MLOGP, RDCHI, GATS1p, and nN with 1-5 of nitrogen atoms. From the observation of this model, it can be concluded that the higher order model without the variables which have low explanatory powers is needed in order to get a better fit.

We also note that in this full model, all parameters of the C-040 explanatory variable is insignificant, however we are keeping it for the sake of having this model be our baseline full linear model that has all the explanatory variables for comparisons.

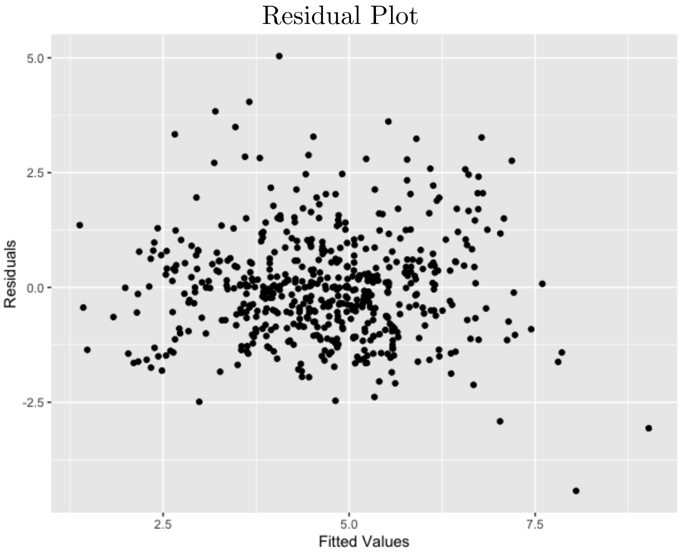


Figure 2: Residual v.s. Fitted value for the above full linear model

From observing the residual plot, it does not seem like there is a obvious pattern except that there are some outliers. Therefore, observing the residuals would not help deciding if the model is valid or not.

2.2 Optimized Simple Linear Model: Exhaustive Search

As introduced in Lab 8, the best subset selection algorithms (exhaustive search) compare all possible combinations of variables. Then, it will select the set of variables performing best for each size by certain criteria such as adjusted-R^2, $Mallows'$C$_p$, Akaike information criterion (AIC), and Sawa's Bayesian information criterion (BIC). Thus, we will conduct an exhaustive search next.

We cannot use regsubsets function from leaps package here since leaps only works with quantitative variables, and there are three categorical variables (H-050, nN, C-040) in our data. Thus, we use an alternative package called olsrr with function ols_step_best_subset(model) to perform the model selection. This package handles dummy variables of a categorical variable as a whole, so the dummy variables from the same category will be kept or dropped all together. The results are summarized in Table 2 and Table 3.

Best Subsets Regression

Model Index	**Predictors**
1	MLOGP
2	TPSA MLOGP
3	TPSA SAacc MLOGP
4	TPSA SAacc MLOGP nN
5	TPSA SAacc H.050 MLOGP nN
6	TPSA SAacc H.050 MLOGP RDCHI nN
7	TPSA SAacc H.050 MLOGP RDCHI GATS1p nN
8	TPSA SAacc H.050 MLOGP RDCHI GATS1p nN C.040

Table 2: Variables included in best model of each size selected by exhaustive search

Subsets Regression Summary

Model	Adj R-Square	C_p	AIC	SBIC
1	0.2855	224.5197	1926.8266	376.0294
2	0.3860	117.4504	1845.0205	294.4650
3	0.4441	56.1603	1791.7542	241.6118
4	0.4680	24.2202	1775.6389	211.9029
5	0.4785	4.8742	1774.4280	193.1055
6	0.4832	1.0270	1770.4712	189.3298
7	0.4933	-8.4400	1760.5969	179.8614
8	0.4929	-10.0000	1766.8753	178.3435

Table 3: Model selection criteria summary

From the summary, we can clearly see that the reduced model 7 with variables TPSA, SAacc, H.050, ML0GP, RDCHI, GATS1p, nN is the best model among all eight best models for the following reasons:

- It has the largest Adjusted R^2 0.4933. This reflects the percentage of variance in data that can be explained by the model is the highest after adjusting number of predictors.
- Its absolute C_p value 8.44 is closest to number of β. This shows that model 7 is relatively precise and unbiased.
- It has the smallest AIC 1760.5969. Smaller AIC, better the model.
- Other statistics, such as SBIC, in the table and R output can all be used for model selection. For simplicity, we choose not to analyze all of them here.

To conclude, the best model with only linear combination of explanatory variables selected by exhaustive search is:

$$y = 2.586791 + 0.027782x_1 + -0.015919x_2 + 0.476436x_3 + 0.471957x_4 - 0.52744x_5 + X_h^T B_h + X_n^T B_n$$

Where:

$$B_h = \begin{bmatrix}0.0598 & -0.0878 & -0.1156 & 0.4805 & -0.0926 & 0.0729 & 2.0073 & -2.1890 & 2.6369 & 2.6665\end{bmatrix}^T$$

$$B_n = \begin{bmatrix}0.254137 & -0.392736 & -0.488275 & 1.003452 & -1.087557 & -0.902561 & -1.548572 & -0.947835\end{bmatrix}^T$$

This model is the same as the result we found by trying and reducing full model in section 2.1, where we found the C-040 explanatory variable to be insignificant. The result is consistent.

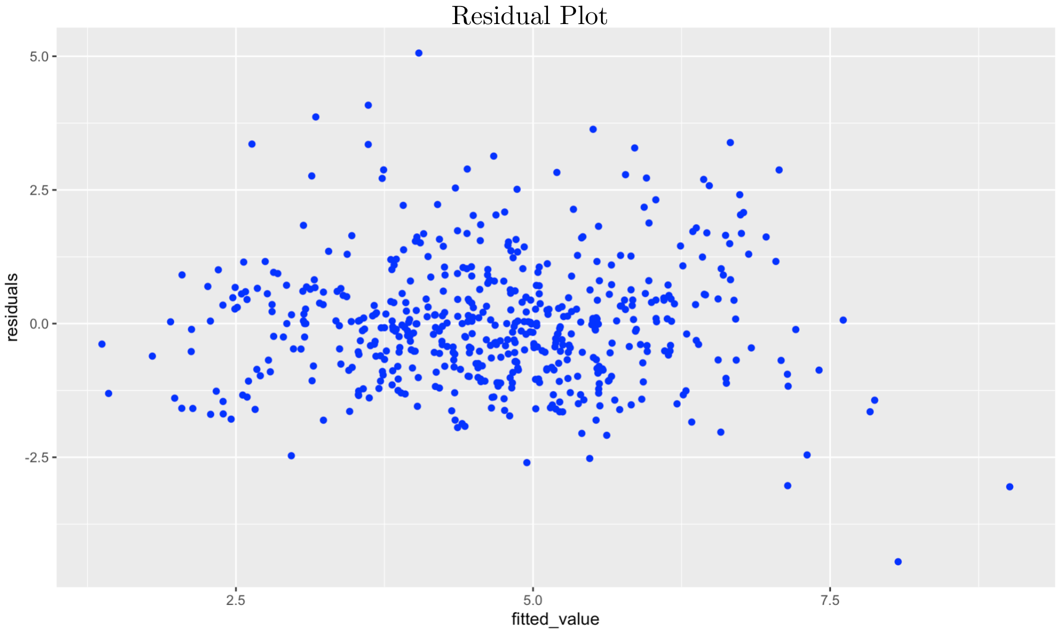


Figure 3: Residual v.s. Fitted value for the above selected model 7

There are two ways we can say about this residual plot (figure 2). If we look at the first sight, although there are some outliers, data are randomly and evenly distributed with no obvious pattern that break homoscedasticity and linear assumptions. If we take a closer look, we can also say there may be a non-linear pattern since more data points are above 0 residual than below 0. The shape looks slightly arched, but definitely this is not an obvious pattern. However, this still give us a hint that we may want to add some polynomial terms next.

2.3 Polynomial Regression

Building off of our optimized simplified linear model, we also considered the possibility of adding polynomial terms to achieve a better fit to the data. We used the simple algorithm of continually adding increasing polynomial terms for each of the continuous explanatory variables (in this case they were RDCHI, SAacc, MLOGP, GATS1p, and TPSA) until we reach an insignificant parameter and removing said insignificant parameter. After running through this algorithm and getting our model, we found that the H-050 categorical variable became insignificant and so we removed it. Despite this reducing our adjusted R^2 value, albeit by an insignificant amount ($\Delta = 0.0002$), the removal of the variable helped improve our other information criterias such as AIC ($\Delta = 9.029096$) and Cp ($\Delta = 8.349001$). This gave us the model of:

$$y = 2.474695 + 0.026110x_1 - 0.013760x_2 + 0.422836x_3 + 0.033874(x_3 - \overline{\text{MLOGP}})^2 \\ + 0.604753x_4 - 0.189366(x_4 - \overline{\text{RDCHI}})^2 - 0.576434x_5 + X_n^T B_n$$

Where:

$$B_n = \begin{bmatrix} 0.205405 & -0.497253 & -0.663693 & -1.080122 & -1.303988 & -0.808131 & -2.439560 & 0.564320 \end{bmatrix}^T$$

Note: we subtract the mean from the quadratic terms to control for collinearity

One thing of note here is the fact that as speculated in our pre-analysis (section 1.3.1) RDCHI does seem to have a negative quadratic relationship with LC50 with a small negative quadratic term but a larger positive linear term.

From this model:

- We get an adjusted R^2 value of 0.5022 compared to 0.4933 from model 2.2
- We have a Cp value of -7.75908 compared to -8.44 from model 2.2 while also only requiring 16 parameters compared to 24
- We get an AIC value of 1743.214 compared to 1760.5969 from model 2.2

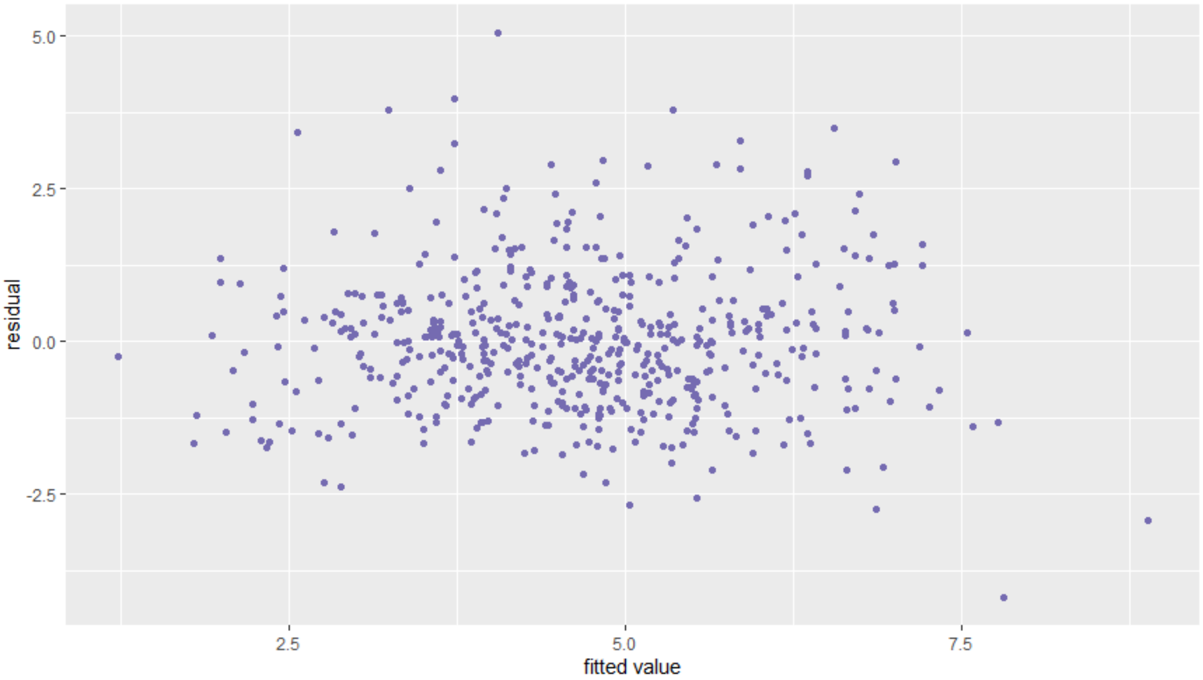


Figure 4: residual V.S. fitted value for model 2.3

Looking at the residual vs fitted, aside from the two extremely low points on the far right of the graph and the one unusually high residual in the middle, the graph is mostly random and does not have any major patterns. The unusual points could simply be outliers in our data and so they should not be enough evidence to invalidate the model.

2.4 Interaction Model

In addition, we fitted a model without quadratic term but included some interaction terms. We tried adding some new interaction terms to the full model. The method we used to find the best interaction model was backward model selection. We first added all possible interaction terms to our best simple linear model, and rejected the interaction terms with relatively high p-value. By using this selection, we found that

$TPSA * GATS1p$, GATS1p*C.040 and TPSA*MLOGP are the three interaction terms we should add to our model and also rejected H.050 from the model.

We chose TPSA*GATS1p, GATS1p*C.040 and TPSA*MLOGP to be the interaction terms for the following reasons:

- It increases our Adjusted R^2 from 0.4933 to 0.5149.
- It decreases the AIC from 1760.5969 to 1731.742, which is the smallest we can get in interaction models.

To conclude, the best model with interaction variables selected by backward model selection is:

$$y = 2.819 + 0.01612x_1 - 0.01575x_2 + 0.5242x_3 + 0.4373x_4 - 0.7769x_5 + 0.01238x_1x_5 - 0.001083x_1x_3 + X_n^T B_n + X_c^T B_c + x_5 X_c^T B_{5c}$$

Where:

$$B_n = \begin{bmatrix} 0.2783 & -0.2882 & -0.4294 & -0.8713 & -1.137 & -0.6211 & -2.296 & 2.931 \end{bmatrix}^T$$
$$B_c = \begin{bmatrix} 0.3537 & 3.297 & 142.5 & -3.002 & 0* \end{bmatrix}^T$$
$$B_{5c} = \begin{bmatrix} -0.4494 & -2.837 & -95.6 & 2.004 & 0* \end{bmatrix}^T$$

* parameter set to 0 due to singularities in our data

After adding these three terms, we first check the residual plot of it, which shows that it's plausible to add these interact terms. The residual plot of our interaction model is shown below:

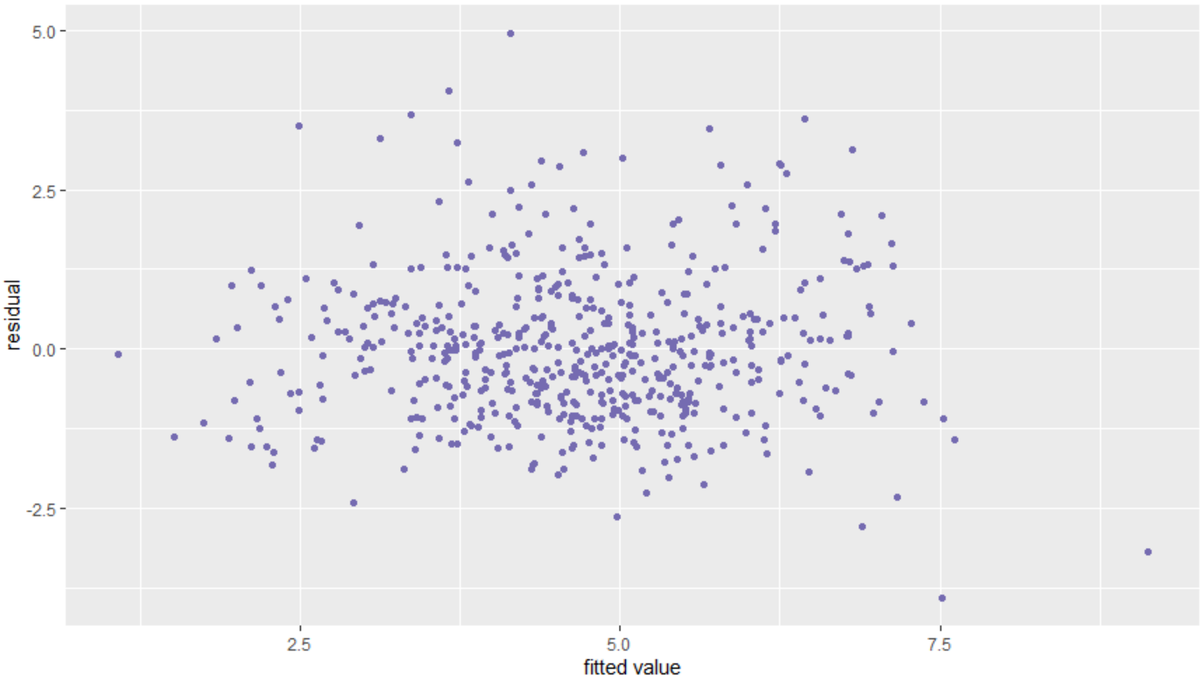


Figure 5: Residual plot of interaction model

From the residual plot, we can see that the graph is very similar to the residual plot for model2.3, which doesn't show any pattern, the points are evenly plotted to each side of the horizontal line.

2.5 Interaction and Polynomial Model

After adding polynomial terms and interaction terms to our model separately, it is sensible for us to try to develop a model that includes both interaction terms and polynomial terms. The easiest way to create this model would be to add the interactions from our interaction model directly to our polynomial model. Since the C-040 explanatory variable was used in the interaction model we will also try to include it in this model. After fitting the model we found that the quadratic term for MLOGP became insignificant and so we removed it, leaving us with the model of:

$$y = 2.691 + 0.01459x_1 - 0.01496x_2 + 0.49x_3 + 0.5709x_4 - 0.1599(x_4 - \overline{\text{RDCHI}})^2 - 0.7672x_5 + 0.0127x_1x_5 - 0.001055x_1x_3 + X_n^T B_n + X_c^T B_c + x_5 X_c^T B_{5c}$$

Where:

$$B_n = \begin{bmatrix}0.2445 & -0.3594 & -0.5461 & -0.9471 & -1.239 & -0.7709 & -2.425 & -1.243\end{bmatrix}^T$$
$$B_c = \begin{bmatrix}0.3096 & 3.461 & 95.08 & -1.195 & 0^*\end{bmatrix}^T$$
$$B_{5c} = \begin{bmatrix}-0.4032 & -2.984 & 64.04 & -0.9057 & 0^*\end{bmatrix}^T$$

* parameter set to 0 due to singularities in our data

Note: we are once again subtracting the mean of the data from the quadratic term to control for collinearity

From this model we get the following we have:

- An adjusted R^2 of 0.5256
- An AIC of 1725.568
- A Cp of -34.63863 compared to our 27 parameters in our model

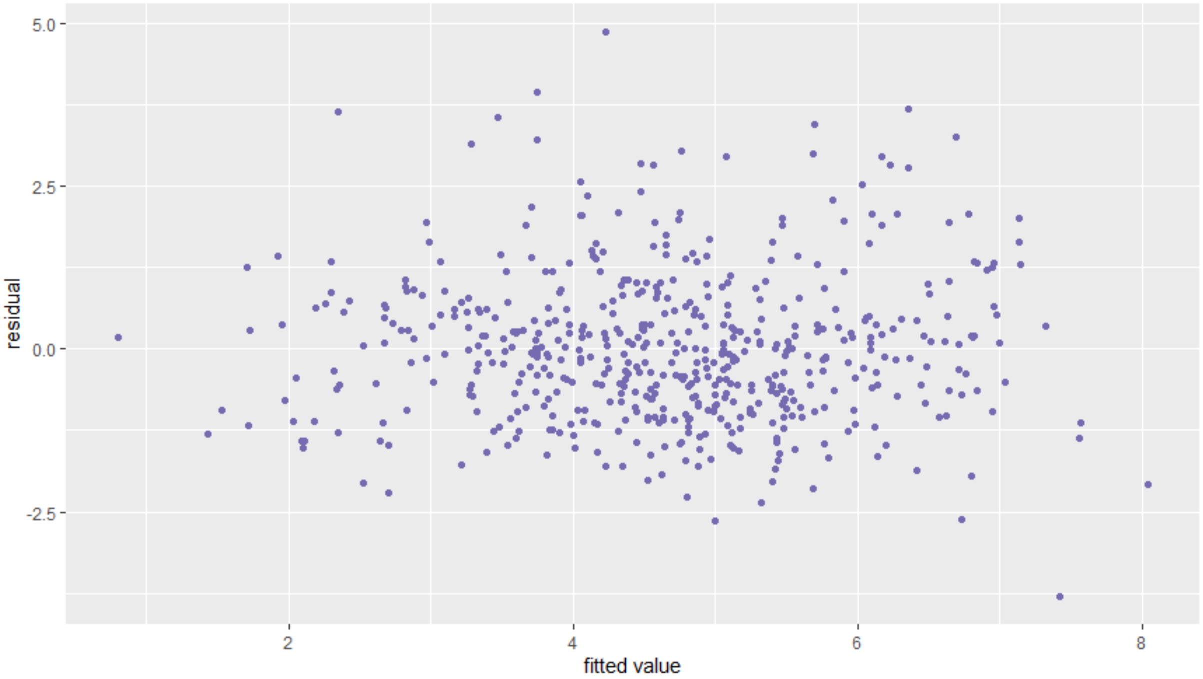


Figure 6: Residual plot of polynomial and interaction model

From this residual vs fitted value graph we can see that, similar to the residual graphs of model 2.3 and 2.4, there are no major patterns and that the residuals are mostly random and scattered around 0, reassuring us that this model is not invalid

2.6 Cross Validation

3-fold cross validation was carried out in order to compare the performance of 5 models mentioned above, and to see which model performs the best in predicting LC50 concentrations.

Procedure of the test is the following:
1.Randomly and equally split the samples into 3 groups
2.Take one group as test set and the other 2 sets as training set
3.Fit the model with the training set and find RMSE values for the test set
4.Repeat steps 2 and 3 for 3 times and take the average of the RMSE values

Model	RMSE	adj R^2	AIC	Cp	# of β
Full linear	1.2830	0.4929	1766.8753	-10	29
Optimized simple linear	1.2520	0.4933	1760.5969	-8.44	24
Polynomial	1.2250	0.5022	1743.214	-7.76	16
Interaction	1.2013	0.5149	1760.5969	-29.29	26
Interaction and Polynomial	1.2005	0.5256	1725.568	-34.63863	27

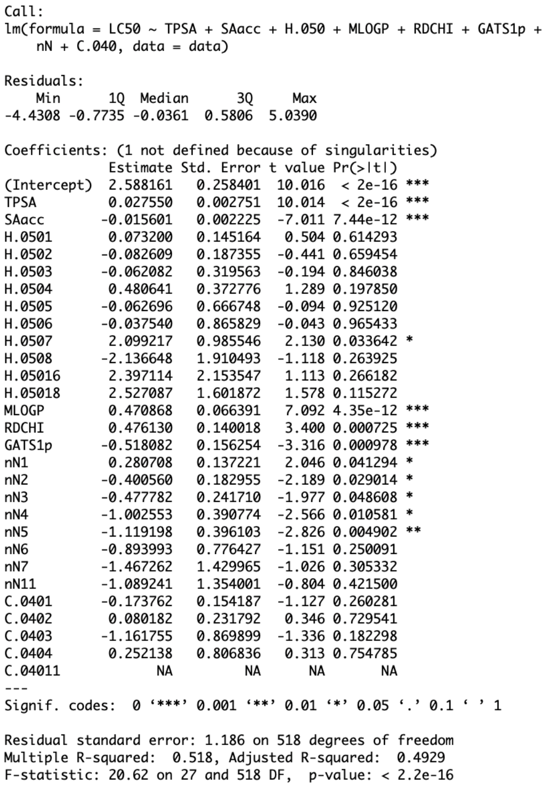
Table 4: RMSE values from 3-fold cross validation + model summary

The result above shows us that in terms of RMSE, the interaction and polynomial model had the best performance. Though this model does have the second most parameters and not the best Cp value when compared to the of parameters, it does has the highest adjusted R^2 value as well as the lowest AIC value out of all of the 5 models, making it the best bet for predicting LC50 of chemicals.

3 Conclusion

To conclude, our journey of finding the best model to predict the lethal concentration 50 (LC50) of chemicals based on a set of molecular descriptors started from fitting full linear model, which then led to trying exhaustive search algorithm before adding polynomial and interaction terms. We end our journey by applying cross validation, which reveals that the model with both interaction and polynomial terms we developed is the best model we can find so far. We believe that further research on the relationship between lethal concentration 50 and chemicals could be done to develop an even better fitted model.

Appendix



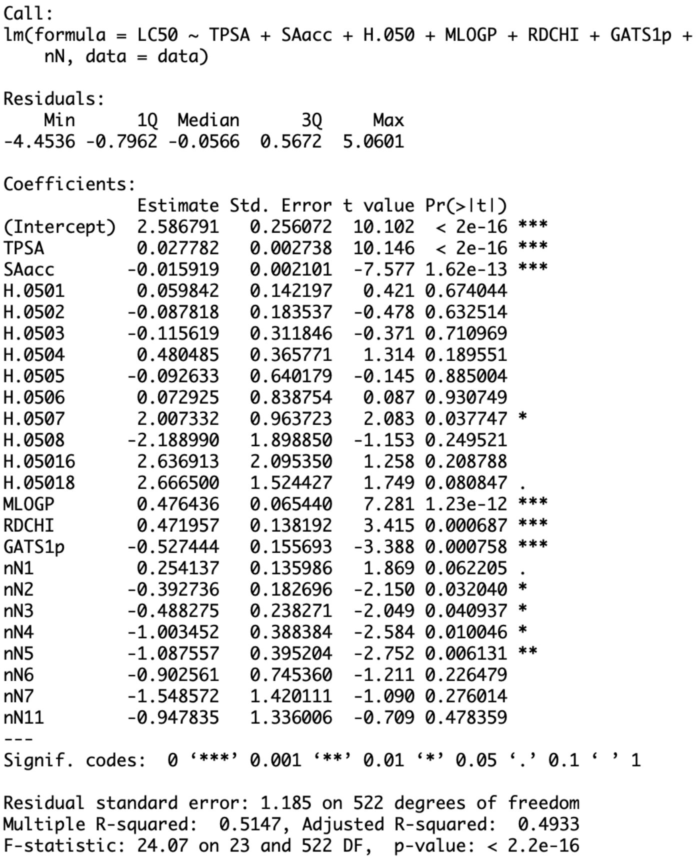
```
Call:
lm(formula = LC50 ~ TPSA + SAacc + H.050 + MLOGP + RDCHI + GATS1p +
    nN + C.040, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4308 -0.7735 -0.0361  0.5806  5.0390 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.588161   0.258401  10.016  < 2e-16 ***
TPSA         0.027550   0.002751  10.014  < 2e-16 ***
SAacc       -0.015601   0.002225  -7.011 7.44e-12 ***
H.0501       0.073200   0.145164   0.504 0.614293    
H.0502      -0.082609   0.187355  -0.441 0.659454    
H.0503      -0.062082   0.319563  -0.194 0.846038    
H.0504       0.480641   0.372776   1.289 0.197850    
H.0505      -0.062696   0.666748  -0.094 0.925120    
H.0506      -0.037540   0.865829  -0.043 0.965433    
H.0507       2.099217   0.985546   2.130 0.033642 *  
H.0508      -2.136648   1.910493  -1.118 0.263925    
H.05016      2.397114   2.153547   1.113 0.266182    
H.05018      2.527087   1.601872   1.578 0.115272    
MLOGP        0.470868   0.066391   7.092 4.35e-12 ***
RDCHI        0.476130   0.140018   3.400 0.000725 ***
GATS1p      -0.518082   0.156254  -3.316 0.000978 ***
nN1          0.280708   0.137221   2.046 0.041294 *  
nN2         -0.400560   0.182955  -2.189 0.029014 *  
nN3         -0.477782   0.241710  -1.977 0.048608 *  
nN4         -1.002553   0.390774  -2.566 0.010581 *  
nN5         -1.119198   0.396103  -2.826 0.004902 ** 
nN6         -0.893993   0.776427  -1.151 0.250091    
nN7         -1.467262   1.429965  -1.026 0.305332    
nN11        -1.089241   1.354001  -0.804 0.421500    
C.0401      -0.173762   0.154187  -1.127 0.260281    
C.0402       0.080182   0.231792   0.346 0.729541    
C.0403      -1.161755   0.869899  -1.336 0.182298    
C.0404       0.252138   0.806836   0.313 0.754785    
C.04011            NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.186 on 518 degrees of freedom
Multiple R-squared:  0.518,	Adjusted R-squared:  0.4929 
F-statistic: 20.62 on 27 and 518 DF,  p-value: < 2.2e-16
```

Figure 7: Section 2.1 full model R output



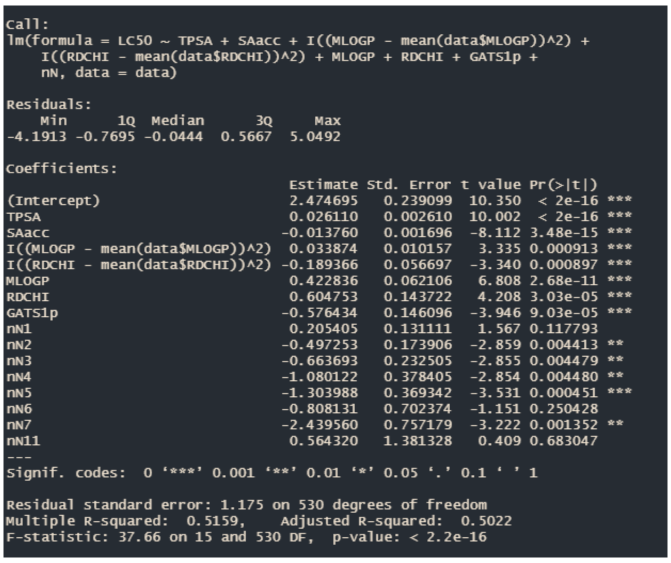
```
Call:
lm(formula = LC50 ~ TPSA + SAacc + H.050 + MLOGP + RDCHI + GATS1p +
    nN, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4536 -0.7962 -0.0566  0.5672  5.0601 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.586791   0.256072  10.102  < 2e-16 ***
TPSA         0.027782   0.002738  10.146  < 2e-16 ***
SAacc       -0.015919   0.002101  -7.577 1.62e-13 ***
H.0501       0.059842   0.142197   0.421 0.674044    
H.0502      -0.087818   0.183537  -0.478 0.632514    
H.0503      -0.115619   0.311846  -0.371 0.710969    
H.0504       0.480485   0.365771   1.314 0.189551    
H.0505      -0.092633   0.640179  -0.145 0.885004    
H.0506       0.072925   0.838754   0.087 0.930749    
H.0507       2.007332   0.963723   2.083 0.037747 *  
H.0508      -2.188990   1.898850  -1.153 0.249521    
H.05016      2.636913   2.095350   1.258 0.208788    
H.05018      2.666500   1.524427   1.749 0.080847 .  
MLOGP        0.476436   0.065440   7.281 1.23e-12 ***
RDCHI        0.471957   0.138192   3.415 0.000687 ***
GATS1p      -0.527444   0.155693  -3.388 0.000758 ***
nN1          0.254137   0.135986   1.869 0.062205 .  
nN2         -0.392736   0.182696  -2.150 0.032040 *  
nN3         -0.488275   0.238271  -2.049 0.040937 *  
nN4         -1.003452   0.388384  -2.584 0.010046 *  
nN5         -1.087557   0.395204  -2.752 0.006131 ** 
nN6         -0.902561   0.745360  -1.211 0.226479    
nN7         -1.548572   1.420111  -1.090 0.276014    
nN11        -0.947835   1.336006  -0.709 0.478359    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.185 on 522 degrees of freedom
Multiple R-squared:  0.5147,	Adjusted R-squared:  0.4933 
F-statistic: 24.07 on 23 and 522 DF,  p-value: < 2.2e-16
```

Figure 8: Section 2.2 reduced model R output



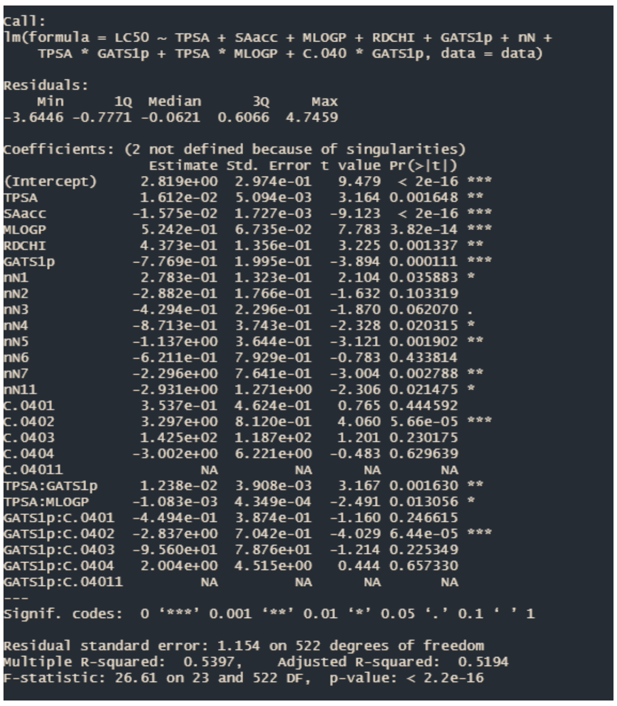
```
Call:
lm(formula = LC50 ~ TPSA + SAacc + I((MLOGP - mean(data$MLOGP))^2) +
    I((RDCHI - mean(data$RDCHI))^2) + MLOGP + RDCHI + GATS1p +
    nN, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1913 -0.7695 -0.0444  0.5667  5.0492

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      2.474695   0.239099  10.350  < 2e-16 ***
TPSA                             0.026110   0.002610  10.002  < 2e-16 ***
SAacc                           -0.013760   0.001696  -8.112 3.48e-15 ***
I((MLOGP - mean(data$MLOGP))^2)  0.033874   0.010157   3.335 0.000913 ***
I((RDCHI - mean(data$RDCHI))^2) -0.189366   0.056697  -3.340 0.000897 ***
MLOGP                            0.422836   0.062106   6.808 2.68e-11 ***
RDCHI                            0.604753   0.143722   4.208 3.03e-05 ***
GATS1p                          -0.576434   0.146096  -3.946 9.03e-05 ***
nN1                              0.205405   0.131111   1.567 0.117793
nN2                             -0.497253   0.173906  -2.859 0.004413 **
nN3                             -0.663693   0.232505  -2.855 0.004479 **
nN4                             -1.080122   0.378405  -2.854 0.004480 **
nN5                             -1.303988   0.369342  -3.531 0.000451 ***
nN6                             -0.808131   0.702374  -1.151 0.250428
nN7                             -2.439560   0.757179  -3.222 0.001352 **
nN11                             0.564320   1.381328   0.409 0.683047
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.175 on 530 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.5022
F-statistic: 37.66 on 15 and 530 DF,  p-value: < 2.2e-16
```

Figure 9: Section 2.3 polynomial model R output



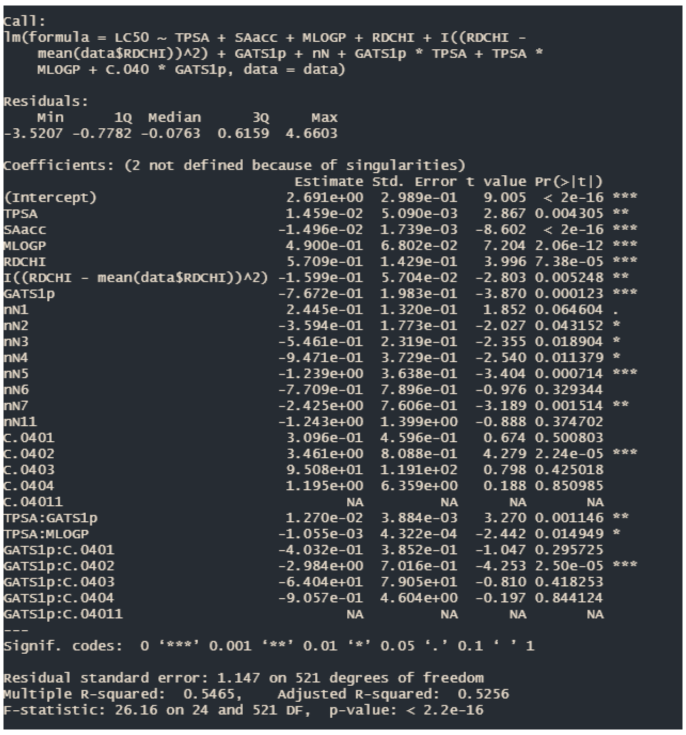
```
Call:
lm(formula = LC50 ~ TPSA + SAacc + MLOGP + RDCHI + GATS1p + nN +
    TPSA * GATS1p + TPSA * MLOGP + C.040 * GATS1p, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6446 -0.7771 -0.0621  0.6066  4.7459

Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.819e+00  2.974e-01   9.479  < 2e-16 ***
TPSA            1.612e-02  5.094e-03   3.164 0.001648 **
SAacc          -1.575e-02  1.727e-03  -9.123  < 2e-16 ***
MLOGP           5.242e-01  6.735e-02   7.783 3.82e-14 ***
RDCHI           4.373e-01  1.356e-01   3.225 0.001337 **
GATS1p         -7.769e-01  1.995e-01  -3.894 0.000111 ***
nN1             2.783e-01  1.323e-01   2.104 0.035883 *
nN2            -2.882e-01  1.766e-01  -1.632 0.103319
nN3            -4.294e-01  2.296e-01  -1.870 0.062070 .
nN4            -8.713e-01  3.743e-01  -2.328 0.020315 *
nN5            -1.137e+00  3.644e-01  -3.121 0.001902 **
nN6            -6.211e-01  7.929e-01  -0.783 0.433814
nN7            -2.296e+00  7.641e-01  -3.004 0.002788 **
nN11           -2.931e+00  1.271e+00  -2.306 0.021475 *
C.0401          3.537e-01  4.624e-01   0.765 0.444592
C.0402          3.297e+00  8.120e-01   4.060 5.66e-05 ***
C.0403          1.425e+02  1.187e+02   1.201 0.230175
C.0404         -3.002e+00  6.221e+00  -0.483 0.629639
C.04011                NA         NA      NA       NA
TPSA:GATS1p     1.238e-02  3.908e-03   3.167 0.001630 **
TPSA:MLOGP     -1.083e-03  4.349e-04  -2.491 0.013056 *
GATS1p:C.0401  -4.494e-01  3.874e-01  -1.160 0.246615
GATS1p:C.0402  -2.837e+00  7.042e-01  -4.029 6.44e-05 ***
GATS1p:C.0403  -9.560e+01  7.876e+01  -1.214 0.225349
GATS1p:C.0404   2.004e+00  4.515e+00   0.444 0.657330
GATS1p:C.04011         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.154 on 522 degrees of freedom
Multiple R-squared:  0.5397,    Adjusted R-squared:  0.5194
F-statistic: 26.61 on 23 and 522 DF,  p-value: < 2.2e-16
```

Figure 10: Section 2.4 polynomial model R output



```
Call:
lm(formula = LC50 ~ TPSA + SAacc + MLOGP + RDCHI + I((RDCHI -
    mean(data$RDCHI))^2) + GATS1p + nN + GATS1p * TPSA + TPSA *
    MLOGP + C.040 * GATS1p, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5207 -0.7782 -0.0763  0.6159  4.6603

Coefficients: (2 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.691e+00  2.989e-01   9.005  < 2e-16 ***
TPSA                            1.459e-02  5.090e-03   2.867 0.004305 **
SAacc                          -1.496e-02  1.739e-03  -8.602  < 2e-16 ***
MLOGP                           4.900e-01  6.802e-02   7.204 2.06e-12 ***
RDCHI                           5.709e-01  1.429e-01   3.996 7.38e-05 ***
I((RDCHI - mean(data$RDCHI))^2) -1.599e-01  5.704e-02  -2.803 0.005248 **
GATS1p                         -7.672e-01  1.983e-01  -3.870 0.000123 ***
nN1                             2.445e-01  1.320e-01   1.852 0.064604 .
nN2                            -3.594e-01  1.773e-01  -2.027 0.043152 *
nN3                            -5.461e-01  2.319e-01  -2.355 0.018904 *
nN4                            -9.471e-01  3.729e-01  -2.540 0.011379 *
nN5                            -1.239e+00  3.638e-01  -3.404 0.000714 ***
nN6                            -7.709e-01  7.896e-01  -0.976 0.329344
nN7                            -2.425e+00  7.606e-01  -3.189 0.001514 **
nN11                           -1.243e+00  1.399e+00  -0.888 0.374702
C.0401                          3.096e-01  4.596e-01   0.674 0.500803
C.0402                          3.461e+00  8.088e-01   4.279 2.24e-05 ***
C.0403                          9.508e+01  1.191e+02   0.798 0.425018
C.0404                          1.195e+00  6.359e+00   0.188 0.850985
C.04011                                NA         NA      NA       NA
TPSA:GATS1p                     1.270e-02  3.884e-03   3.270 0.001146 **
TPSA:MLOGP                     -1.055e-03  4.322e-04  -2.442 0.014949 *
GATS1p:C.0401                  -4.032e-01  3.852e-01  -1.047 0.295725
GATS1p:C.0402                  -2.984e+00  7.016e-01  -4.253 2.50e-05 ***
GATS1p:C.0403                  -6.404e+01  7.905e+01  -0.810 0.418253
GATS1p:C.0404                  -9.057e-01  4.604e+00  -0.197 0.844124
GATS1p:C.04011                         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.147 on 521 degrees of freedom
Multiple R-squared:  0.5465,    Adjusted R-squared:  0.5256
F-statistic: 26.16 on 24 and 521 DF,  p-value: < 2.2e-16
```

Figure 11: Section 2.5 interaction model R output

References

[1] Cassotti, M., Ballabio, D., Consonni, V., Mauri, A., Tetko, I. V., Todeschini, R. (2014). Prediction of Acute Aquatic Toxicity toward Daphnia Magna by using the GA-kNN Method. Alternatives to Laboratory Animals, 42(1), 31–41. https://doi.org/10.1177/026119291404200106

[2] Center for Chemical Process Safety. (2021, November 27). *Lethal Concentration 50 (LC50).* https://www.aiche.org/ccps/resources/glossary/process-safety-glossary/lethal-concentration-50-lc50

[3] UCI Machine Learning Repository. (2021, November 24). *QSAR aquatic toxicity Data Set.* https://archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity